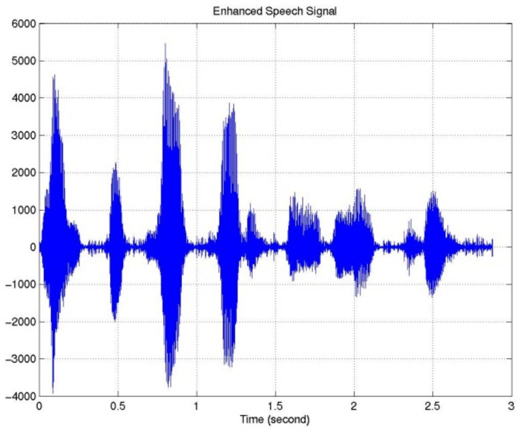
Speech Signal Basics



Nimrod Peleg

Updated: Feb. 2010

Course Objectives

- To get familiar with:
  - Speech coding in general
  - Speech coding for communication (military, cellular)
- Means:
  - Introduction the basics of speech processing
  - Presenting an overview of speech production and hearing systems
  - Focusing on speech coding: LPC codecs:
    - Basic principles
    - Discussing of related standards
    - Listening and reviewing different codecs.

הנחיות לשימוש נכון בטלפון,

פלשתינה-א"י, 1925

פ. ת. 144

הנהלת הדאר, הטלגרף והטלפון של פלשתינה (א"י)

כיצד משתמשים בטלפון האוטומתי.

כדי לקבל מספר טלפון בירושלים או בבית לחם. הרם את השפופרת והקשב לקול המסמן שיש להתחיל בסבוב. הכנס אצבע בחור המראה את הספרה הראשונה של המספר הדרוש לך, סובב את החוגה עד שהאצבע תגיע למעצור ואח"כ הוצא את האצבע כדי שהחוגה תחזור למקומה. עשה כן בכל ספרה של המספר המופיע במדריך הטלפון.

כדי להזמין שיחת־חוץ, סובב את הספרה „19" ומסור לטלפוניסט את שם מרכז הטלפונים שלך ואת מספרך יחד עם שם המרכז הדרוש לך ומספר הטלפון.

במקרה שתרצה לפנות בשאלה או להודיע על איזה קושי שהוא בשרות, סובב את המספרה „0".

כדי להודיע על קלקול, סובב את הספרה „16". יש לסובב ספרה זו בכדי למסור על שבירת מכשירים, קלקול פעמונים, חוטים וענפים ובמקרה שתנאי השמיעה יהיו לקויים או שלא יהא אפשר לקבל את המספרים הנכונים על ידי סבוב החוגה.

באור הקולות האוטומתיים

שם-הטון	הקול	באורו
קול הסבוב	קול נהימה דקה ממושך פררררררררררררררררר	יש להתחיל בסבוב החוגה.
קול הצלצול	.בררר בררר — בררר בררר	המספר שהוזמן מקבל צלצול
קול מספר תפוס	קול רם נשמע לסרוגין זזזז זזזז זזזז זזזזז	המספר שהוזמן תפוס
קול שאין להשיג את המספר	קול רם ממושך זזזזזזזזזזזזזזזזז	המספר שהוזמן הוא מקולקל, מנותק או ששום מנוי לא מחובר אליו.

Subscribers Instruction Card — Jerusalem and Bethlehem

30579—1800—17/2/38—Com. P.

What is Speech ???

- Meaning: Text, Identity, Punctuation, Emotion
  - prosody : rhythm, pitch, intensity
- Statistics: sampled speech is a string of numbers
  - Distribution (Gamma, Laplace) Quantization
- Information Theory: statistical redundancy
  - ZIP, ARJ, compress, Shorten, Entropy coding…

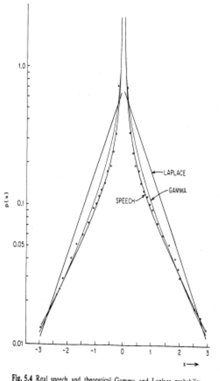


Fig. 5.4 Real speech and theoretical Gamma and Laplace probability densities. (After Paez and Glisson [3].)

...1001100010100110010...

- Perceptual redundancy
  - Temporal masking , frequency masking (mp3, Dolby)
- Speech production Model
  - Physical system: Linear Prediction

The Speech Signal

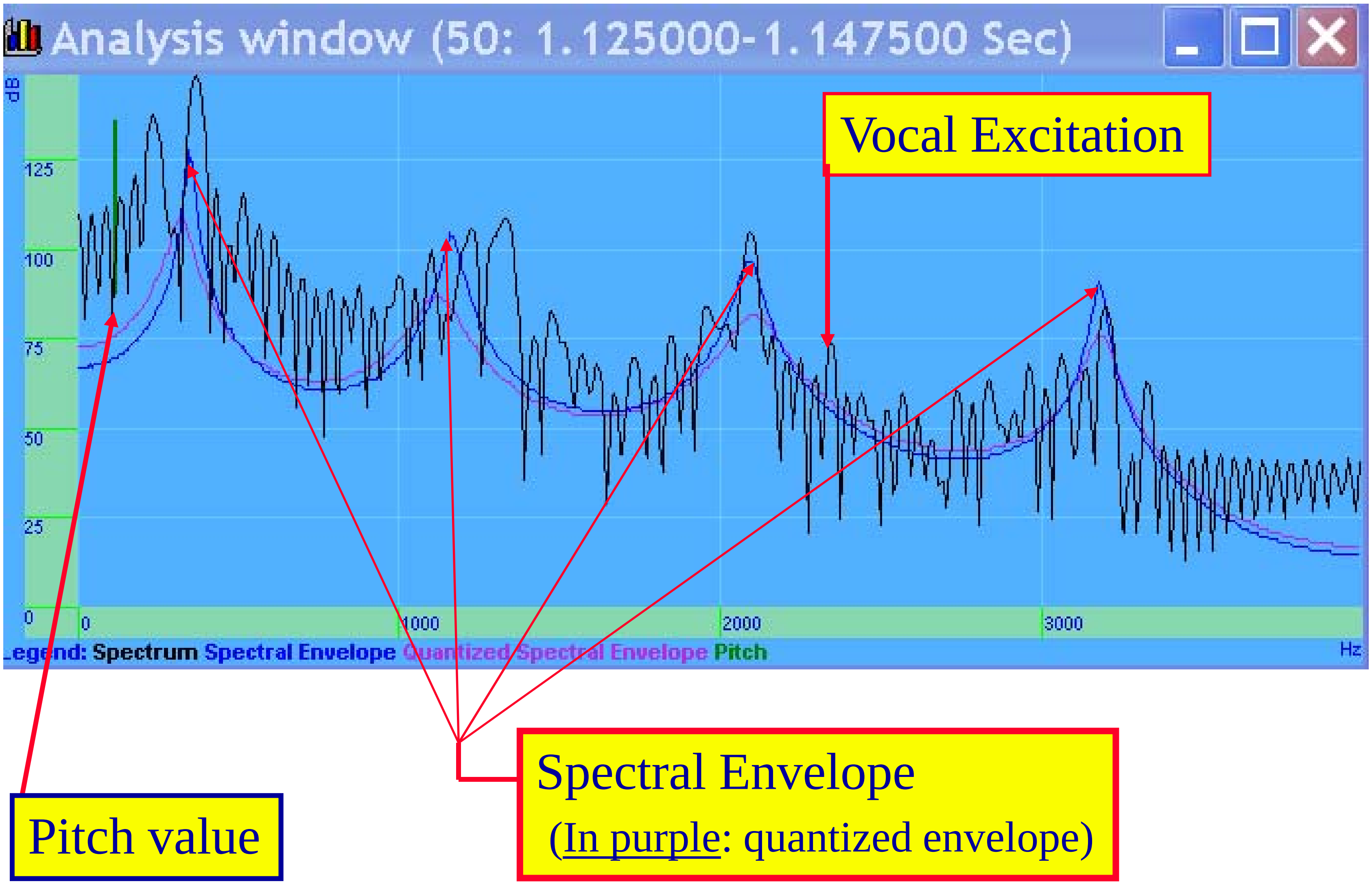
- Created at the Vocal cords, Travels through the Vocal tract, and Produced at speakers mouth
- Gets to the listeners ear as a pressure wave
- Non-Stationary, but can be divided to sound segments which have some common acoustic properties for a short time interval
- Two Major classes: *Vowels* and *Consonants*

Speech Production

- A sound source excites a (vocal tract) filter
  - *Voiced*: Periodic source, created by vocal cords
  - *UnVoiced*: Aperiodic and noisy source

- The *Pitch* is the fundamental frequency of the vocal cords vibration (also called F_0) followed by 4-5 *Formants* (F_1 - F_5) at higher frequencies.

