ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

Ψηφιακή Επεξεργασία Φωνής

Ενότητα 2η: Παραγωγή και Ταξινόμηση Σημάτων Φωνής

Στυλιανού Ιωάννης

Τμήμα Επιστήμης Υπολογιστών

CS578- Speech Signal Processing

Lecture 2: Production and Classification of Speech Sounds

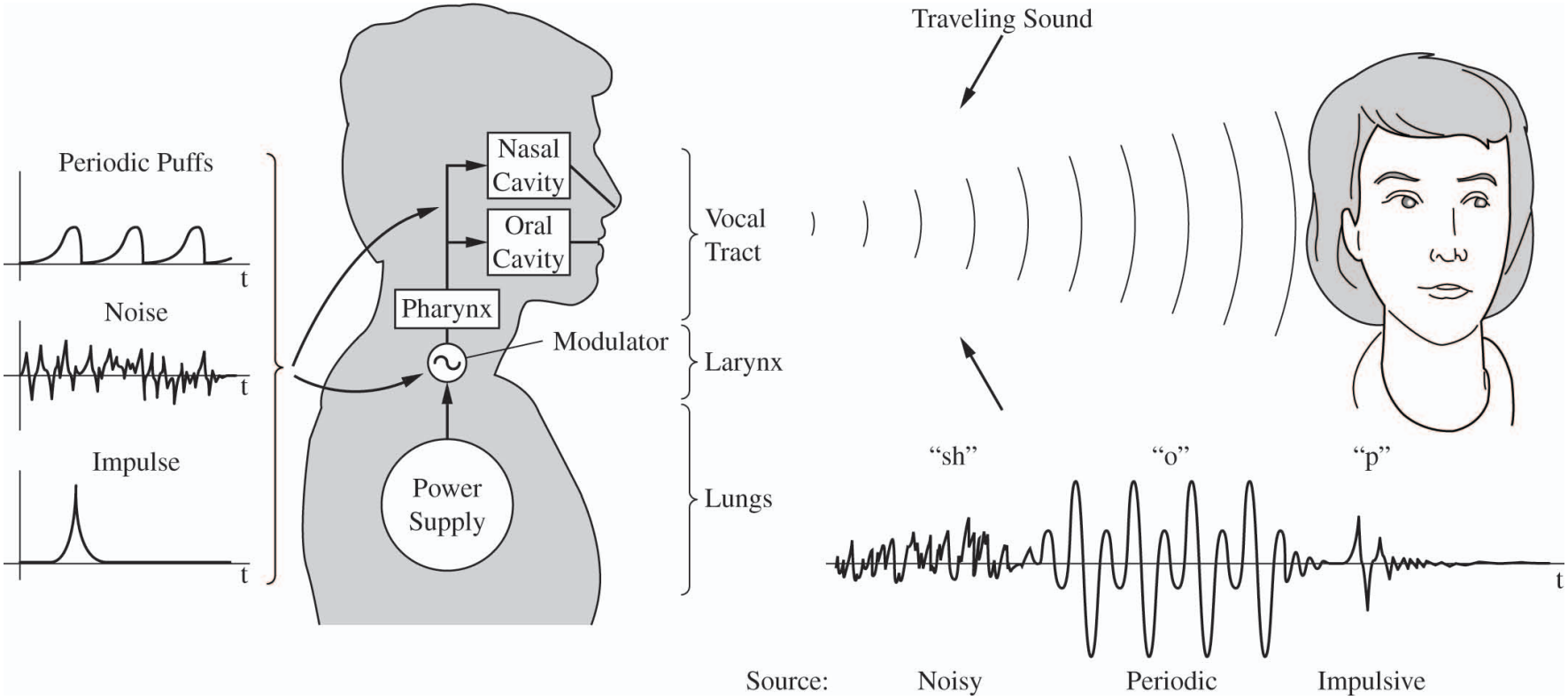
Yannis Stylianou

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

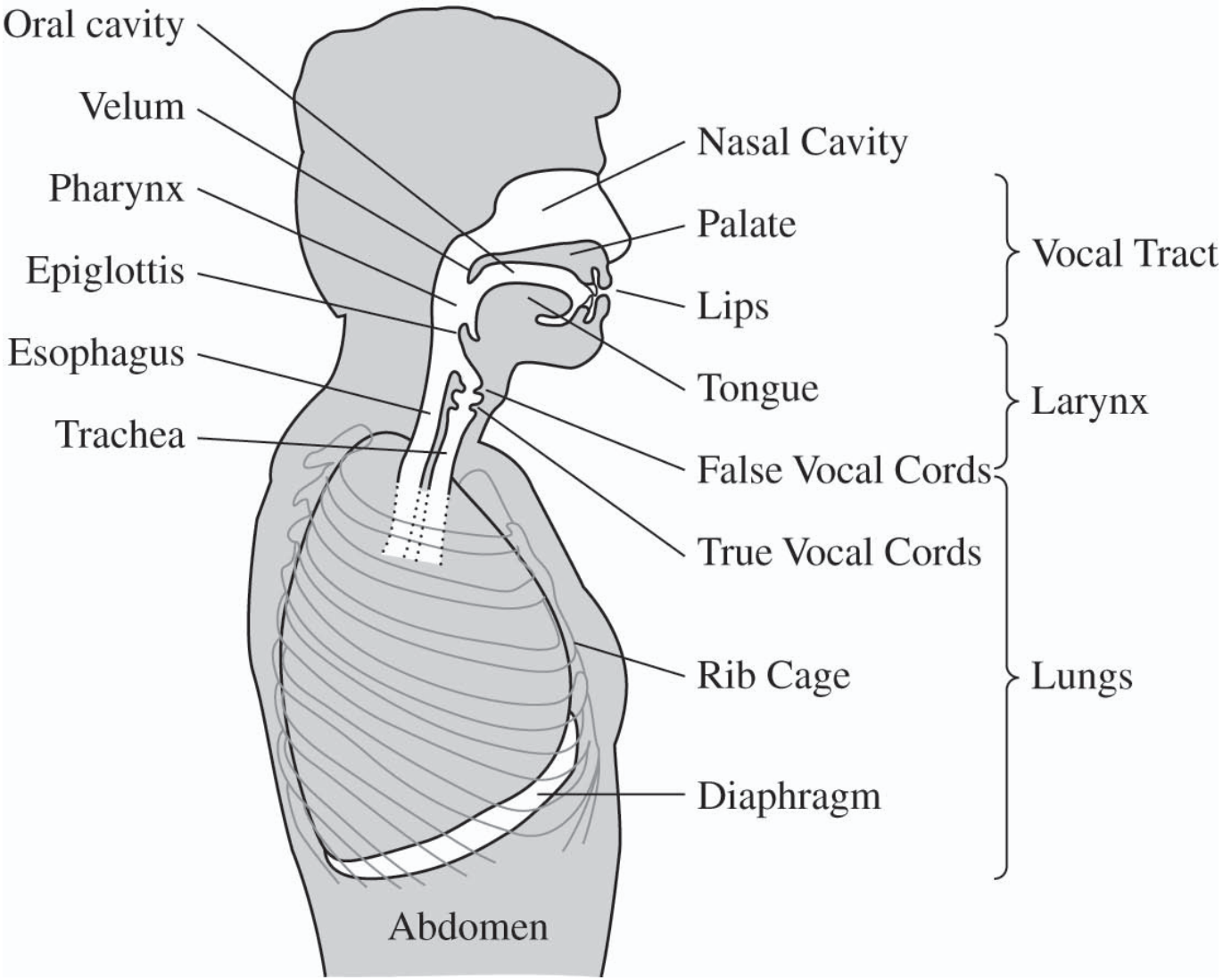
Univ. of Crete, 2008 Winter Period

Outline

A Simple view

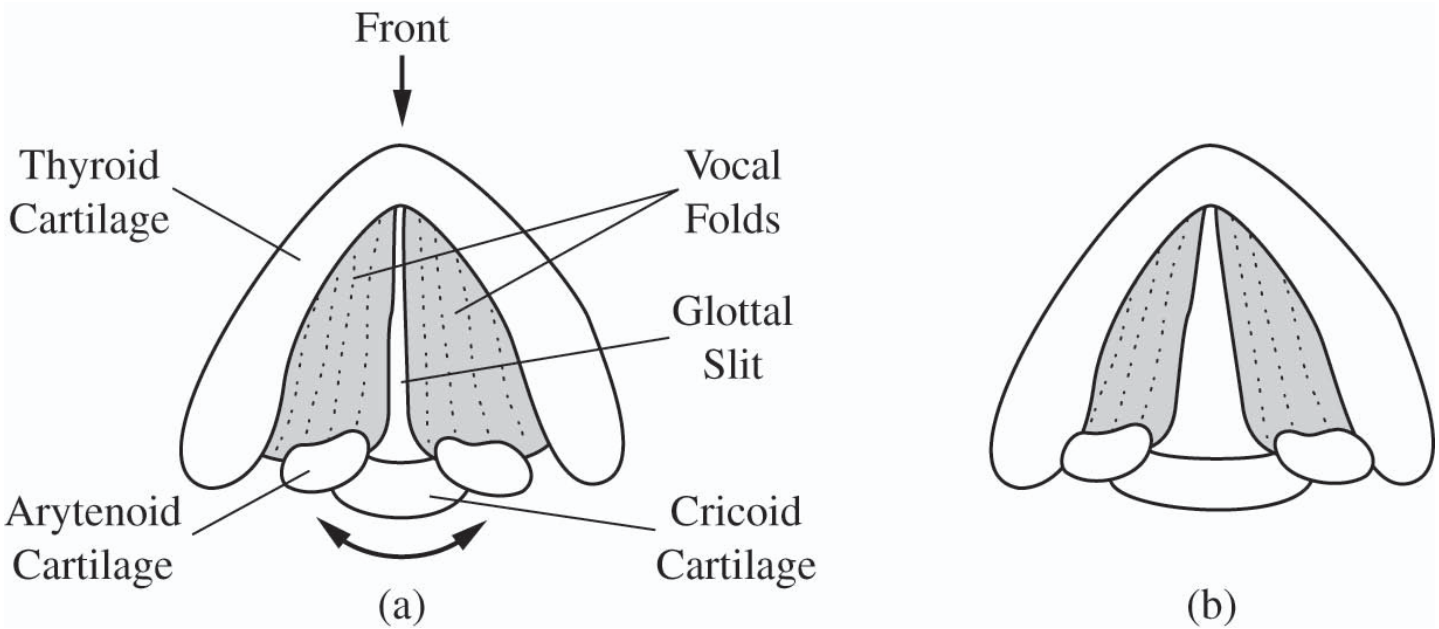


Cross sectional view

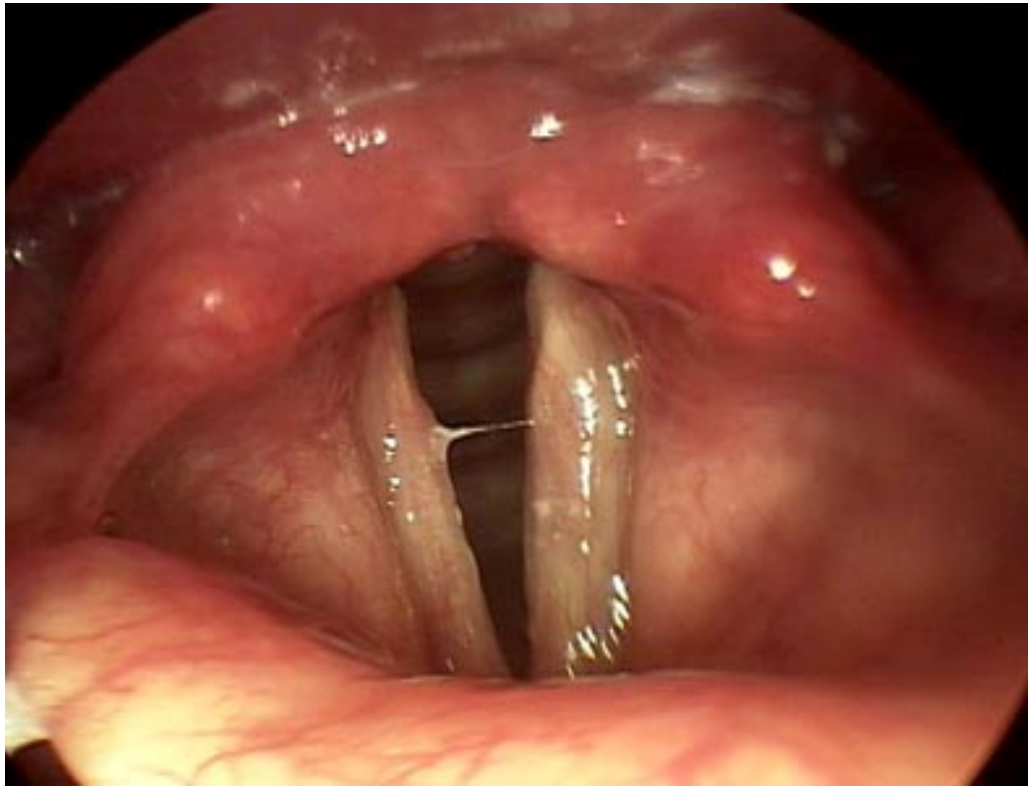