Spectral look of “Oooooohhhh”



Schematic Diagram of Vocal Mechanism

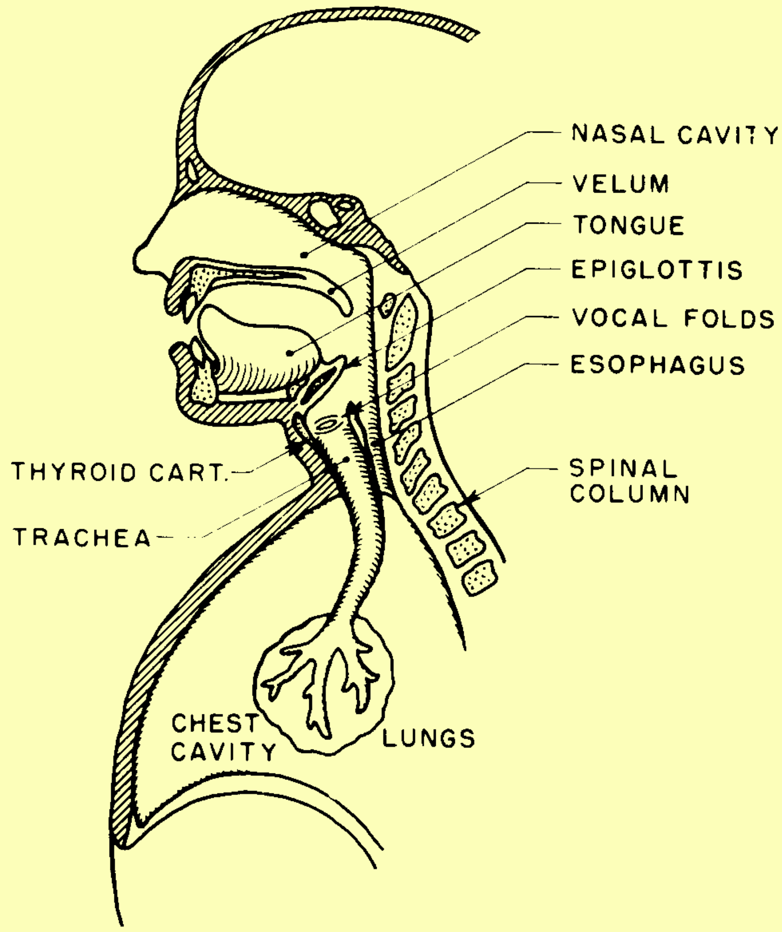
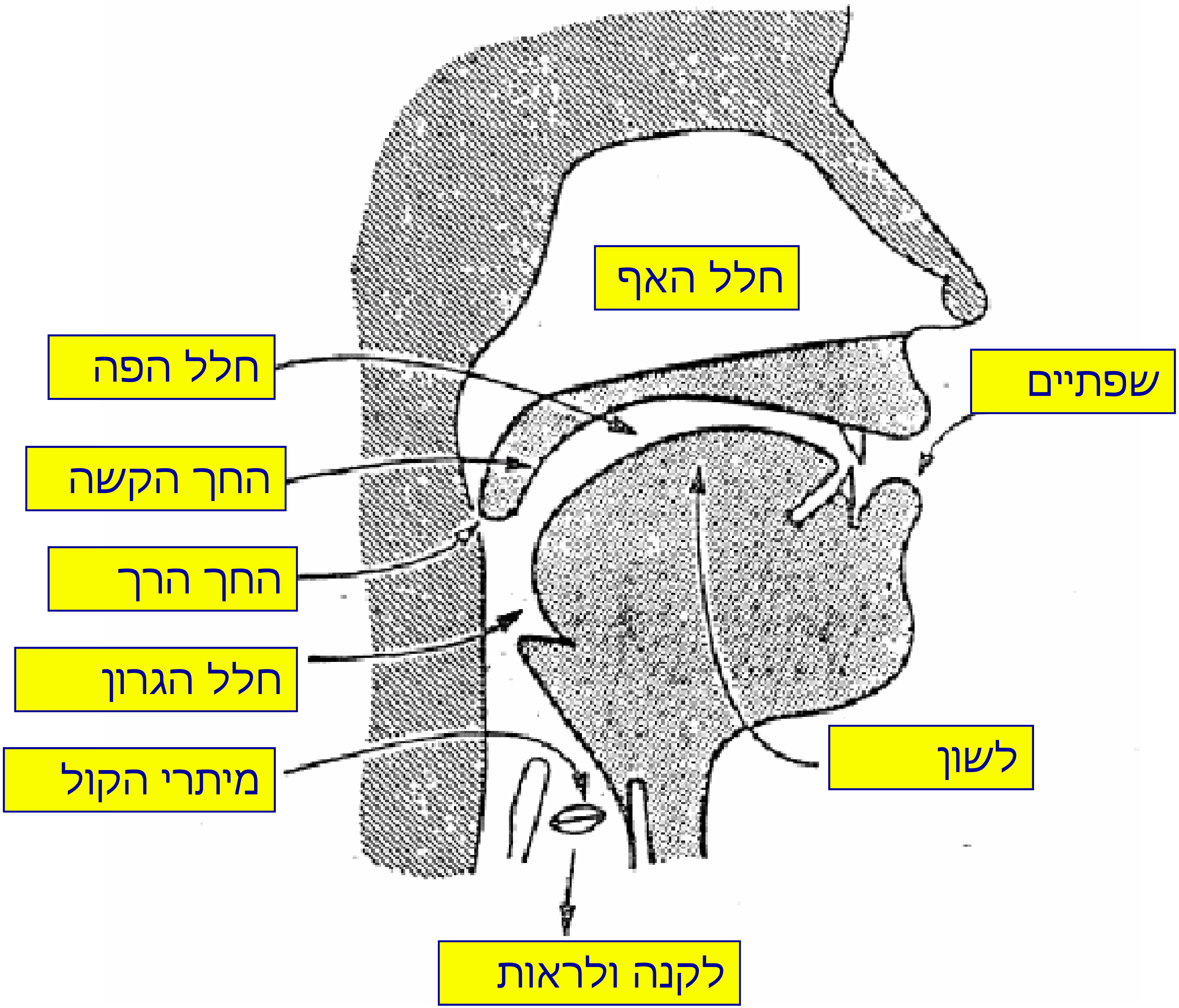
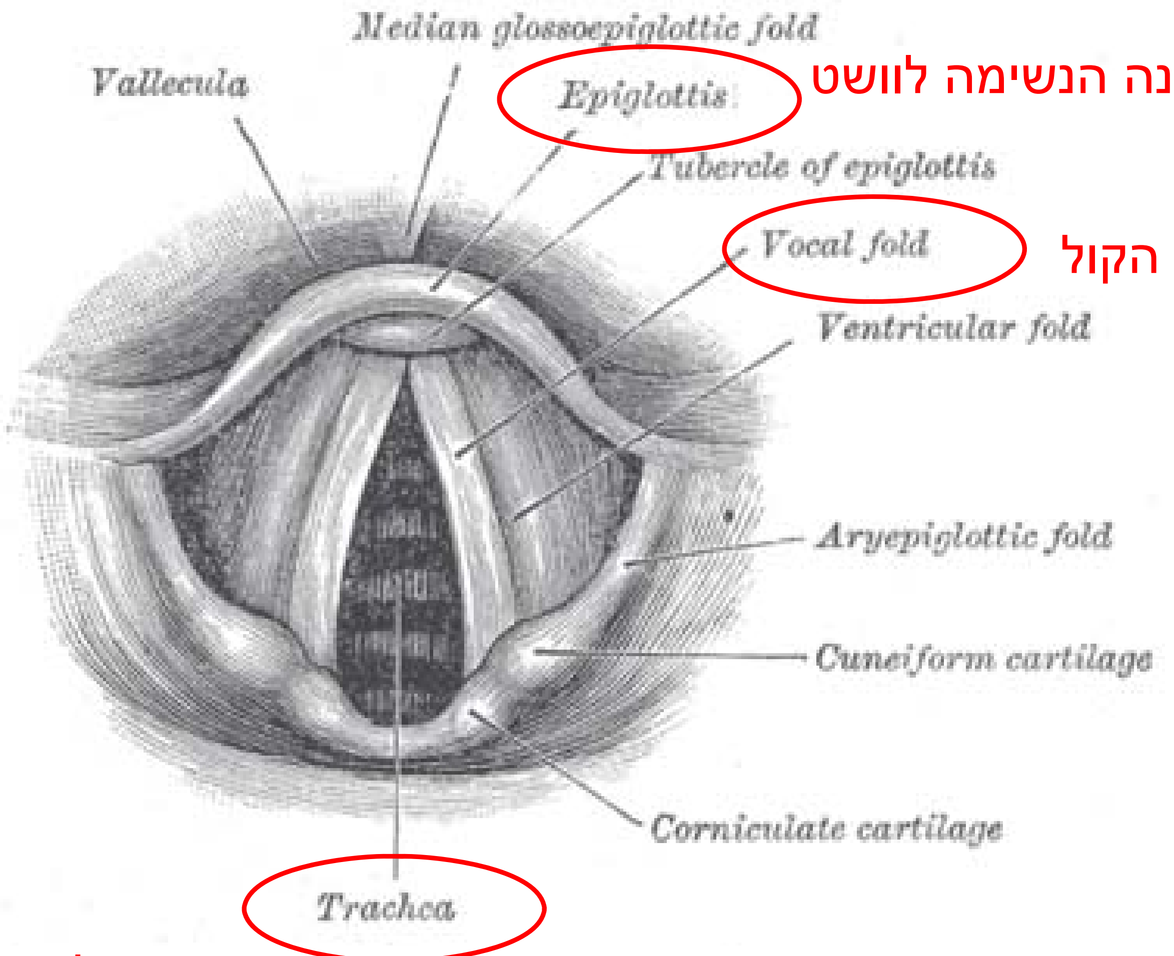


Figure 2.4 Schematic view of the human vocal mechanism (after Flanagan [3]).