Downward-looking into the larynx: Vocal folds

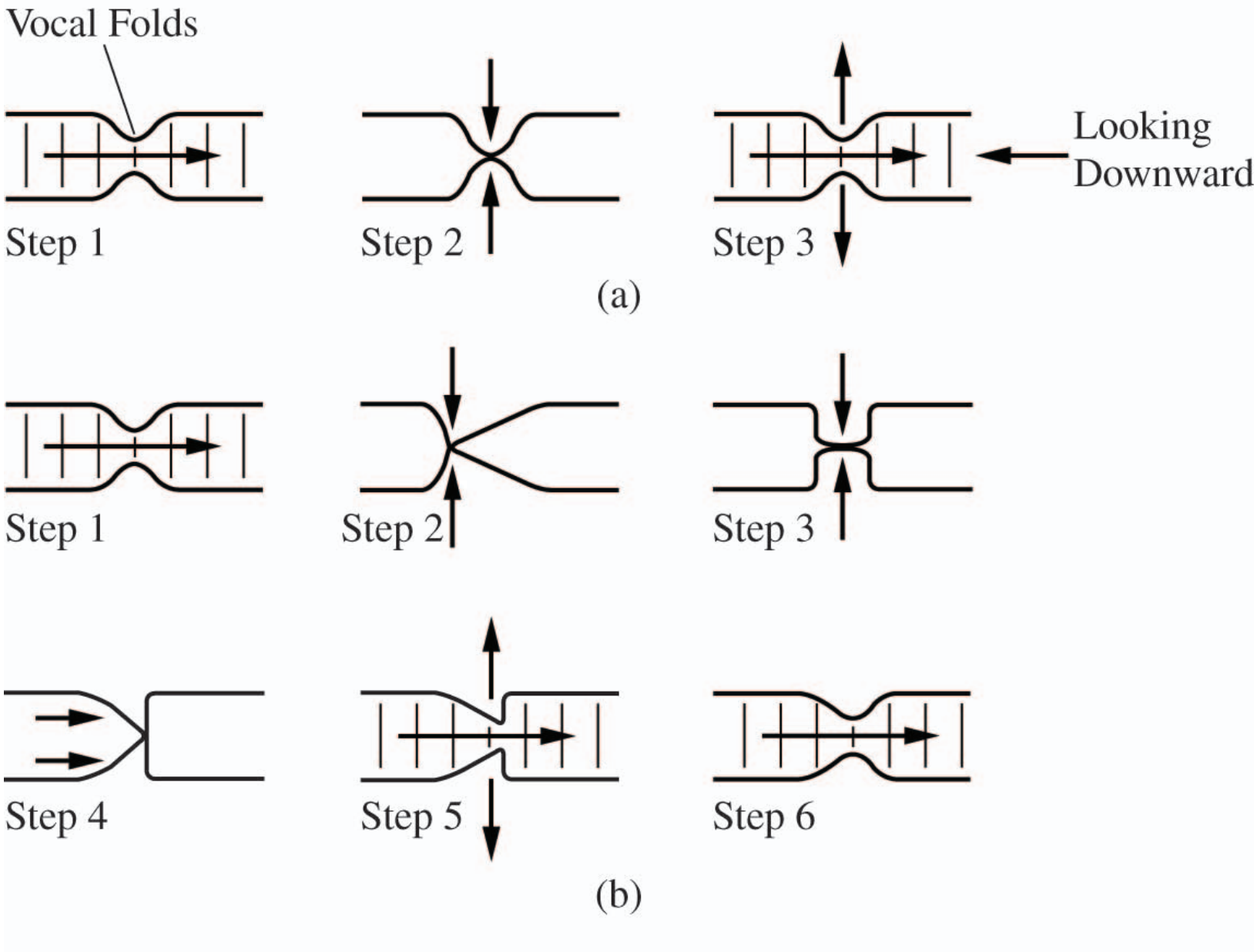
Left: Voicing, **Right:** Breathing



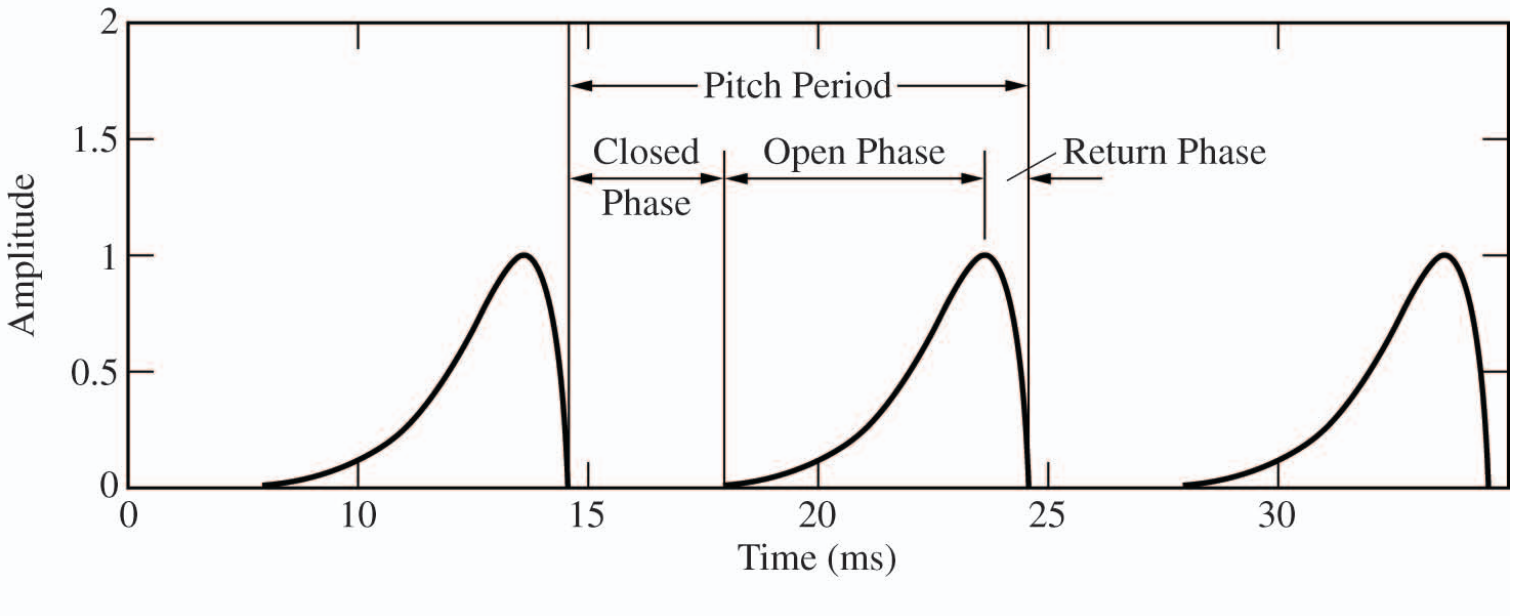
Vocal Folds vibration



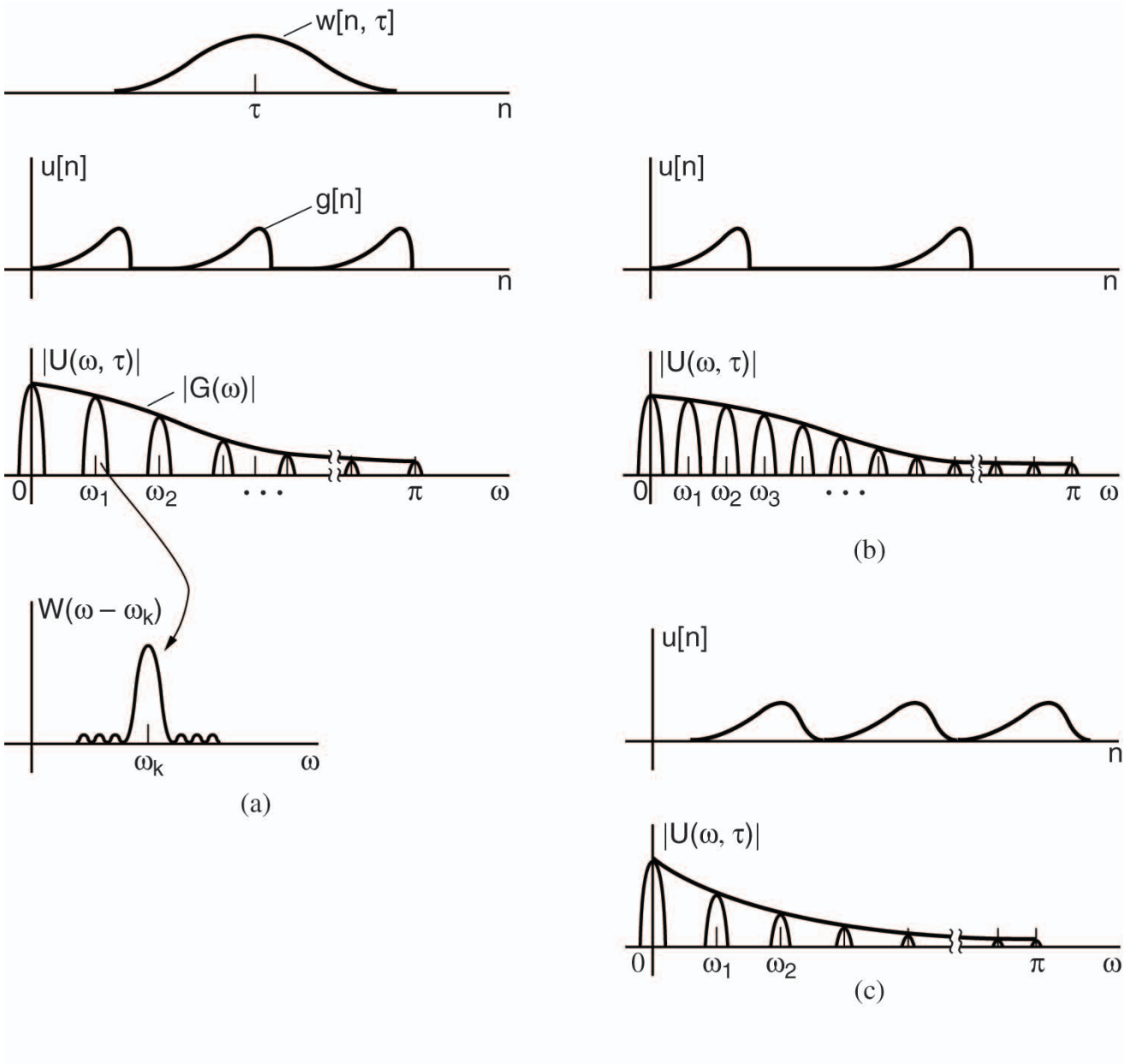
Bernoulli's principle in the glottis



Glottal airflow velocity

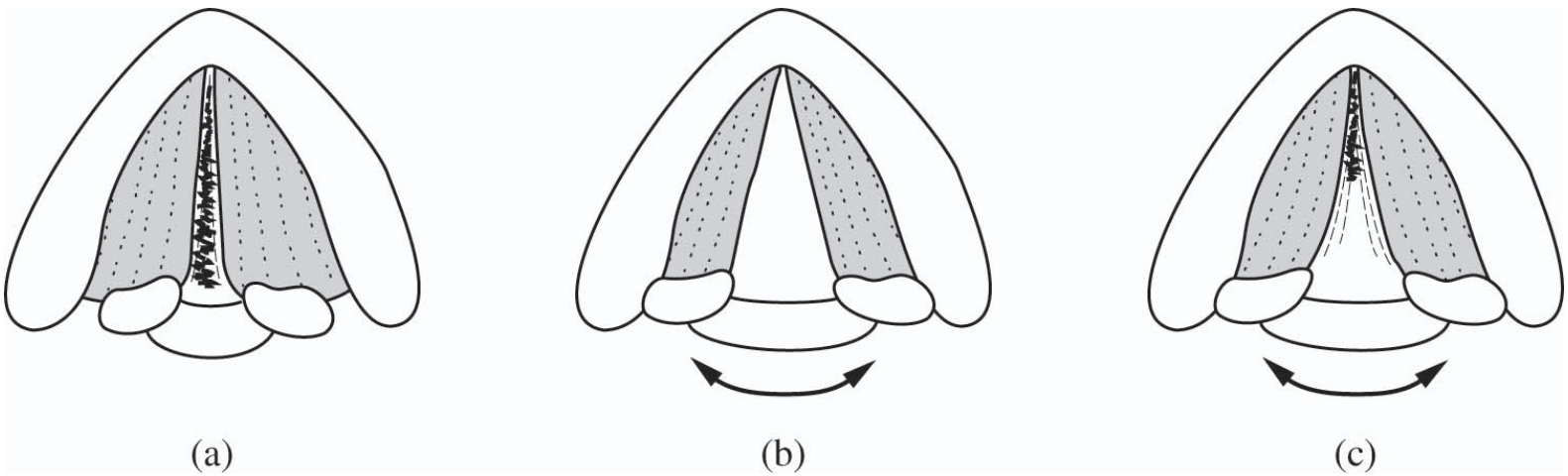


Softer, typical, and relaxed glottal flow



Other vocal folds configurations

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



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

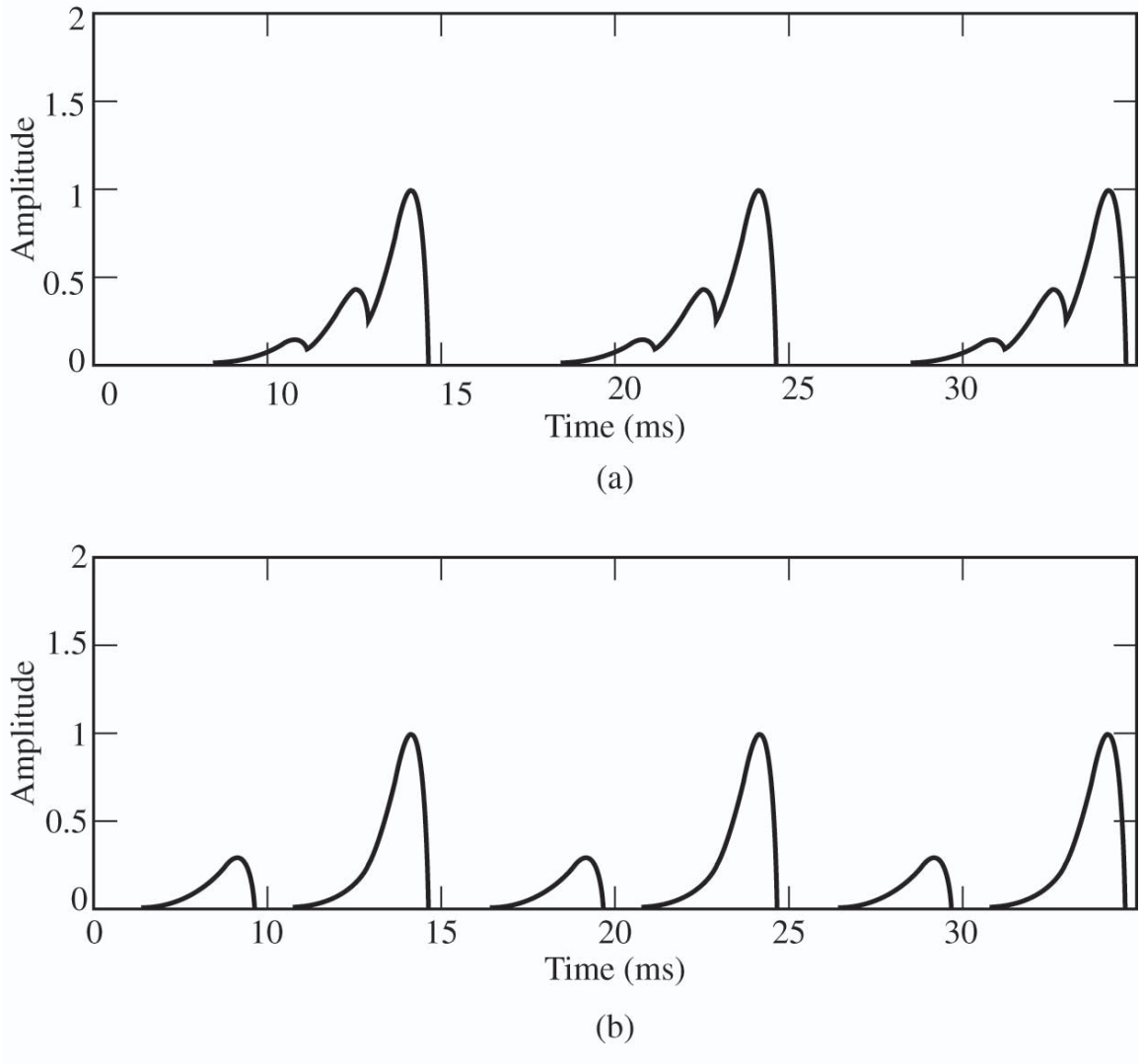
  folds are massy and relaxed