חלל האף
חלל הפה
שפתיים
החך הקשה
החך הרך
חלל הגרון
מיתרי הקול
לשון
לקנה ולראות

The Vocal Cords



Human voice [Wikipedia]

Glottal Volume and Mouth Sound

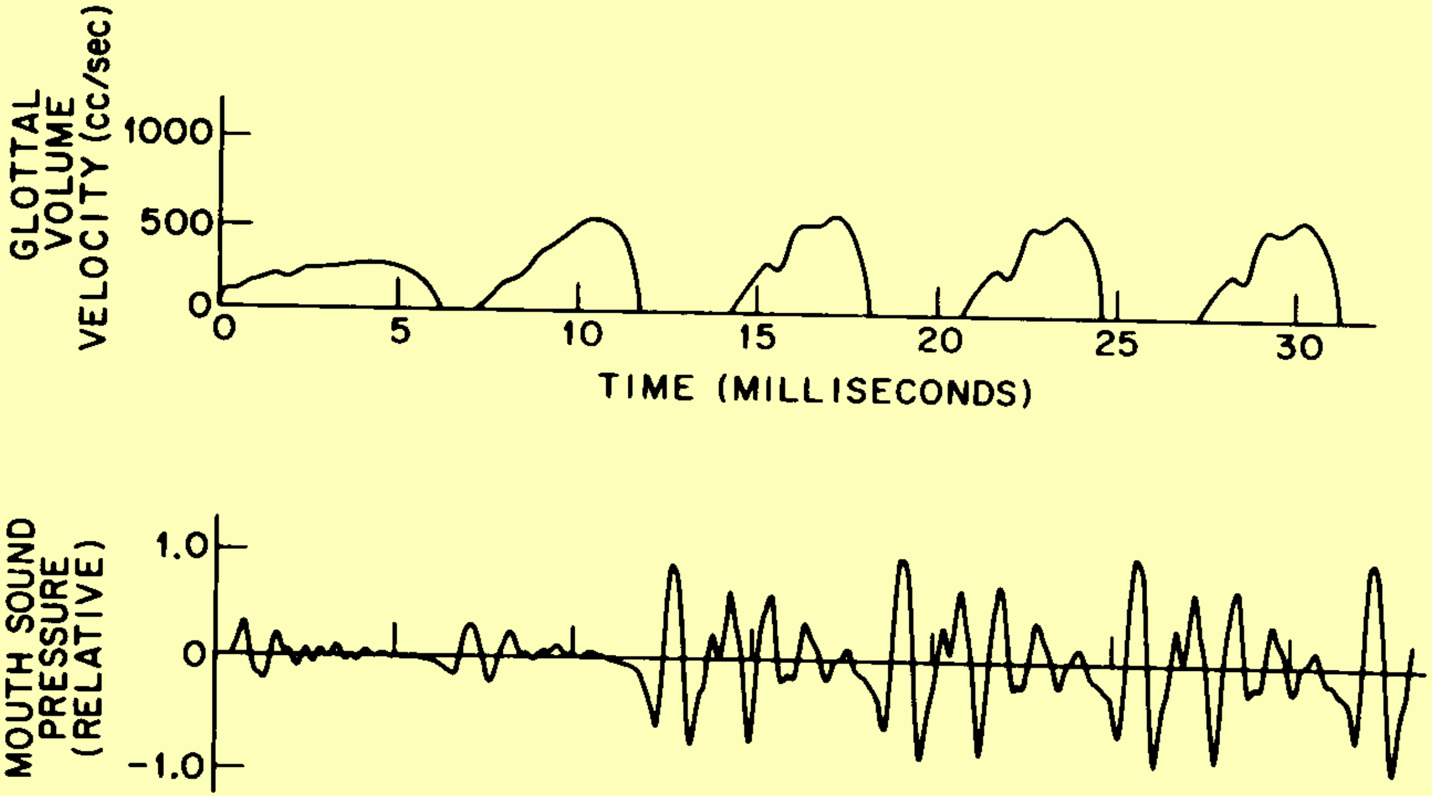


Figure 2.5 Glottal volume velocity and resulting sound pressure at the start of a voiced sound (after Ishizaka and Flanagan [4]).

Pipelines Model

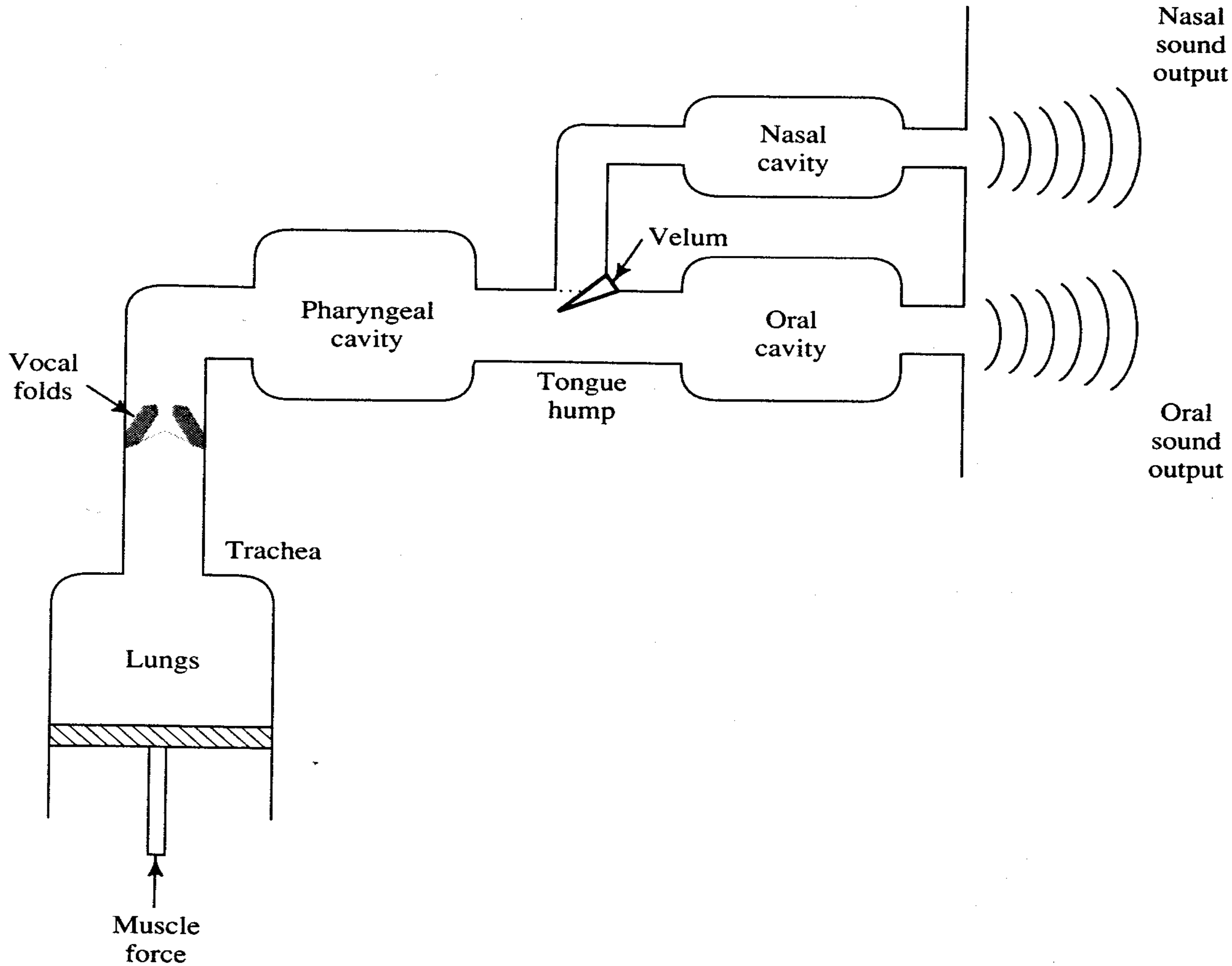
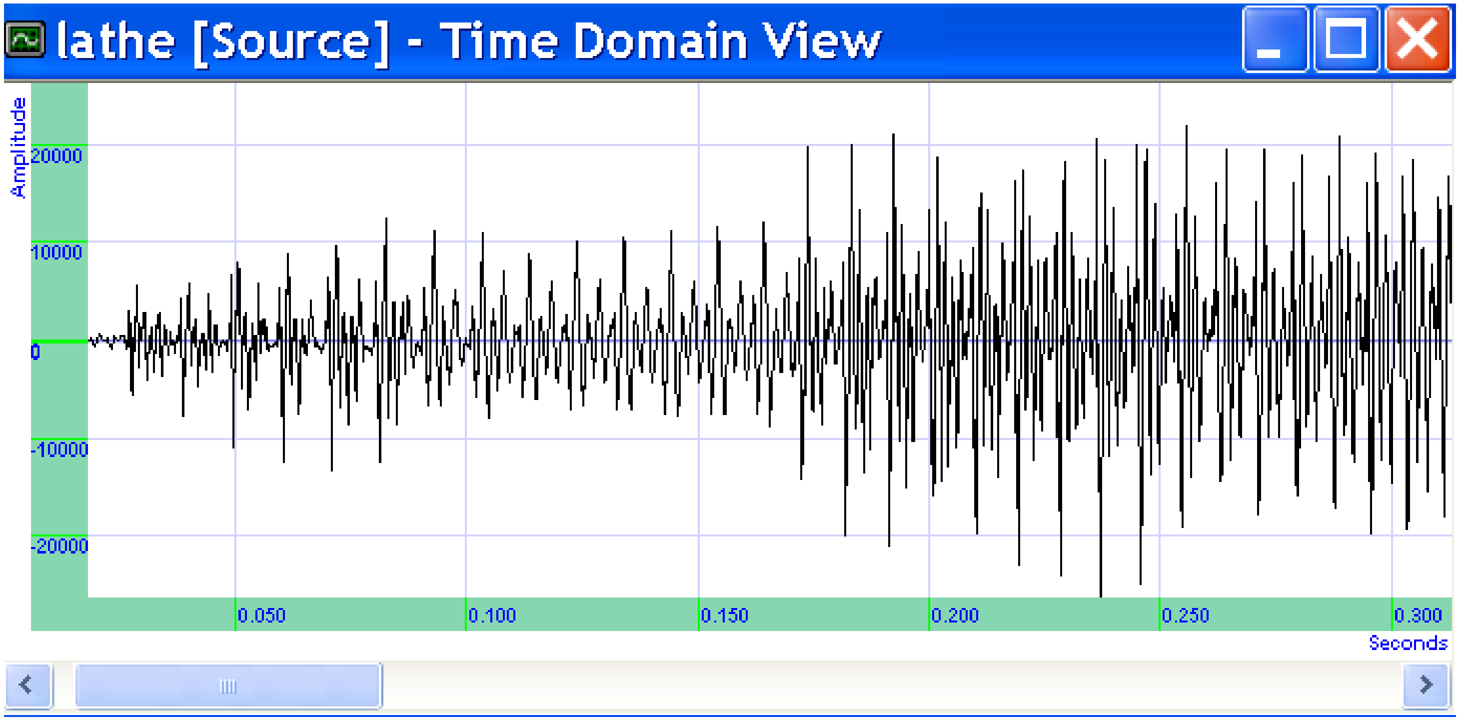


FIGURE 2.2. A block diagram of human speech production.