  abnormally low and irregular pitch
  secondary pulses during open phase

- Diplophonia

  extra flaps
  secondary pulses during the closed phase

Examples

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

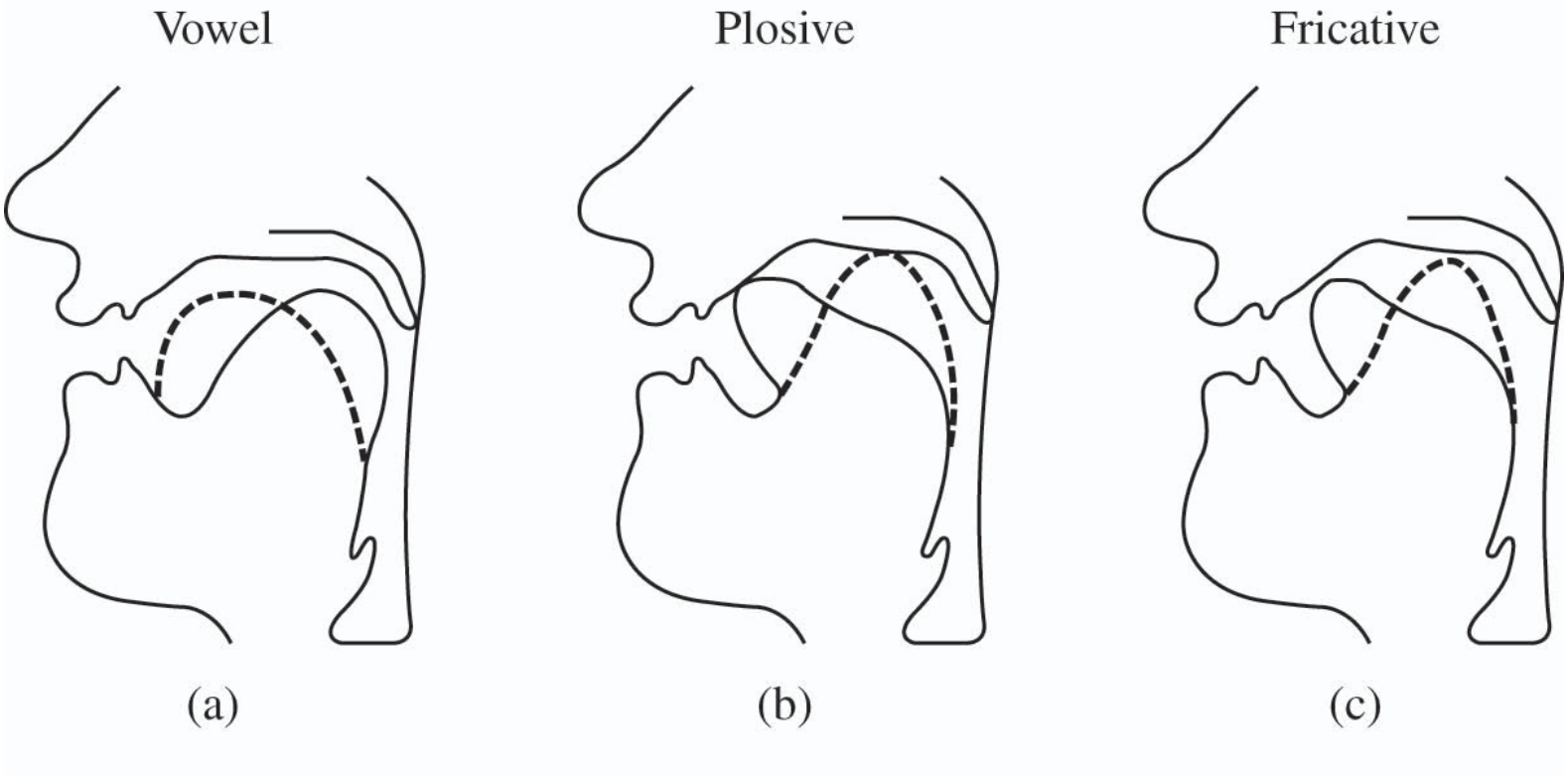


Vocal Tract

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:

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

Example

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

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

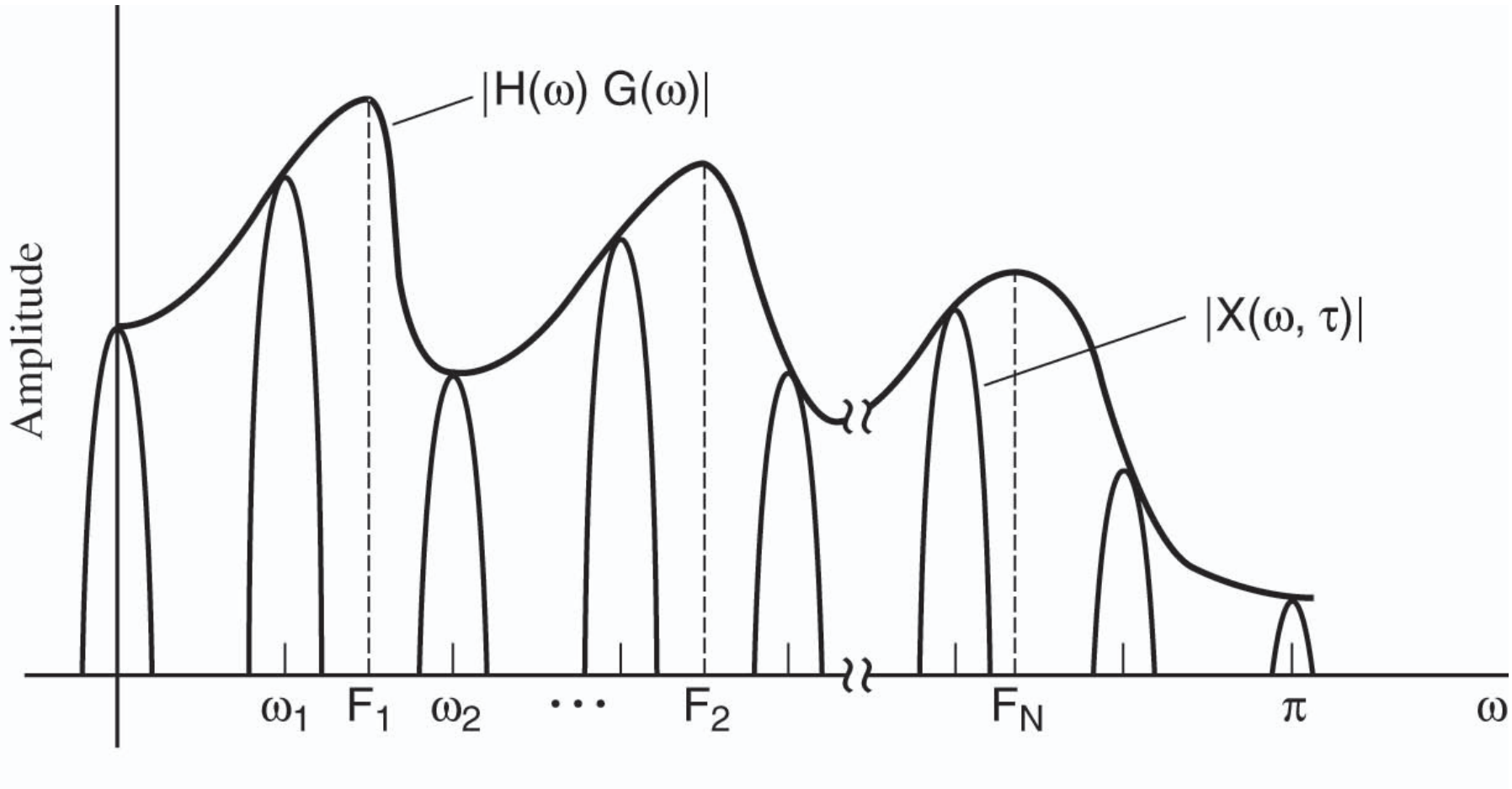
then, the output speech, $x[n, \tau]$, is given by:

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

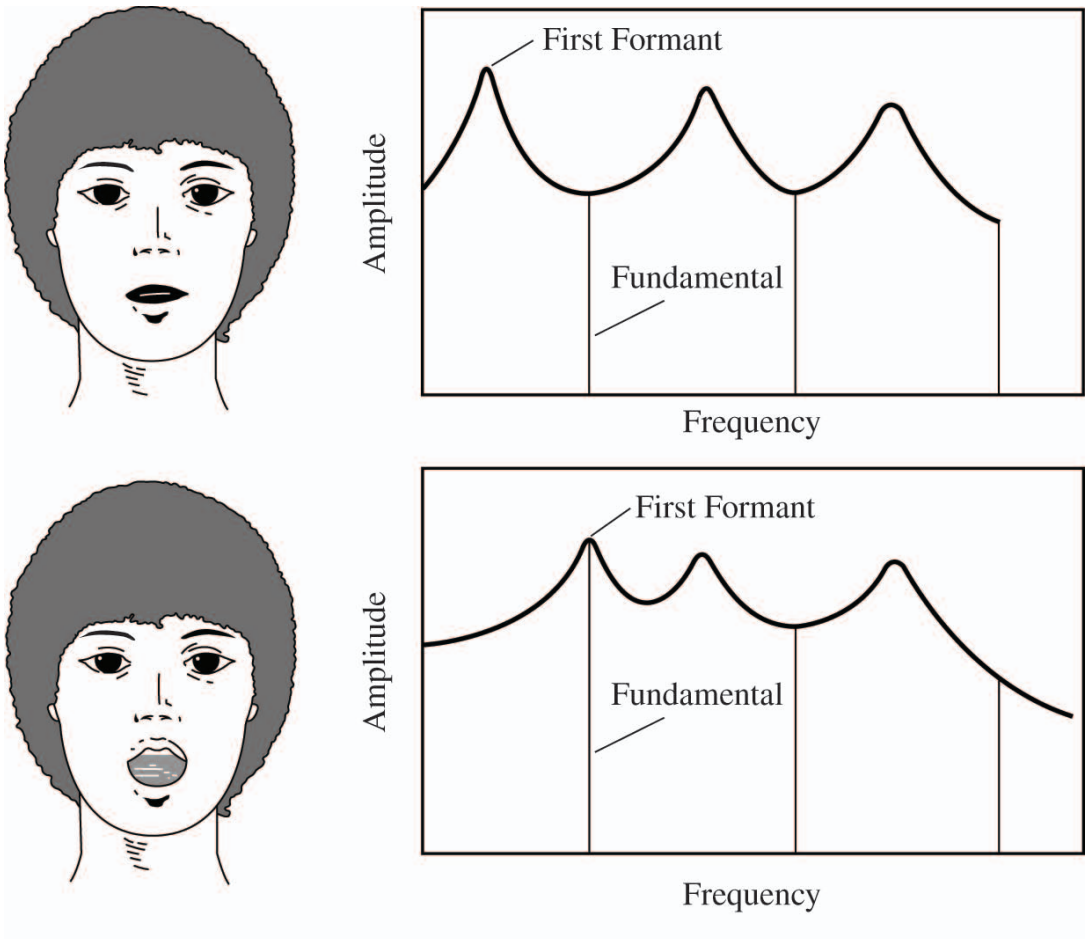
and

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

Harmonics and Formants



Soprano

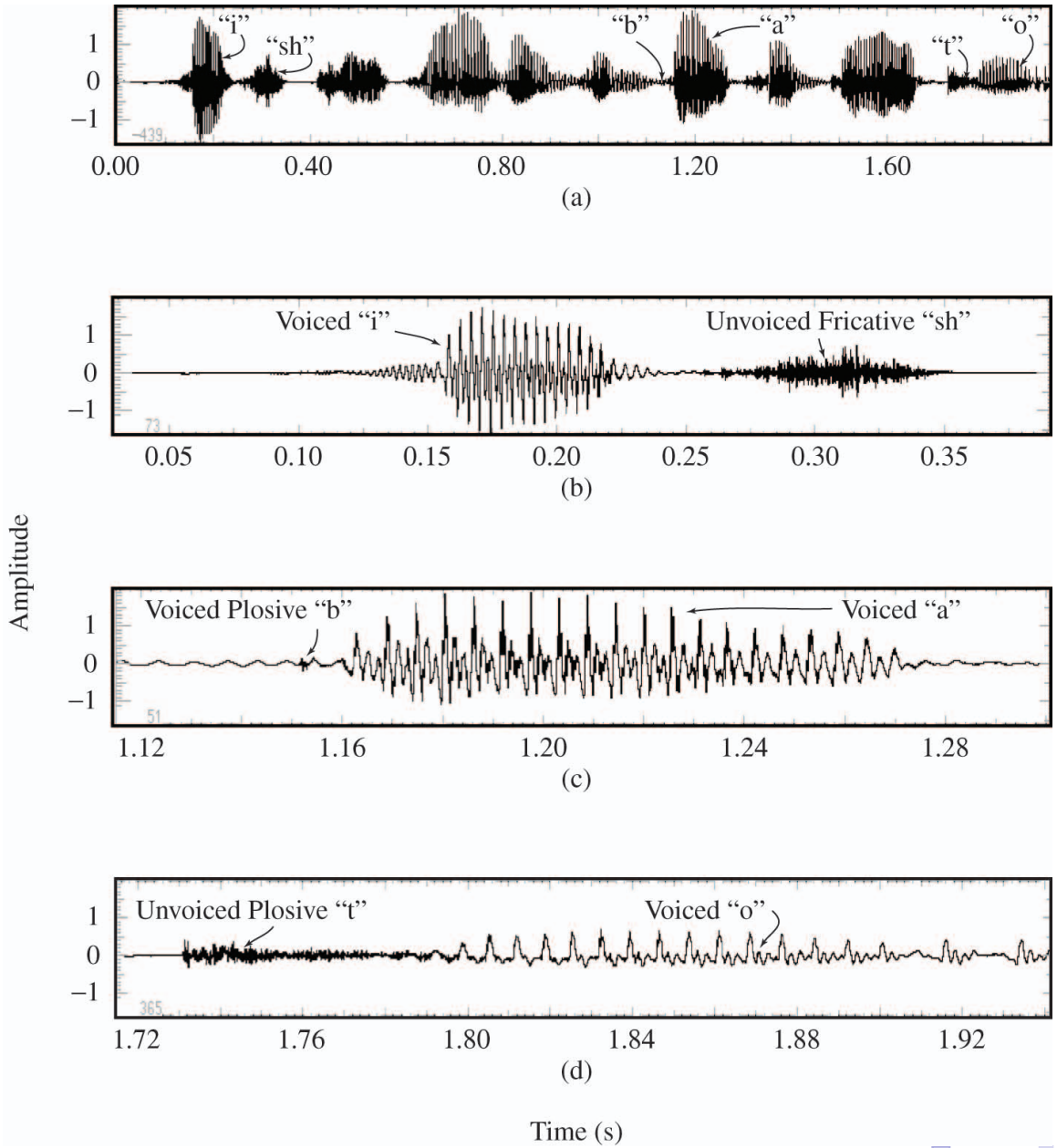

First Formant
Amplitude
Fundamental
Frequency
First Formant
Amplitude
Fundamental
Frequency

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:

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

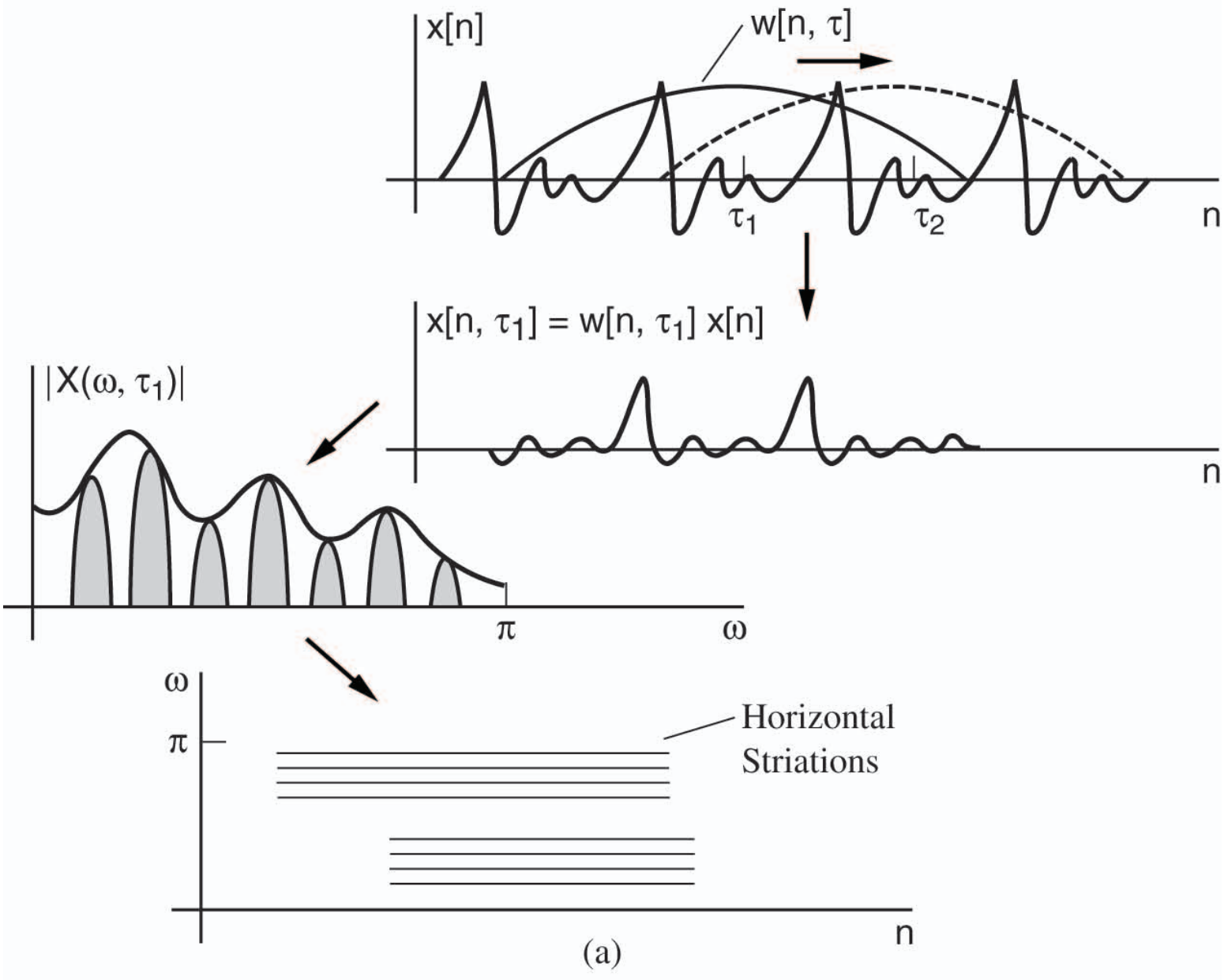
where

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

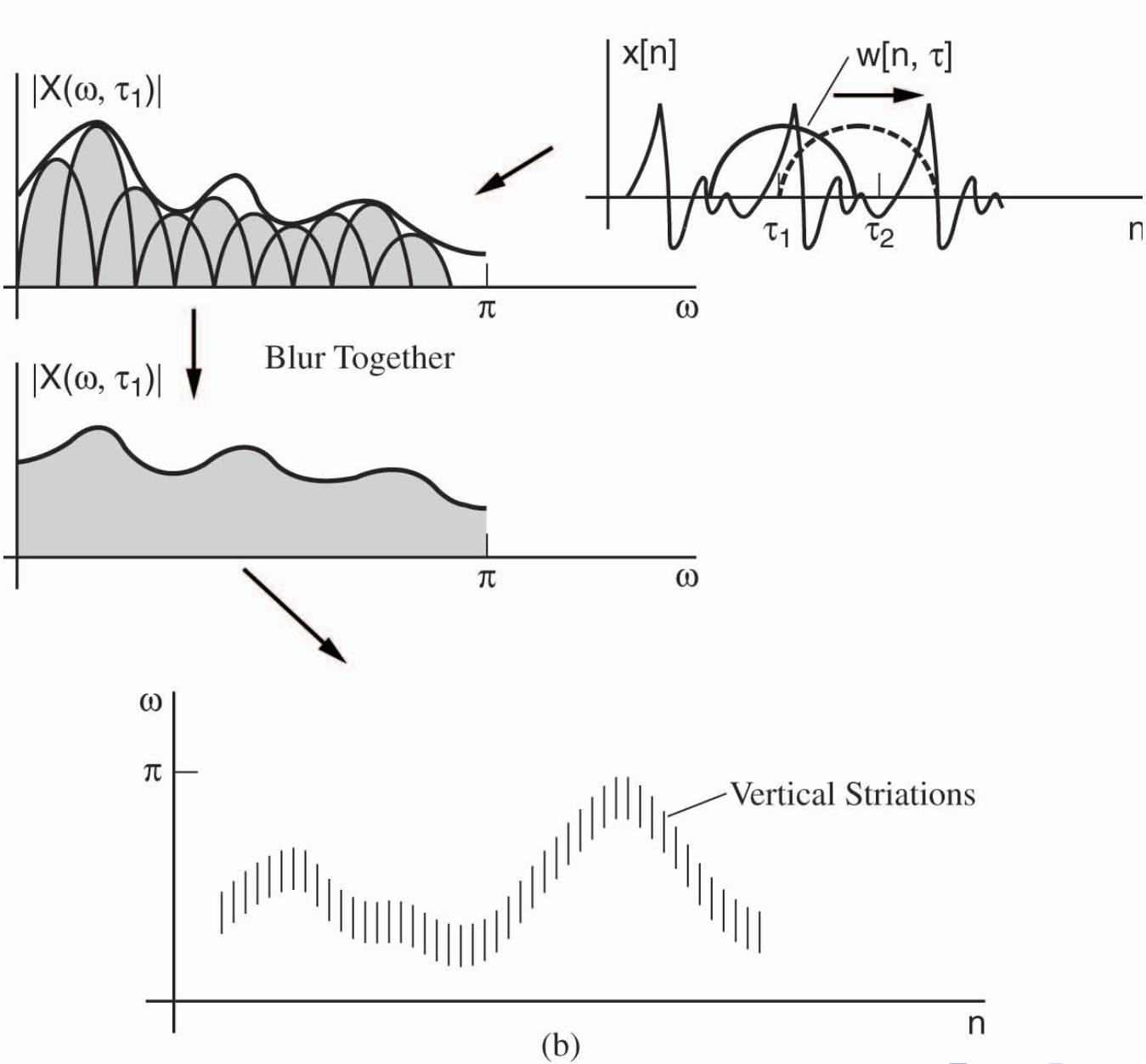
Spectrogram:

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

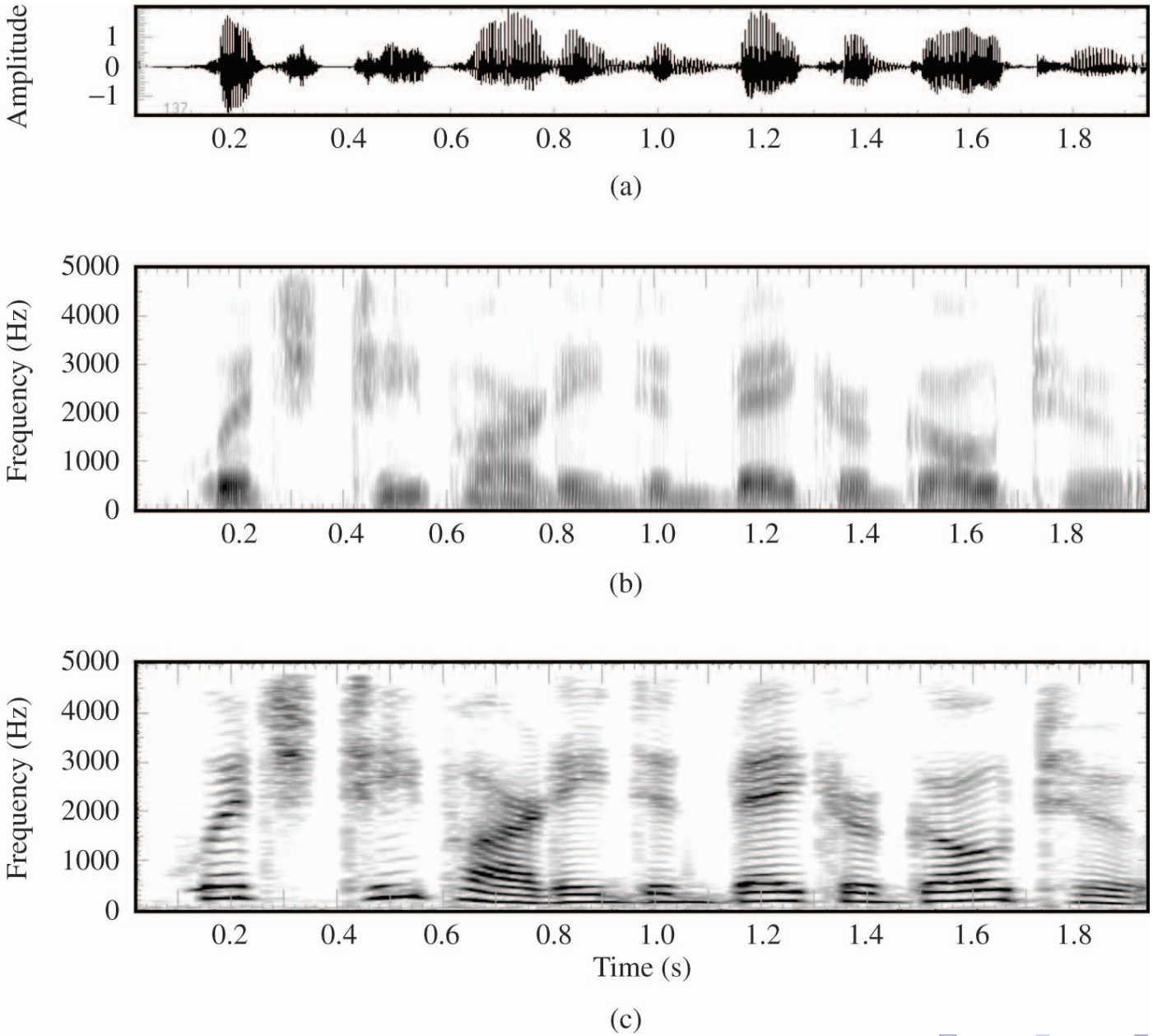
Narrowband Spectrogram



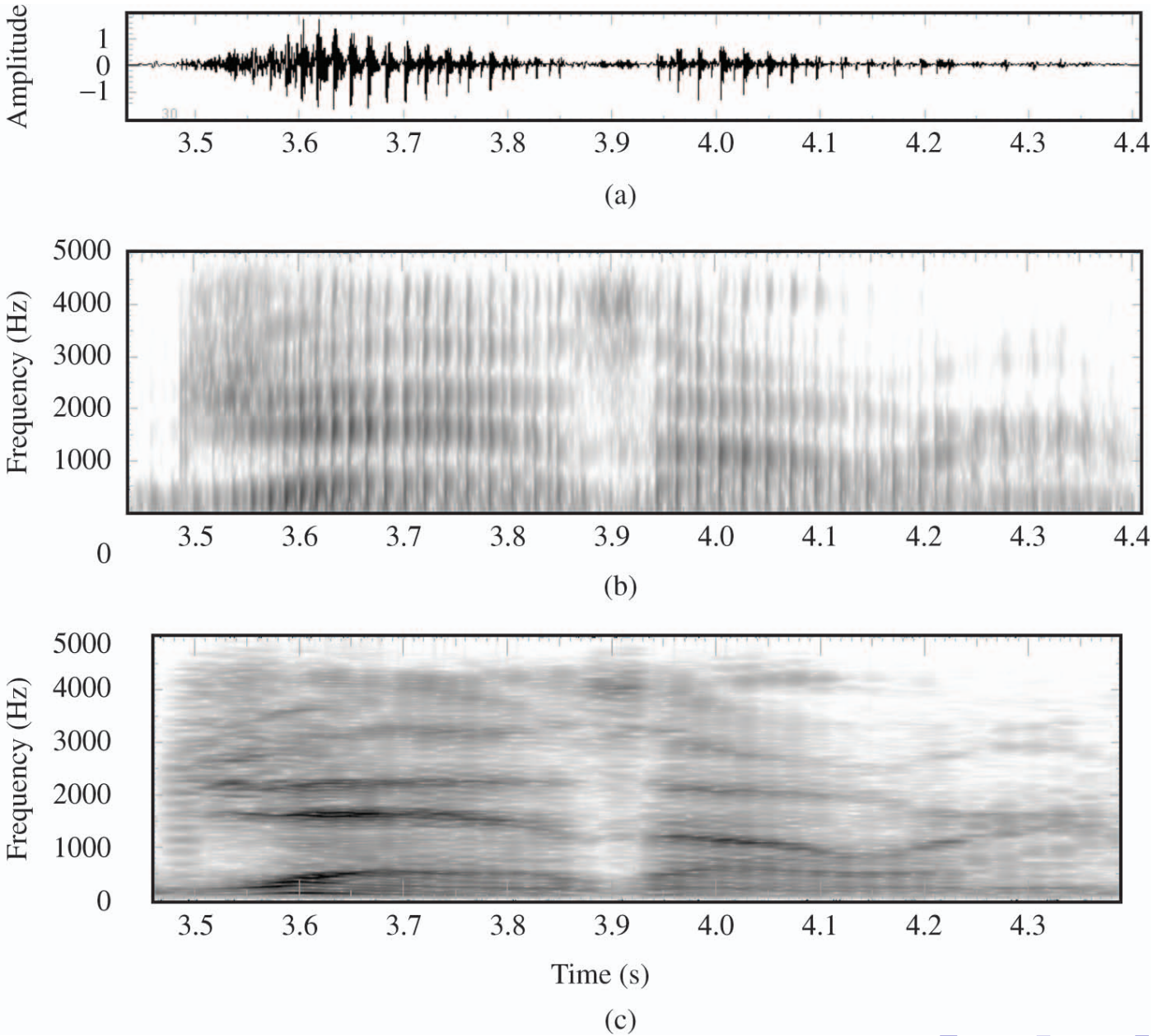
Wideband Spectrogram



Spectrogram on speech



Spectrogram on speech; another example



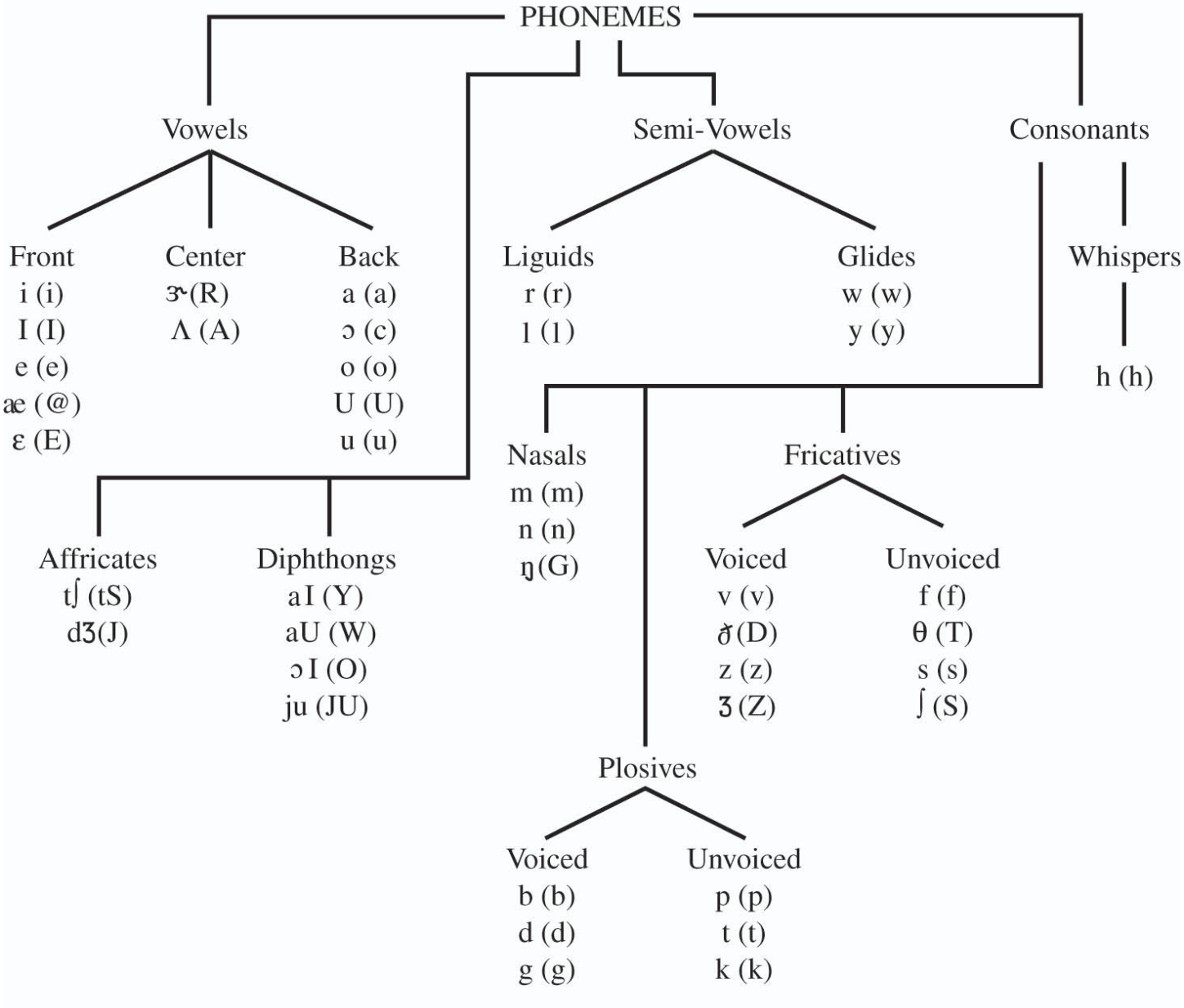
Do we know better now?