Typical Voiced Sound

1Sec, 10,000 Samples, 8bps, Voiced ("Ahhhhh")



A Quasi-Periodic Signal

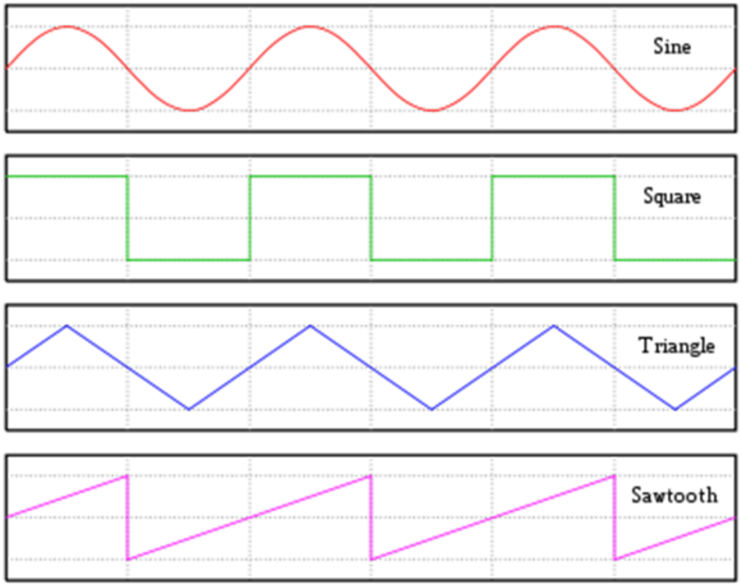
Typical Pitch

- Speech: male ~ 85-155 Hz;
  female ~ 165-255 Hz;

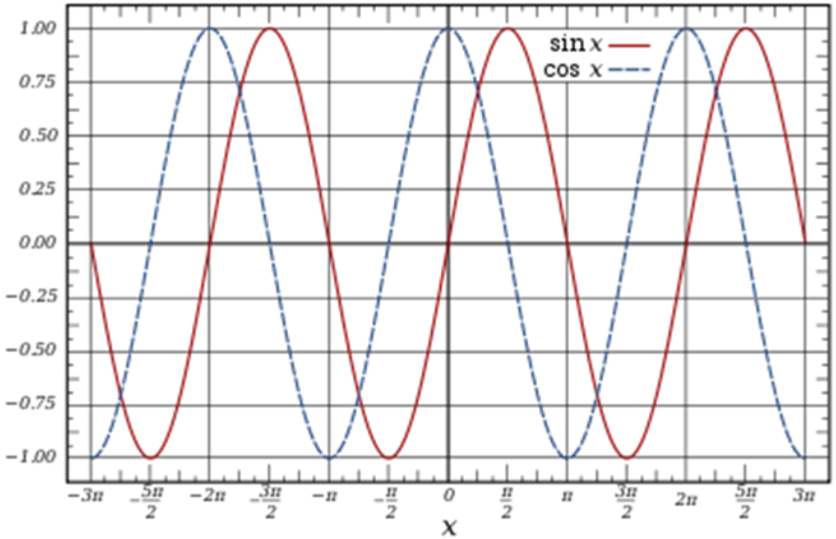
- **Note the overlap !**

- Singer's vocal range:
  from bass to soprano: 80 Hz-1100 Hz

Waves: Sine/Cosine



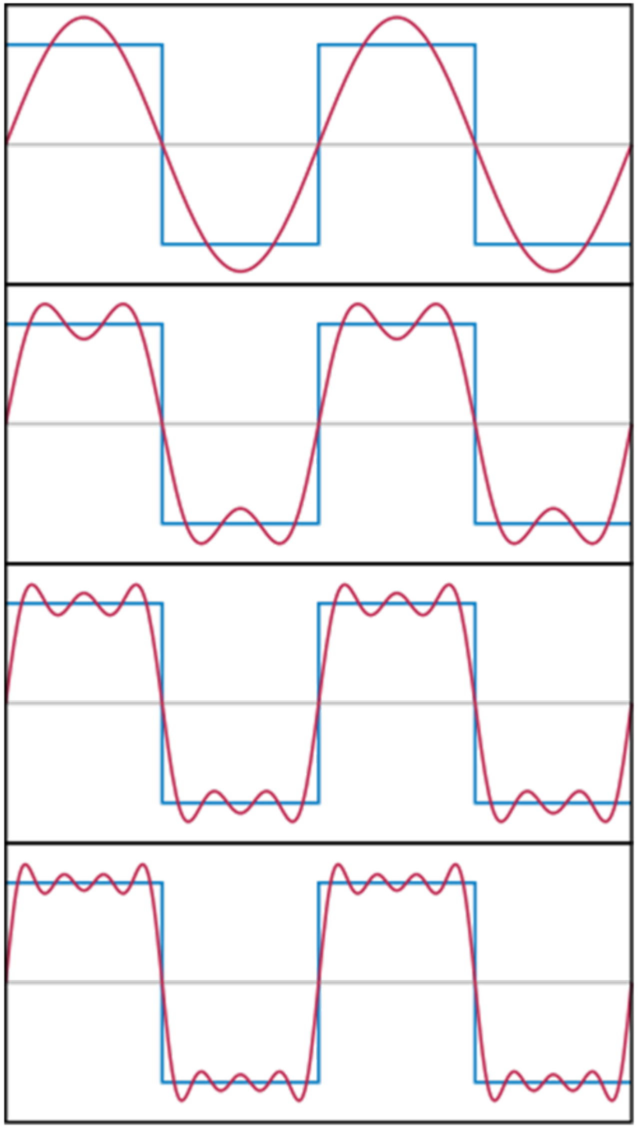
- The graphs of the sine and cosine functions are sinusoids of different phases



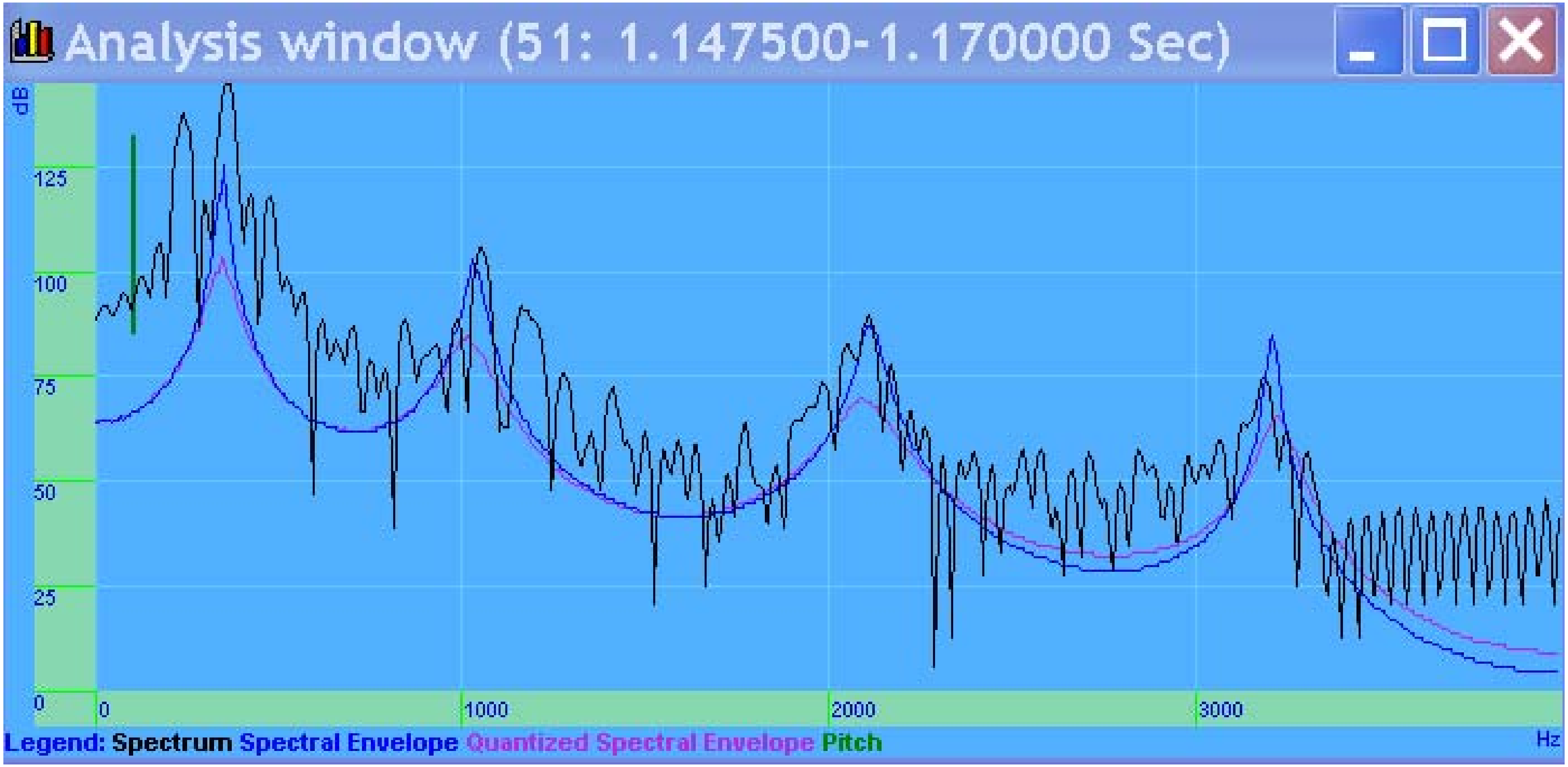
Frequency Analysis

- The first four Fourier series approximations for a square wave

- Analysis performed by Fourier Transform



Power Spectrum: Voiced Speech



Vowel Production

- In vowel production, air is forced from the lungs by contraction of the muscles around the lung cavity

- Air flows through the vocal cords, which are two masses of flesh, causing periodic vibration of the cords whose rate gives the pitch of the sound.

- The resulting periodic puffs of air act as an excitation input, or source, to the vocal tract.

Typical Vowels

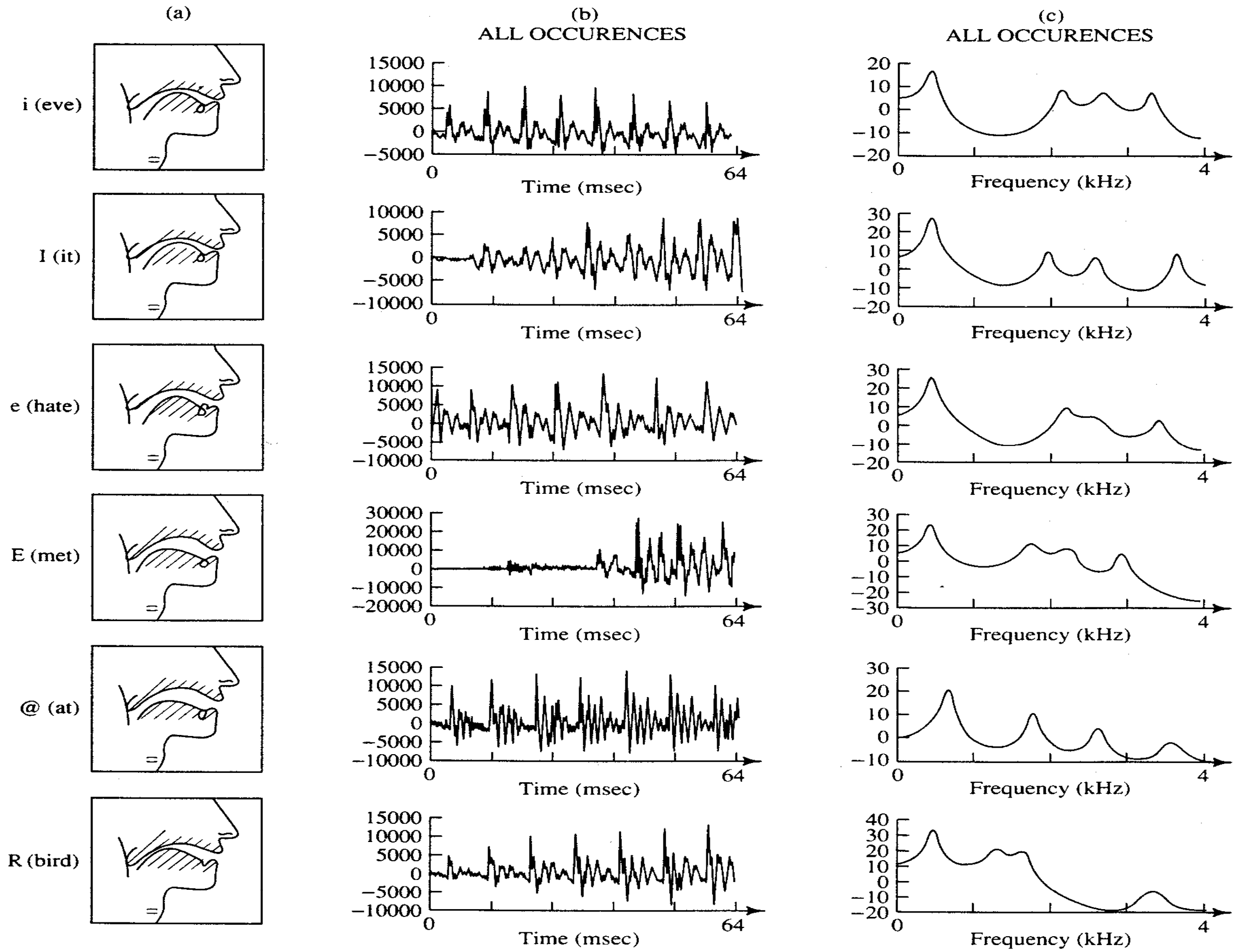


FIGURE 2.10. A collection of features for vowels in American English. Column (a) represents schematic vocal-tract profiles, (b) typical acoustic waveforms, and (c) the corresponding vocal-tract magnitude spectrum for each vowel.

Average Formant Locations

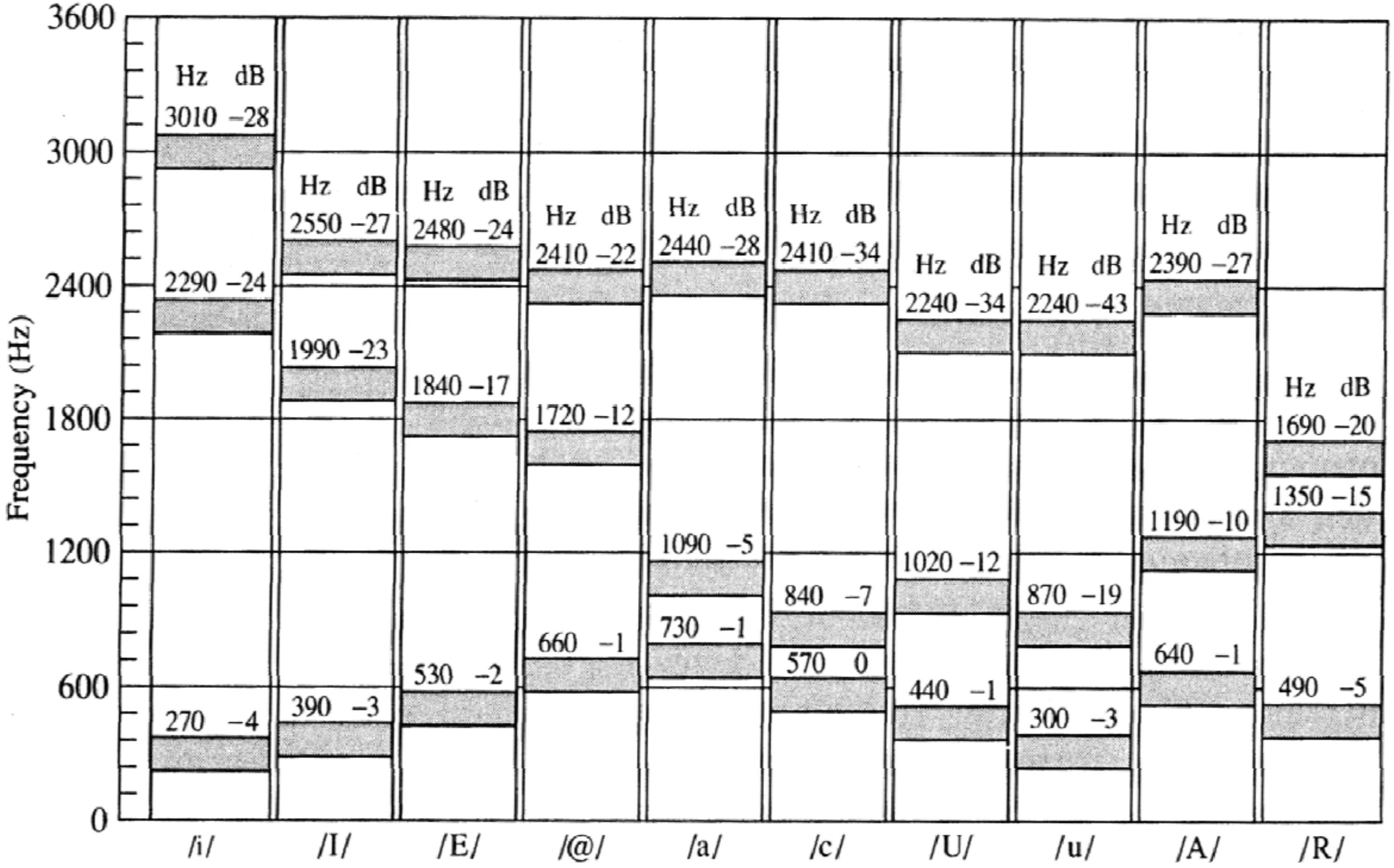


FIGURE 2.11. Average formant locations for vowels in American English (Peterson and Barney, 1952).

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	ɝ	(bird)	490	1350	1690

IPA	Worldbet	Example
iː	i:	beet
ɪ	I	bit
ɛ	E	bet
æ	@	bat
ʌ	^	above
u	u	boot
ʊ	U	book
ə	&	above
ɑ	A	father
ɝ	3r	bird
ɚ	&r	butter