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

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

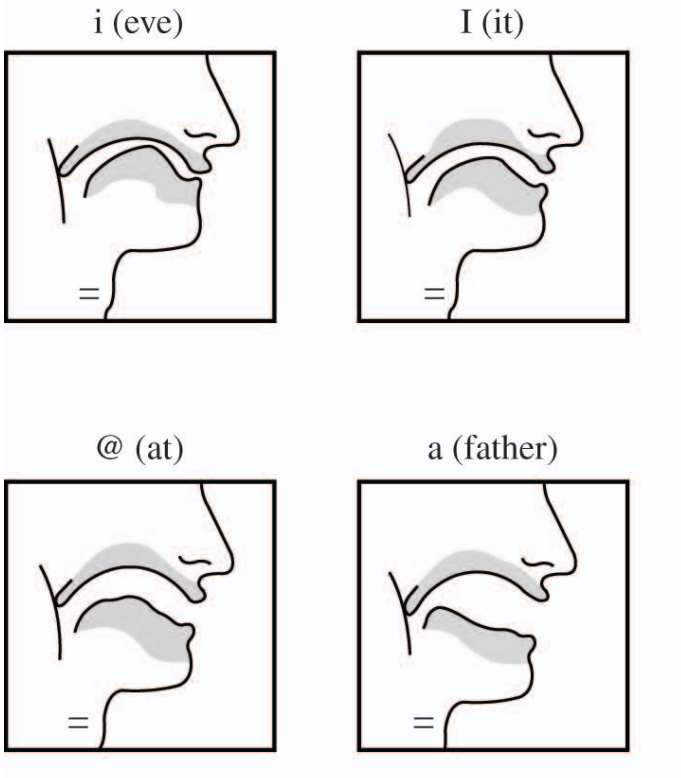
Outline

Phonemes' map

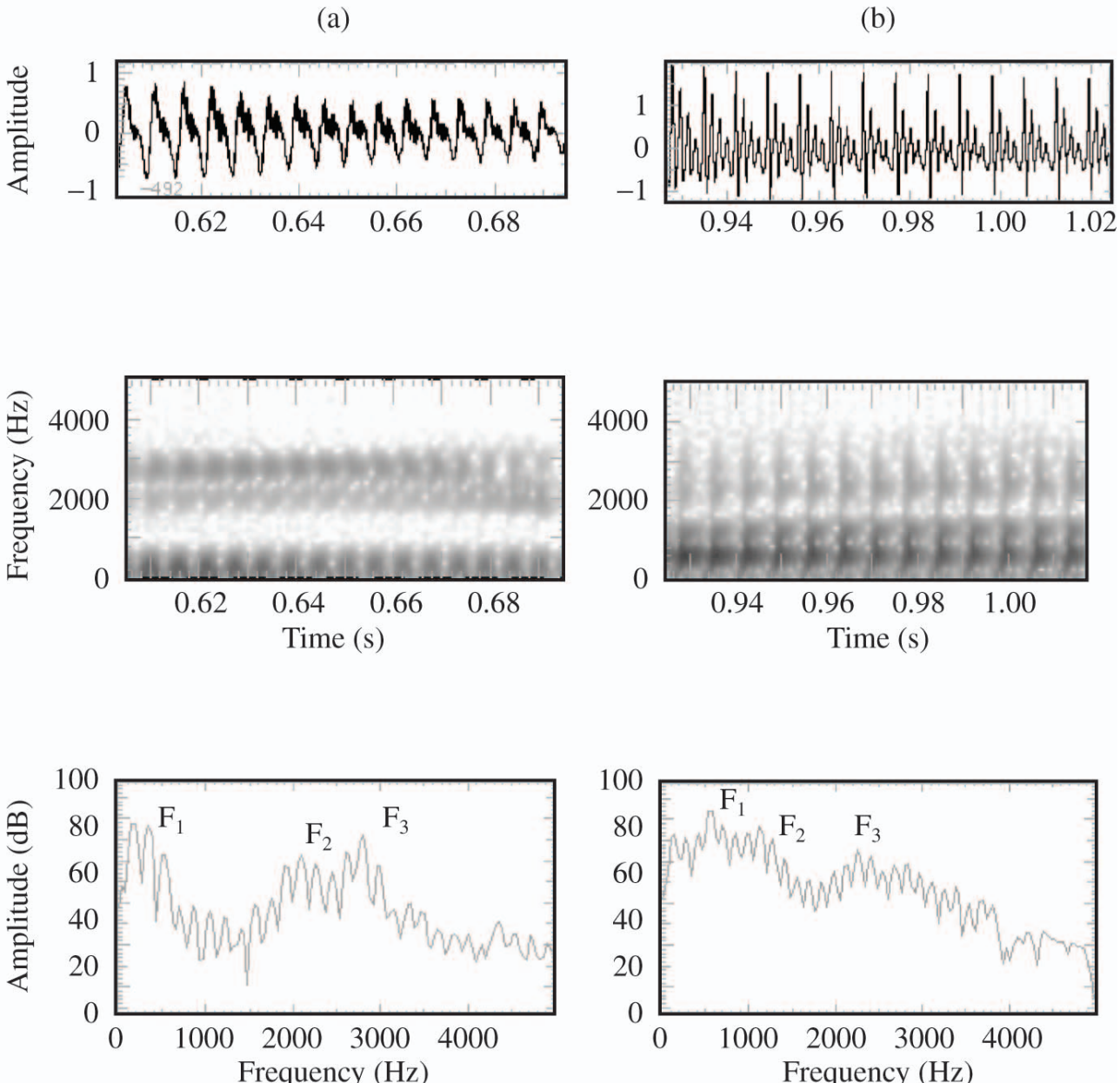


Vowels

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

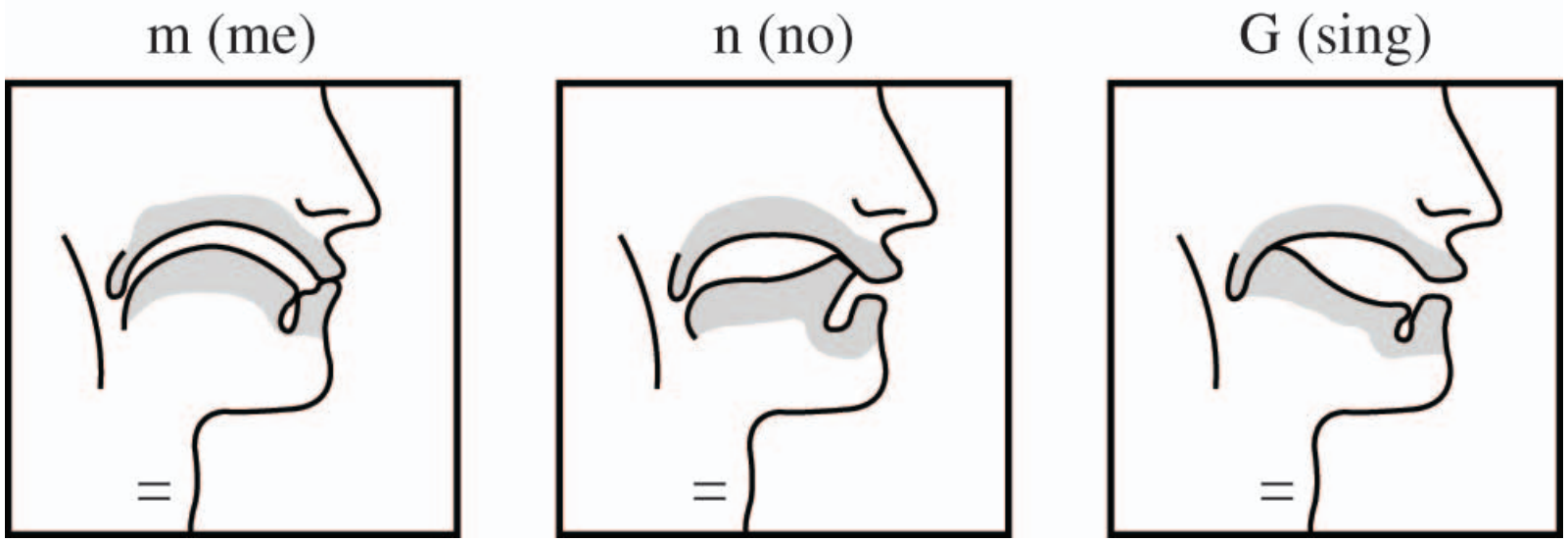


Vowels: time and spectrogram

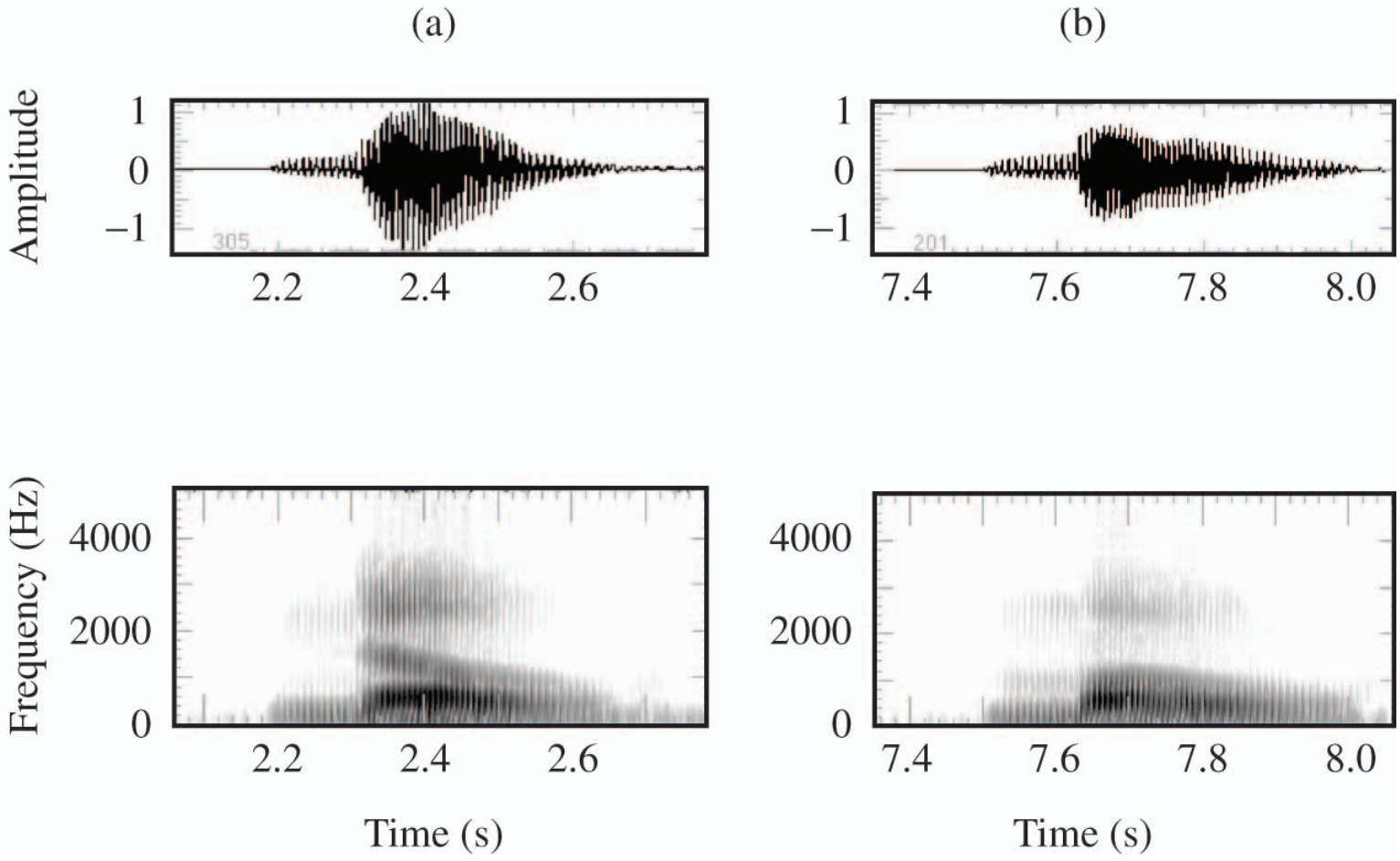


Nasals