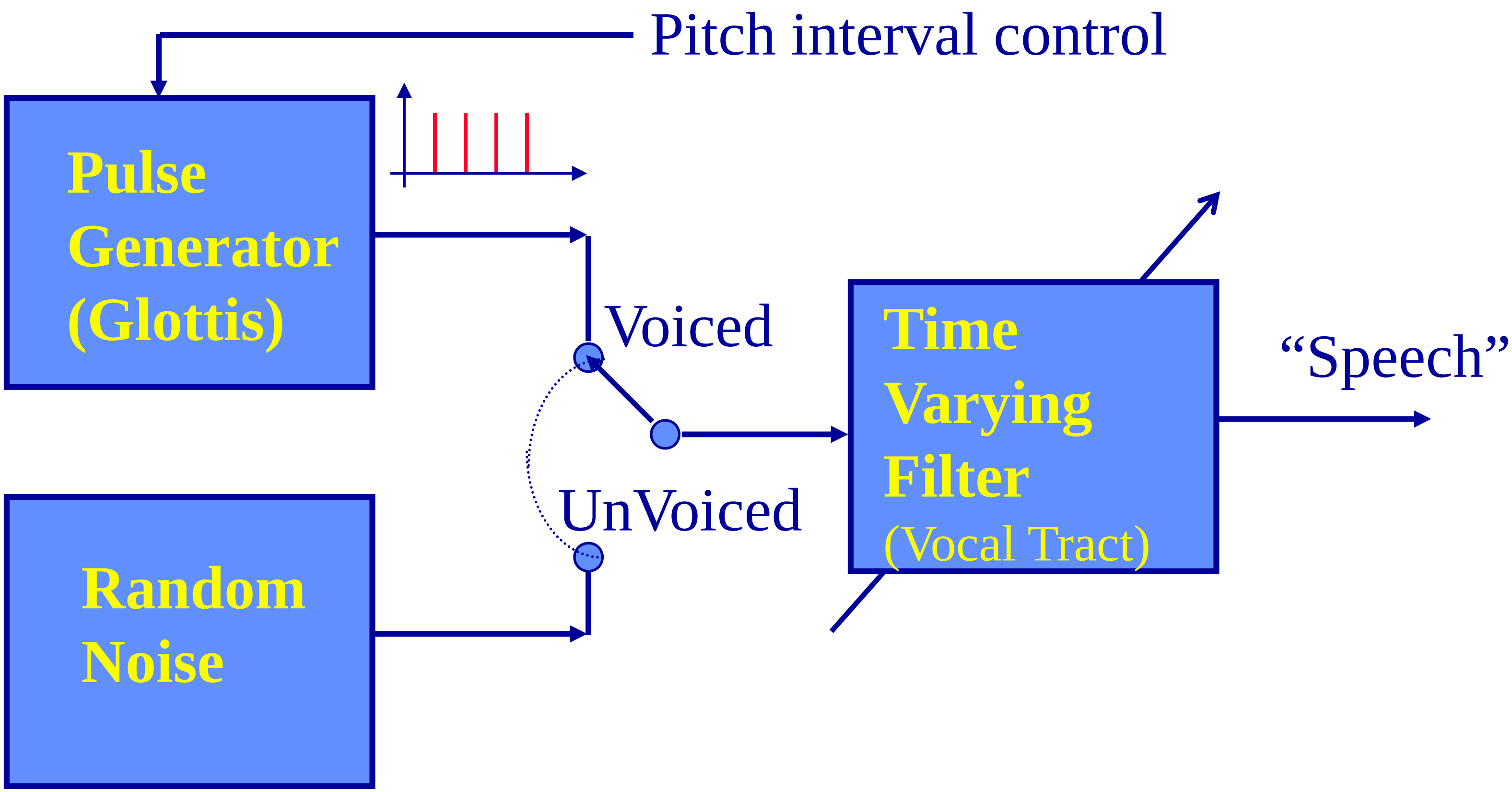
From: Rabiner & Schafer, Digital Processing of Speech Signals

The Vocal Tract

- The vocal tract is the cavity between the vocal cords and the lips, and acts as a resonator that **spectrally shapes the periodic input**, much like the cavity of a musical wind instrument.

- A simple model of a steady-state vowel regards the vocal tract as a linear time-invariant (LTI) filter with a periodic **impulse-like input**.

Speech Production Model



Phonemes

- The basic sounds of a language (e.g. "a" in the word "father“) are called *phonemes*.
- A typical speech utterance consists of a string of vowel and consonant phonemes whose temporal and spectral characteristics **change with time** .
- In addition, the time-varying source and system can also *nonlinearly* interact in a complex way:

our simple model is correct for a **steady vowel**,

but the sounds of speech are not always well represented by linear time-invariant systems !

Speech Sounds Categories

- *Periodic* (Sonorants, Voiced: קולי)
- *Noisy* (Fricatives , Un-Voiced: א-קולי)
- *Impulsive* (Plosive: פוצץ)

- Example:

In the word “shop,” the “sh,” “o,” and “p” are generated from a noisy, periodic, and impulsive source, respectively.

All you need to know about phonemes:

CENTER *for* SPOKEN LANGUAGE UNDERSTANDING @ OGI

speech dialog
synthesis•recognition•enhancement
facial animation
SIGNAL PROCESSING

OREGON HEALTH & SCIENCE UNIVERSITY

OGI SCHOOL OF SCIENCE & ENGINEERING

http://cslu.cse.ogi.edu/tutordemos/SpectrogramReading/spectrogram_reading.html

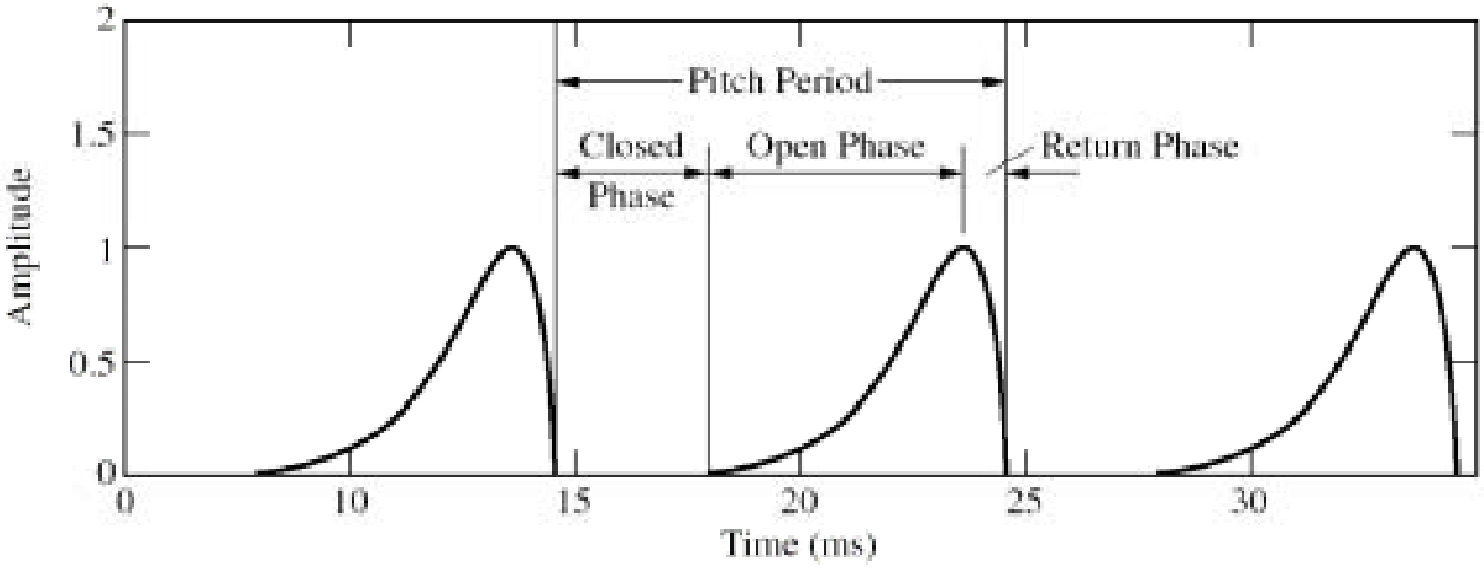
The Pitch

- Pitch period:

The time duration of one glottal cycle.

- Pitch (fundamental frequency):

The reciprocal of the pitch period.



Typical Pitch & Intensity Variations

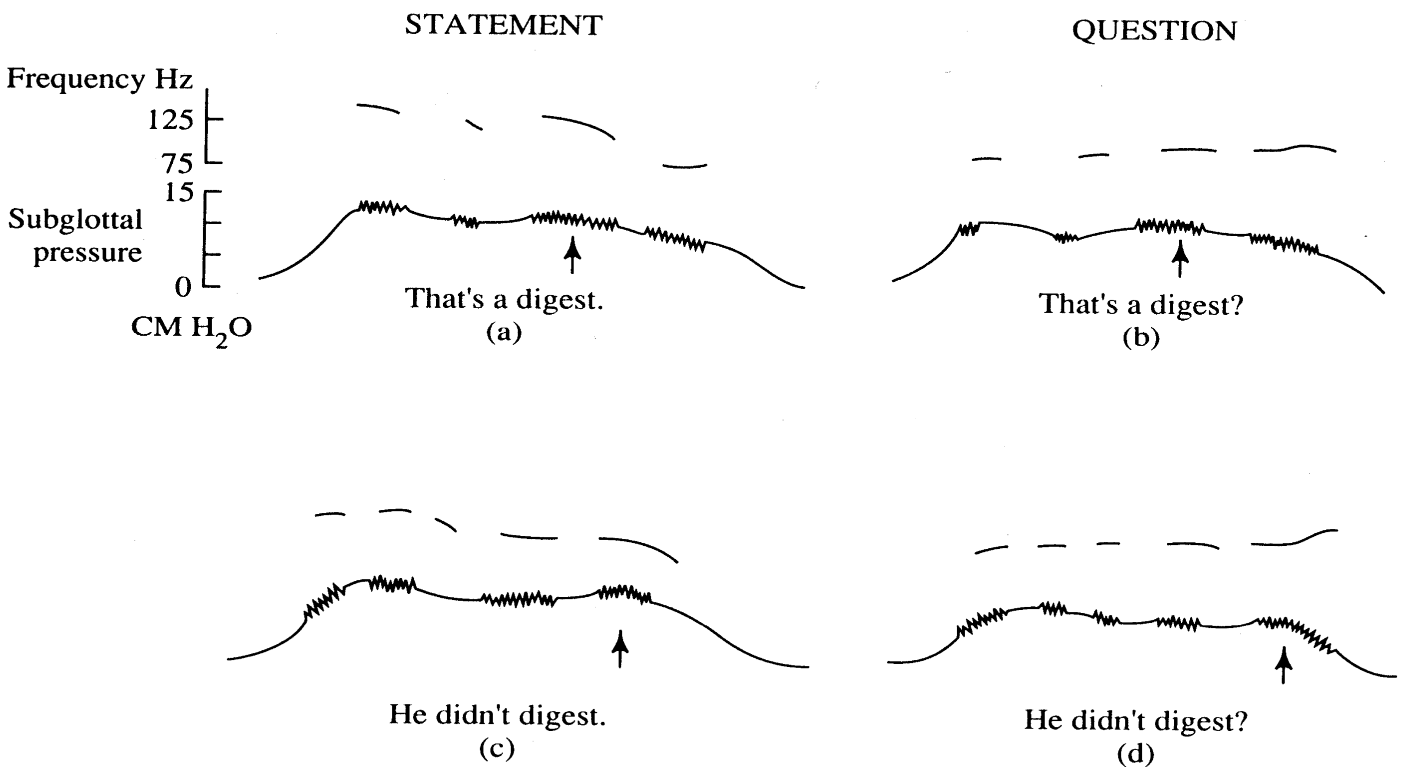


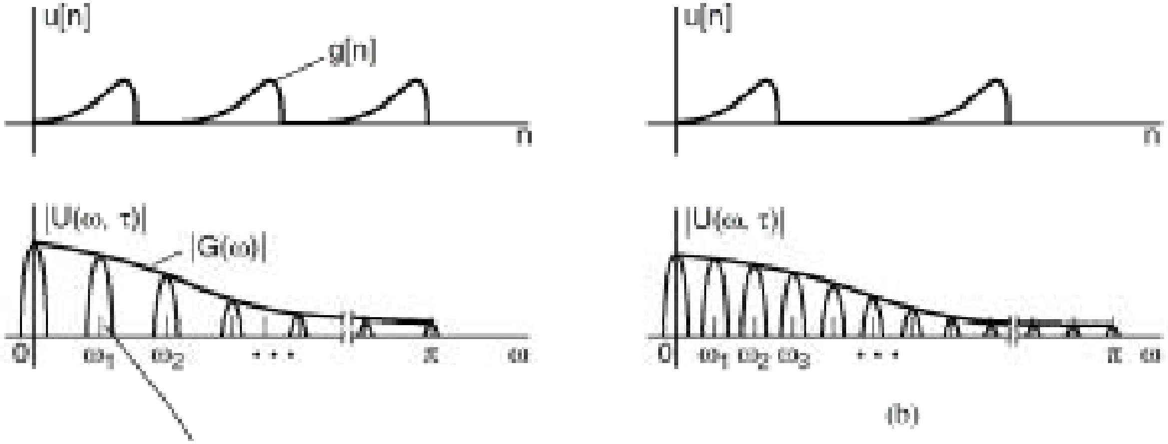
FIGURE 2.21. Relations between fundamental frequency contours and subglottal air pressure for statements and questions with two different word stress patterns. Adapted from Ladefoged (1963).

Harmonics

- The frequencies $\omega = \frac{2\pi}{p} k$ are referred to as the *harmonics* of the glottal waveform

($\frac{2\pi}{p}$ is the fundamental frequency, pitch)

As the pitch period *P* decreases, the spacing between the harmonics increases:



Pitch Detection

- The pitch period and V/UV decisions are elementary to many speech coders

- Many methods for the calculation:
  - Autocorrelation function

Pitch Calculation utility

SPDemo demonstration NOW

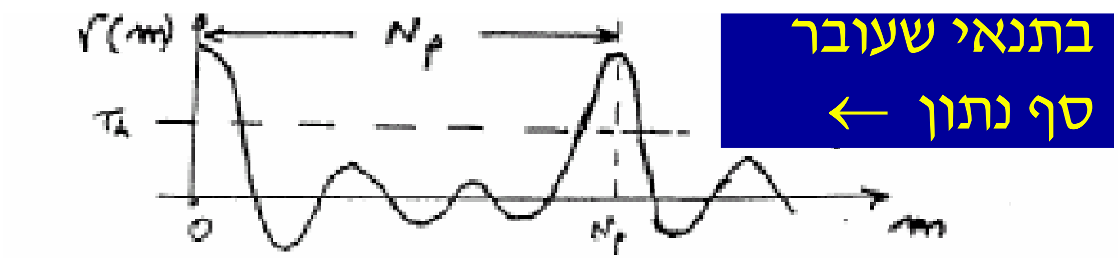
SP DEMO

Speech and Audio Processing Learning Tool

Autocorrelation Pitch Detector

- Calculate the autocorrelation function of the signal within estimated range
  - For speech signal, sampled at 8KHz, the range in samples varies between 20-130 (~2.5-16mSec)
  - Mathematical definition:

$$r(m) = \sum_{i=-\infty}^{\infty} x(i)x(i+m)$$



בתנאי שעובר
סף נתון ←

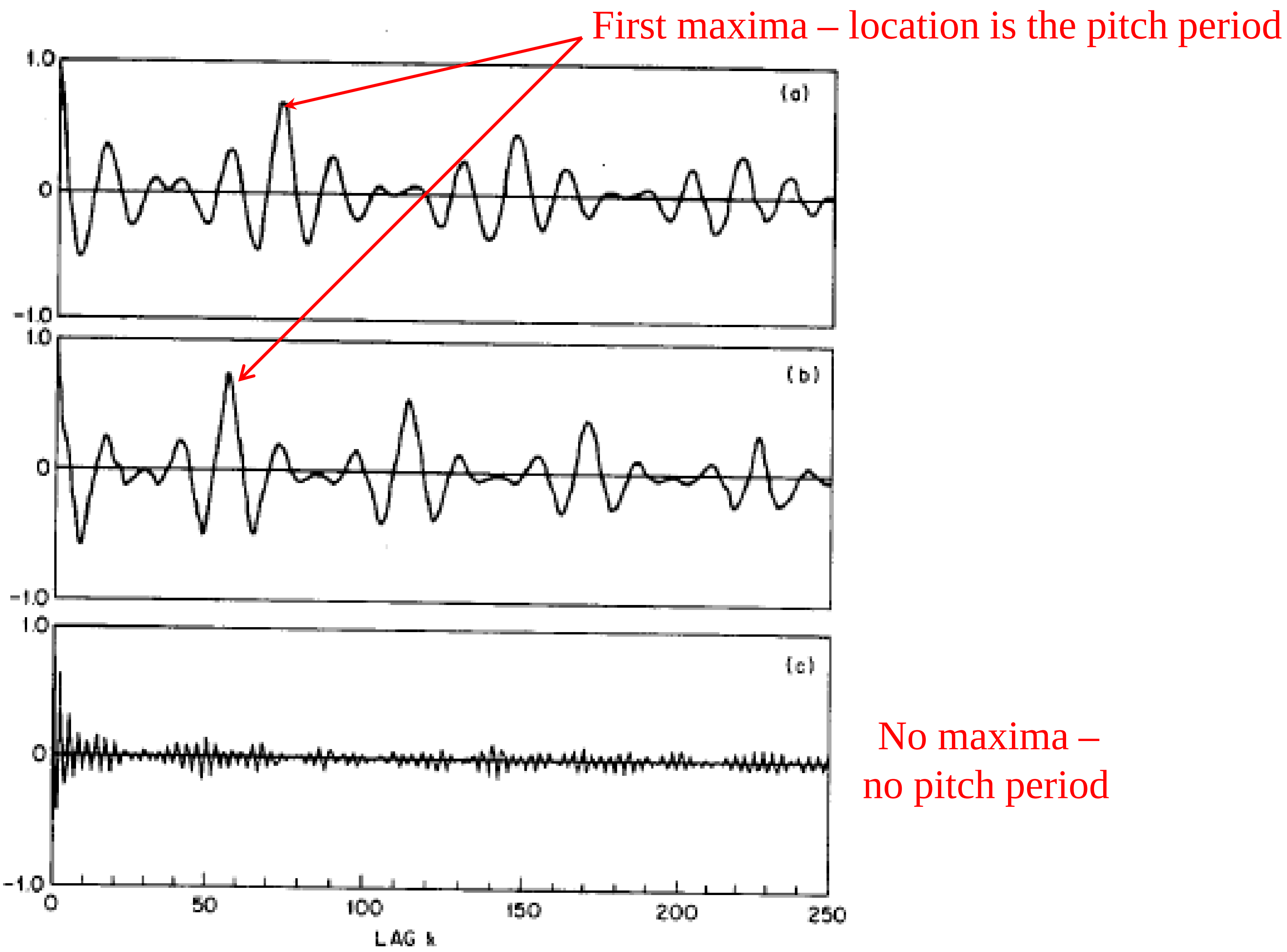


Fig. 4.24 Autocorrelation function for (a) and (b) voiced speech; and (c) unvoiced speech, using a rectangular window with $N = 401$.

The Vocal Tract and Formants

- The relation between a glottal airflow velocity *input* and vocal tract airflow velocity *output* can be approximated by a **linear filter** with resonances called Formants. (like resonances of organ pipes and wind instruments)
- Formants change with different vocal tract configurations corresponding to different resonant cavities and thus different phonemes.

The Vocal Tract and Formants (cont'd)

- Generally, the frequencies of the formants decrease as the vocal tract length increases.
- Therefore, a male speaker tends to have lower formants than a female, and a female has lower formants than a child.
- Under a vocal tract linearity and time-invariance (LTI) assumption, and when the sound source occurs at the glottis, the speech waveform (i.e., the airflow velocity at the vocal tract output) can be expressed as the convolution of the glottal flow input and vocal tract impulse response.