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



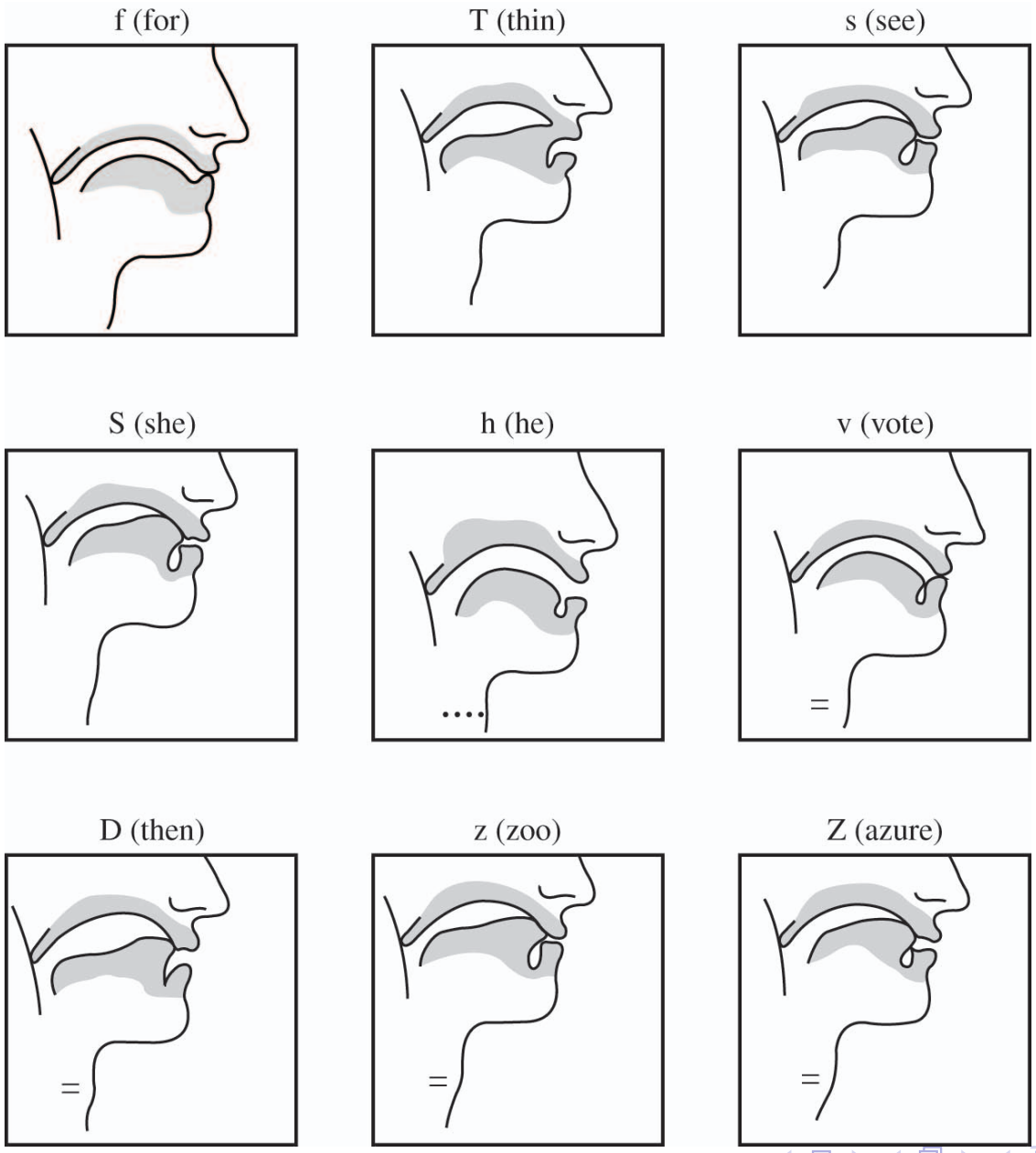
Nasals: time and spectrogram



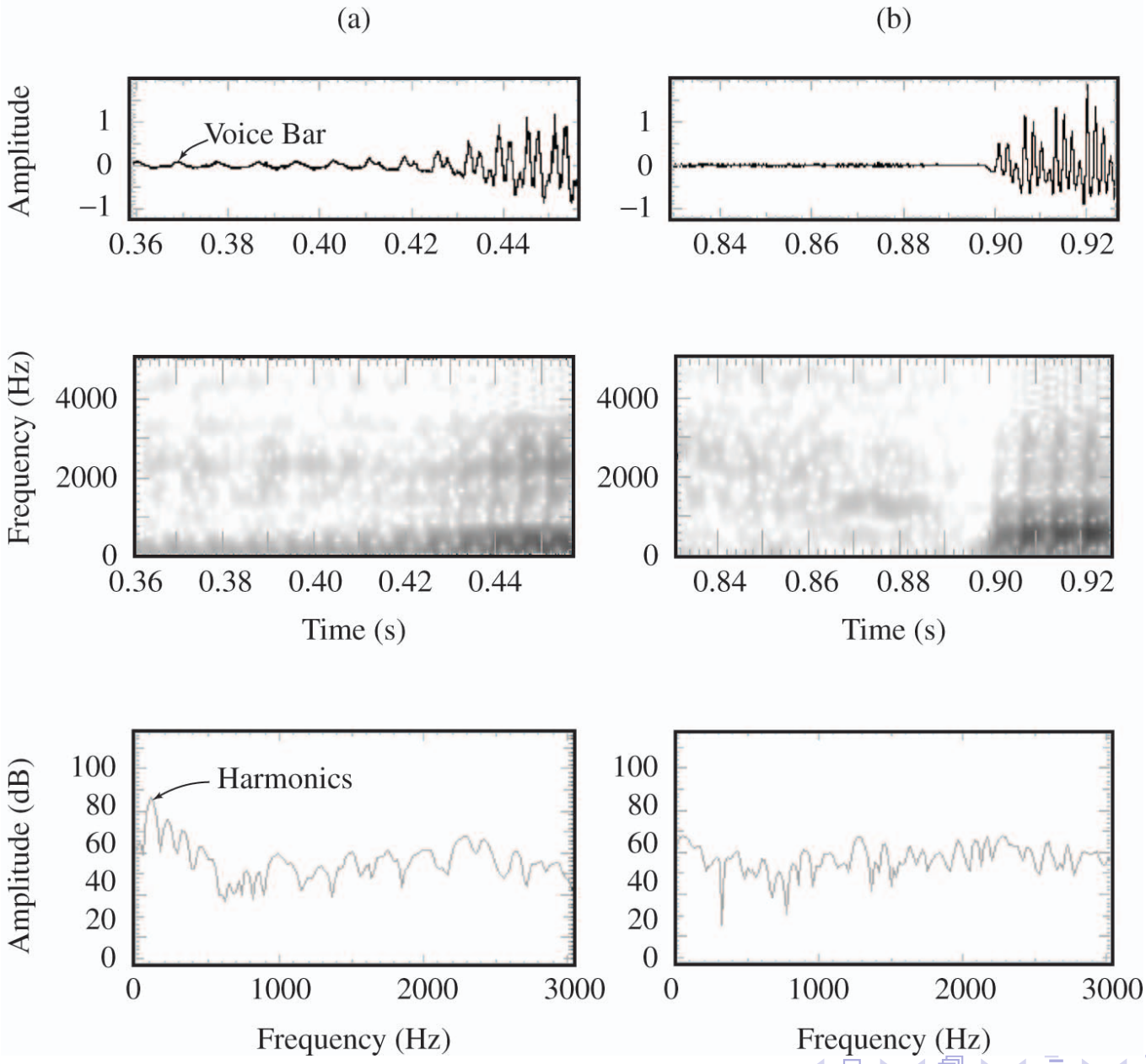
Fricatives

- **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

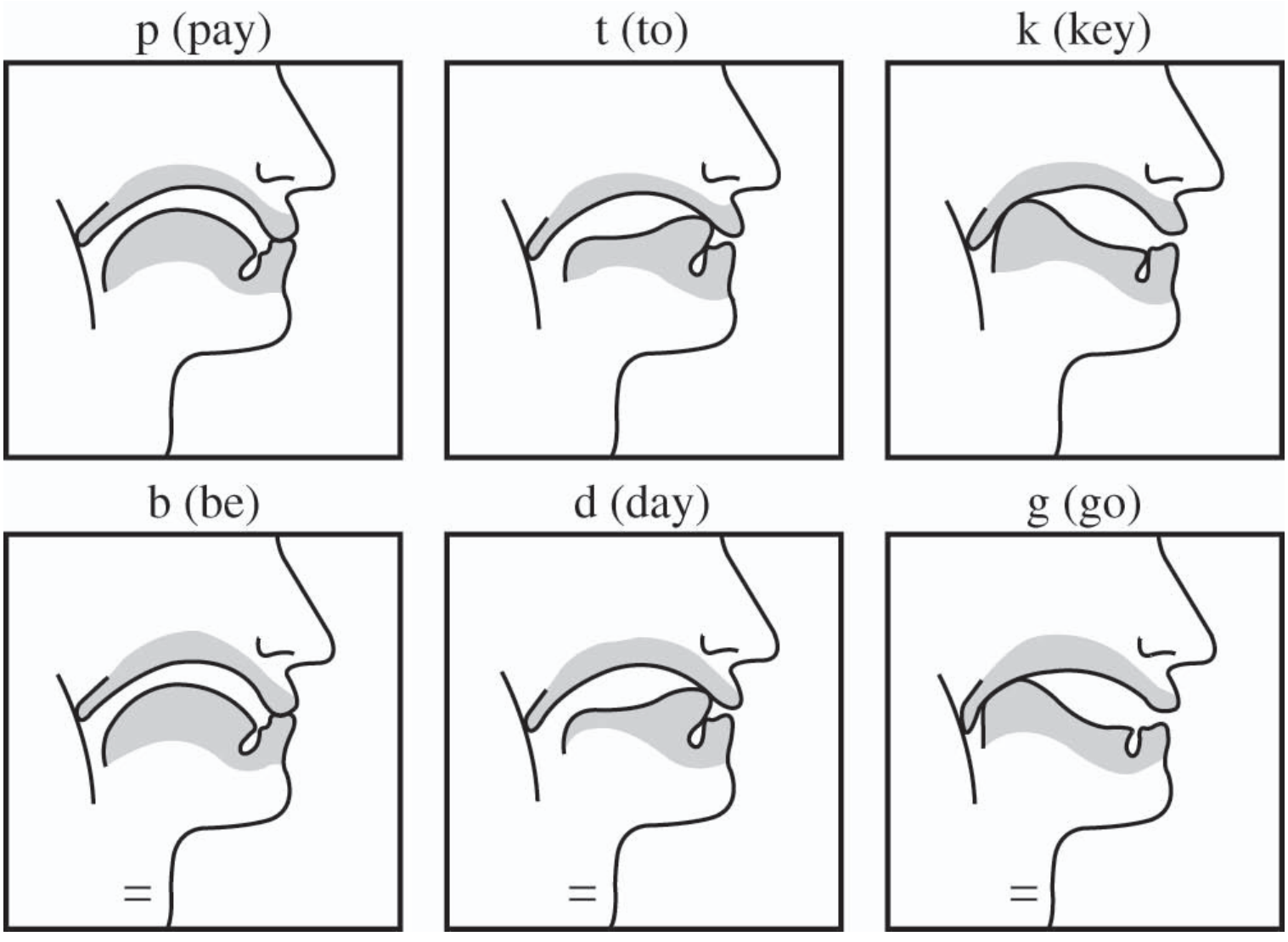
Voiced:

- **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

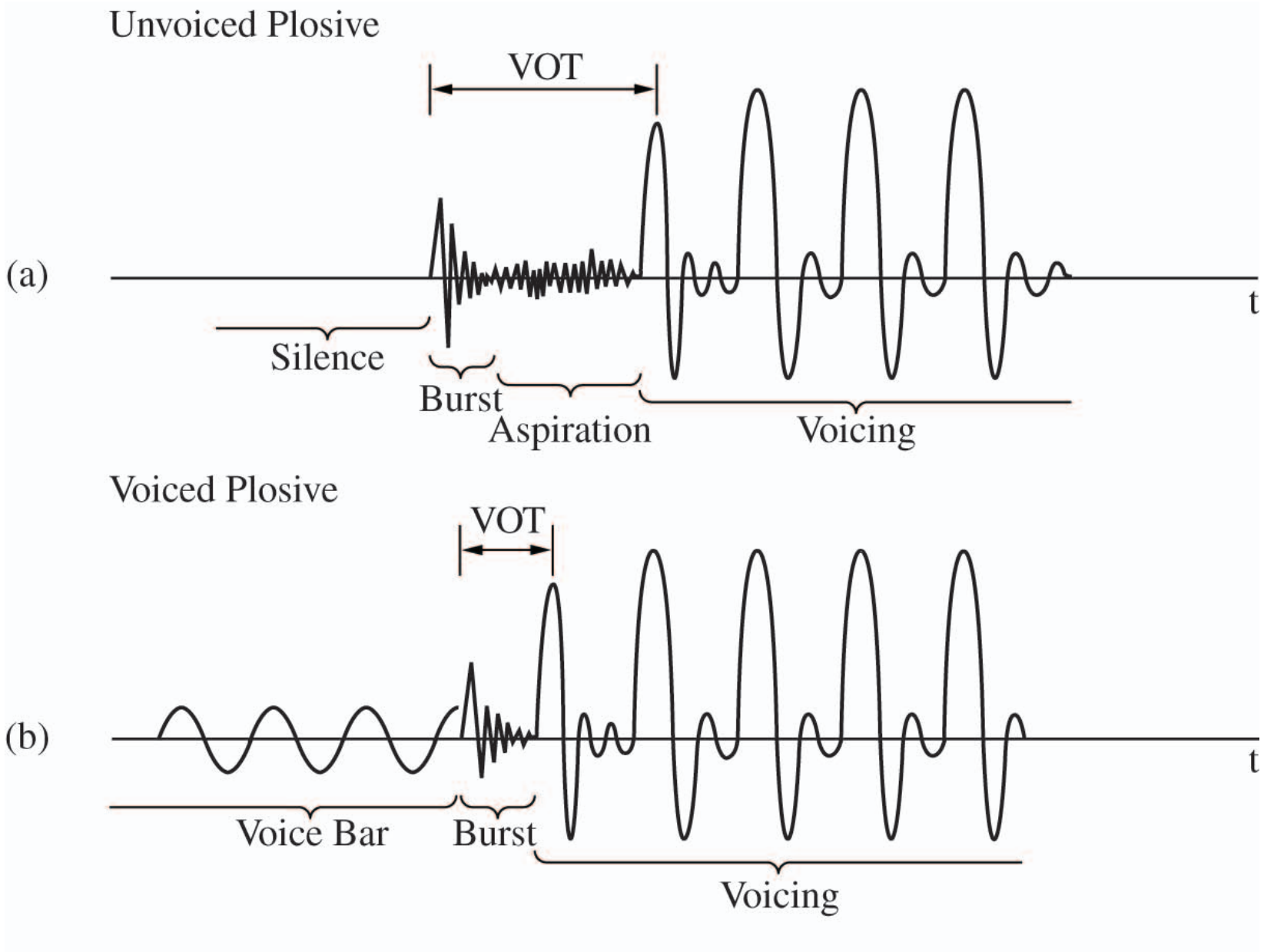
Unvoiced:

- **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

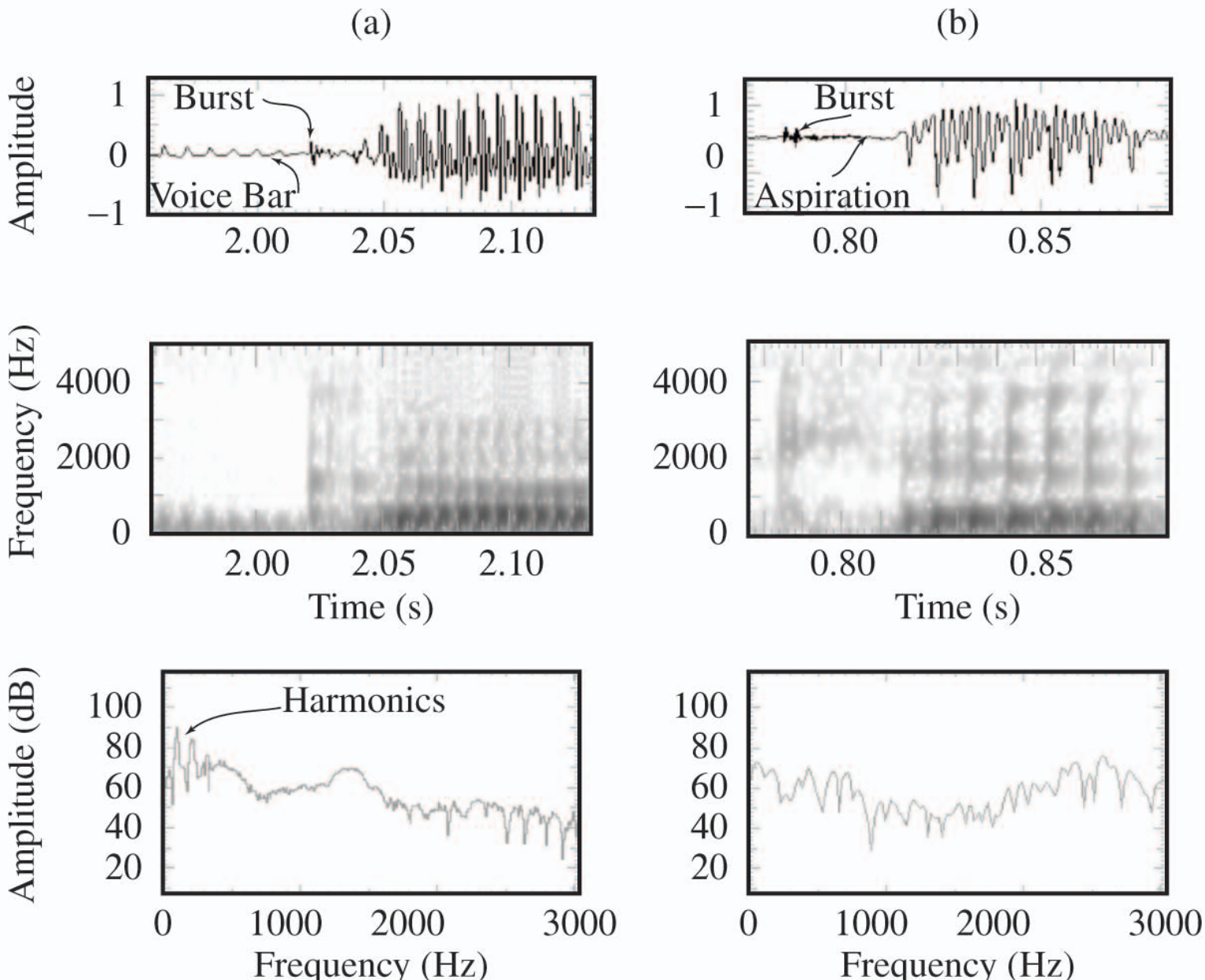
Plosives' profile



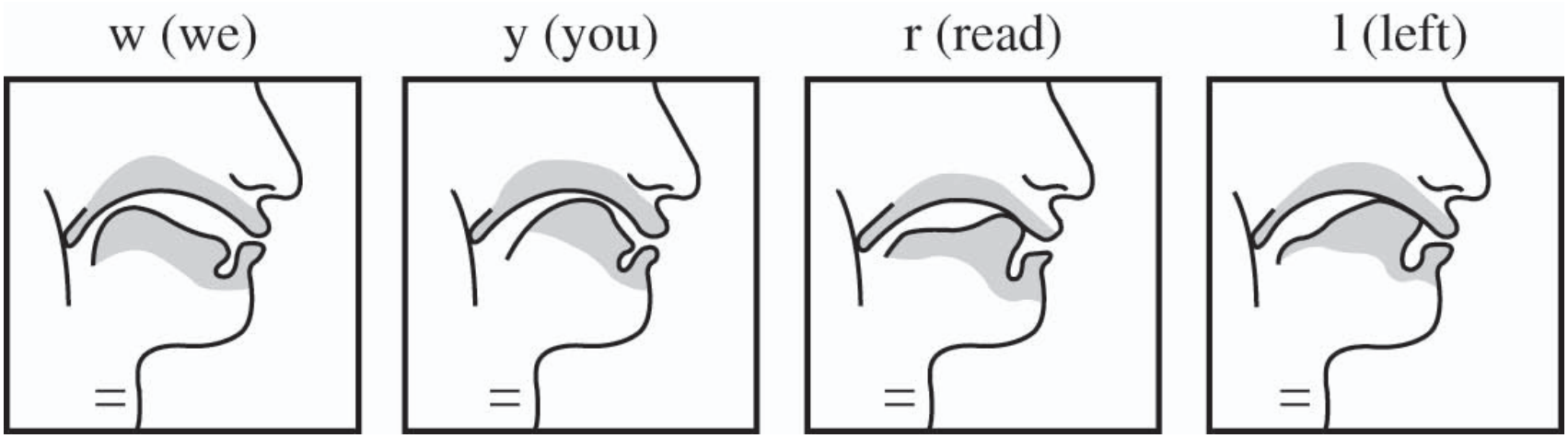
Voice Onset Time



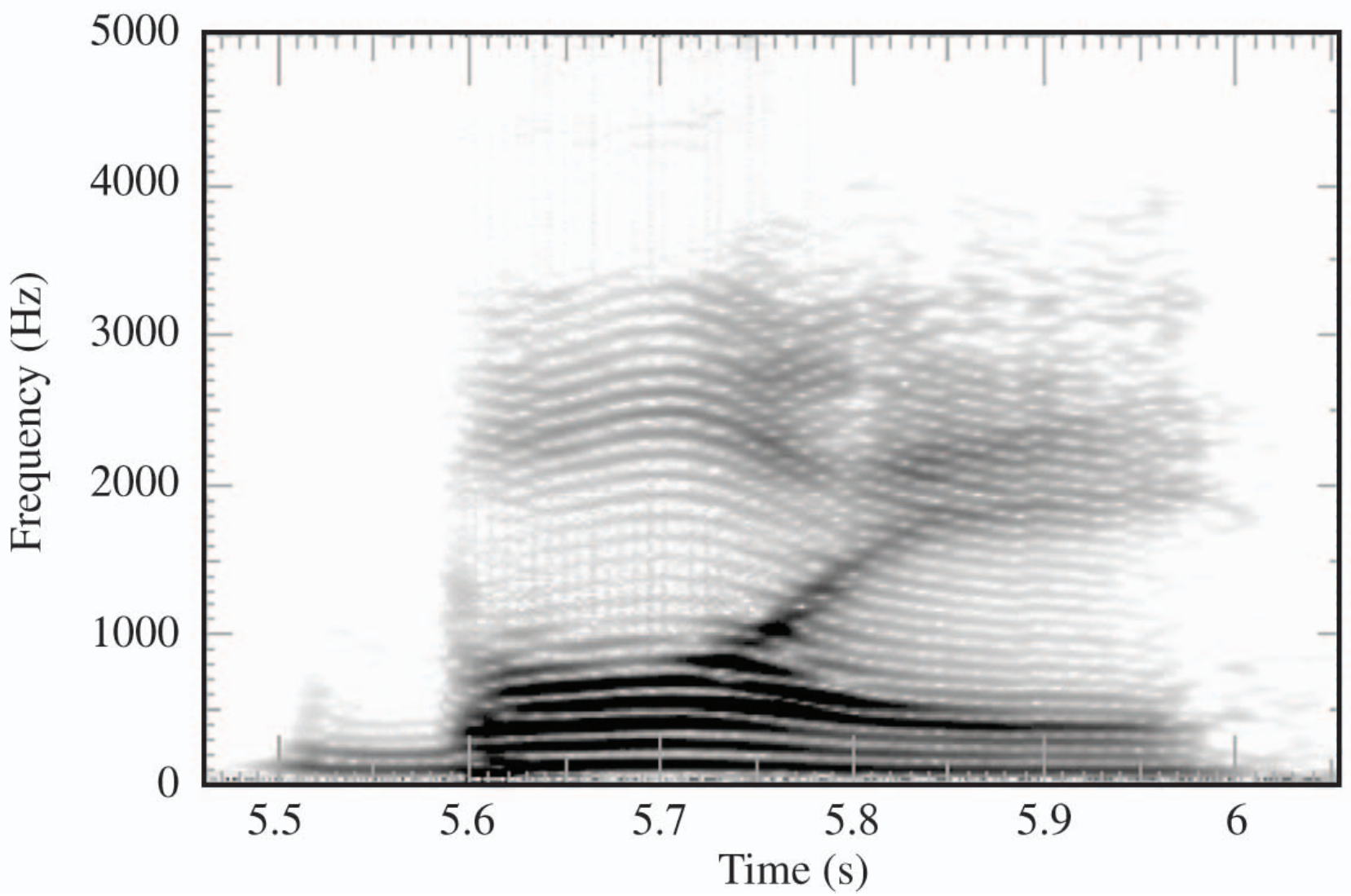
Plosives: time and spectrogram



Semi-vowels



Transitional Speech Sounds: "boy"



Outline

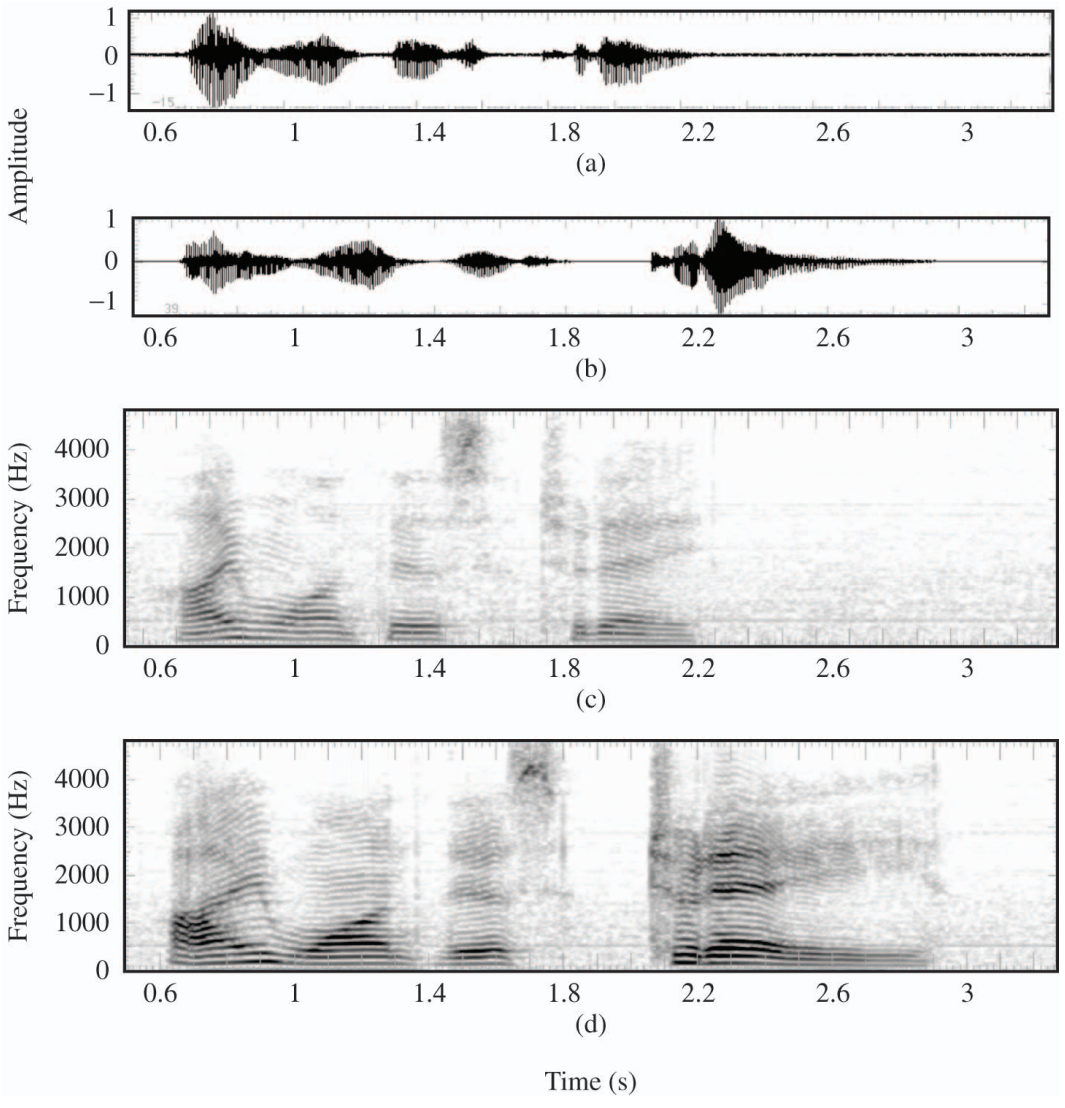
Prosody of speech

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

Τέλος Ενότητας

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Κρήτης**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα αδειοδότησης

- Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά, Μη Εμπορική Χρήση, Όχι Παράγωγο Έργο 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/

- Ως **Μη Εμπορική** ορίζεται η χρήση:
  - που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
  - που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
  - που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

- Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

.

Σημείωμα Αναφοράς

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.

Σημείωμα Χρήσης Έργων Τρίτων

Το Έργο αυτό κάνει χρήση των ακόλουθων έργων:

Εικόνες/Σχήματα/Διαγράμματα/Φωτογραφίες

Εικόνες/σχήματα/διαγράμματα/φωτογραφίες που περιέχονται σε αυτό το αρχείο προέρχονται από το βιβλίο:

Τίτλος: *Discrete-time Speech Signal Processing: Principles and Practice*

Prentice-Hall signal processing series, ISSN 1050-2769

Συγγραφέας: Thomas F. Quatieri

Εκδότης: Prentice Hall PTR, 2002

ISBN: 013242942X, 9780132429429

Μέγεθος: 781 σελίδες

και αναπαράγονται μετά από άδεια του εκδότη.