Glottal source harmonics, Vocal tract formant and Spectral envelope

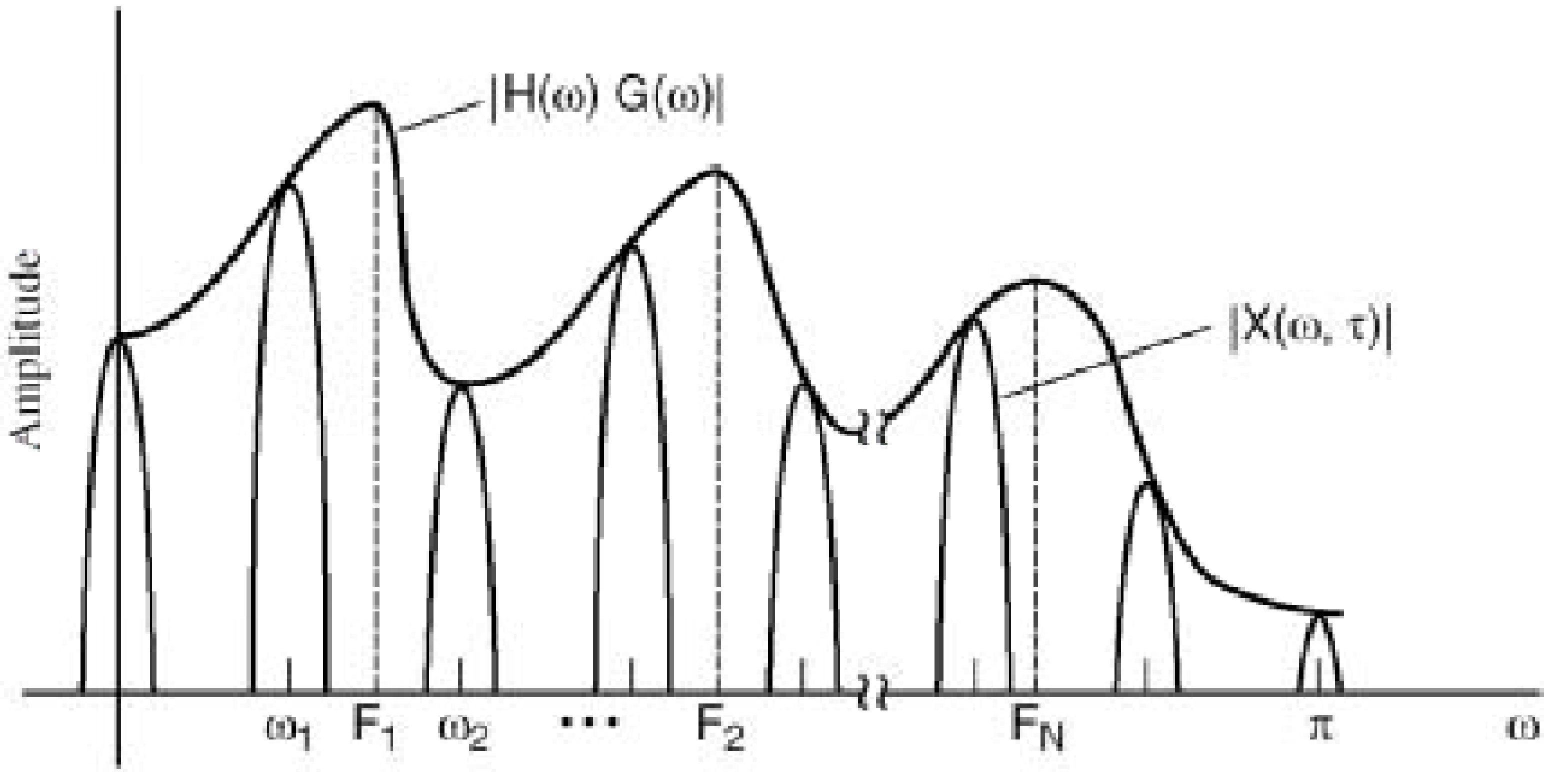


Figure 3.11 Illustration of relation of glottal source harmonics $\omega_1, \omega_2, \ldots \omega_N$, vocal tract formants $F_1, F_2, \ldots F_M$, and the spectral envelope $|H(\omega)G(\omega)|$.

Categorization of Speech Sounds

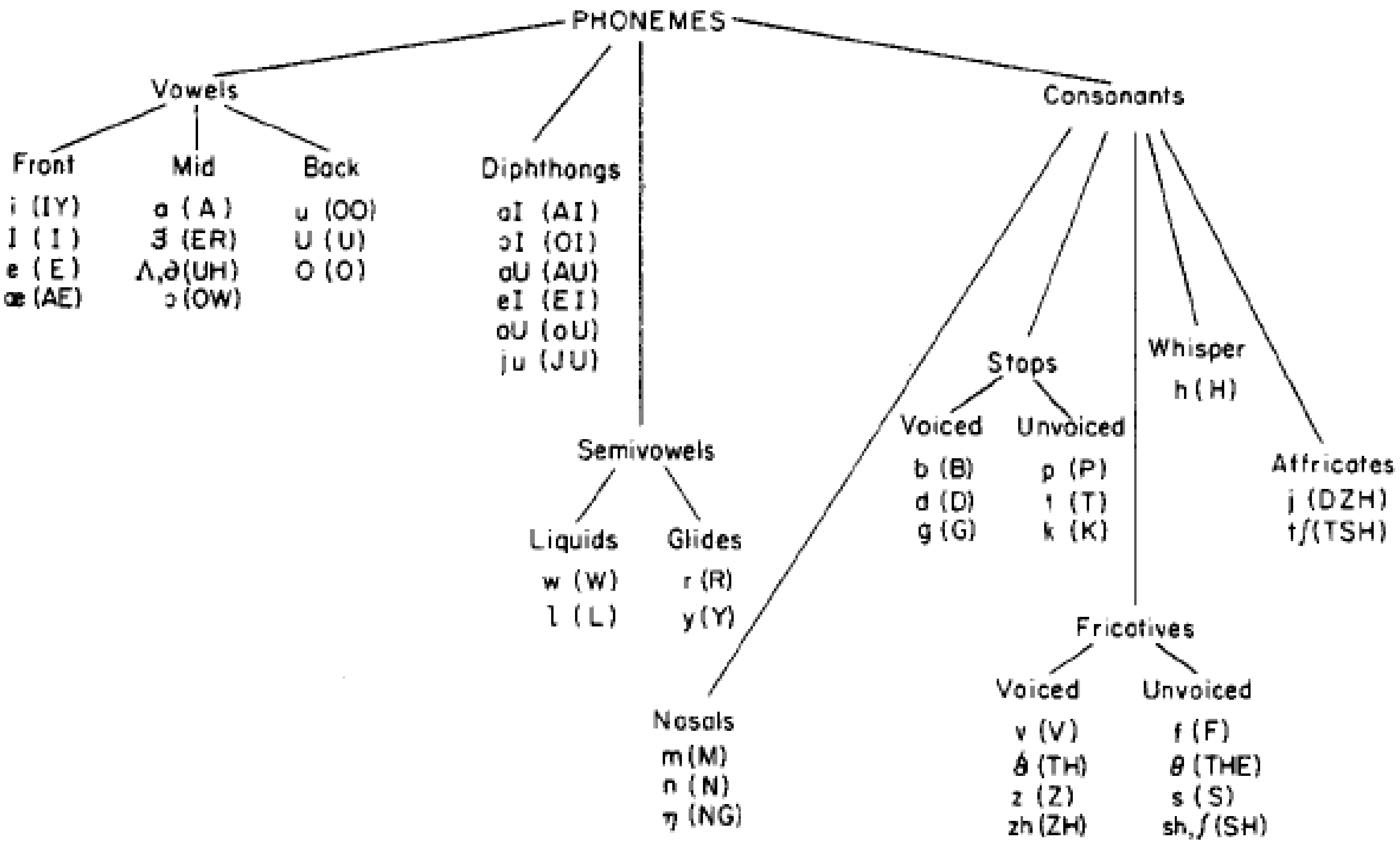
- Speech sounds are studied and classified from the following perspectives:

1) The nature of the source: periodic, noisy, or impulsive, and combinations of the three.

<u>More optional classes:</u>

2) The shape of the vocal tract.
3) The time-domain waveform, which gives the pressure change with time at the lips output.
4) The time-varying spectral characteristics revealed through the spectrogram.

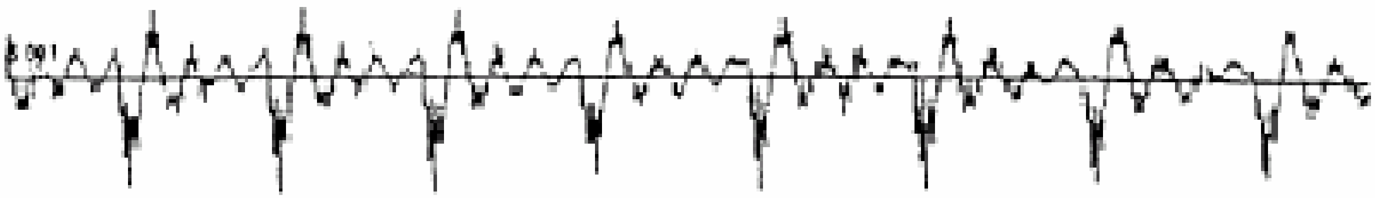
Table 3.1 Phonemes in American English.



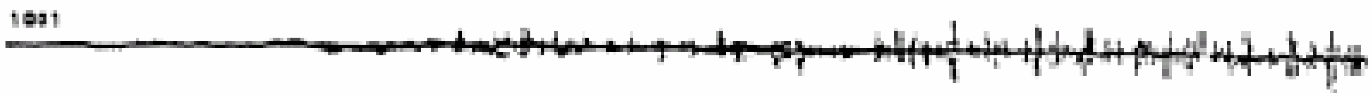
From: Rabiner & Schafer, Digital Processing of Speech Signals

Waveforms Examples

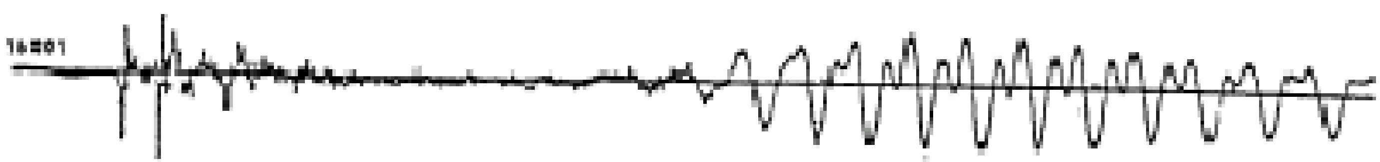
Voiced
(a,e,u,o,i)



Un-Voiced
(s,f,sh)



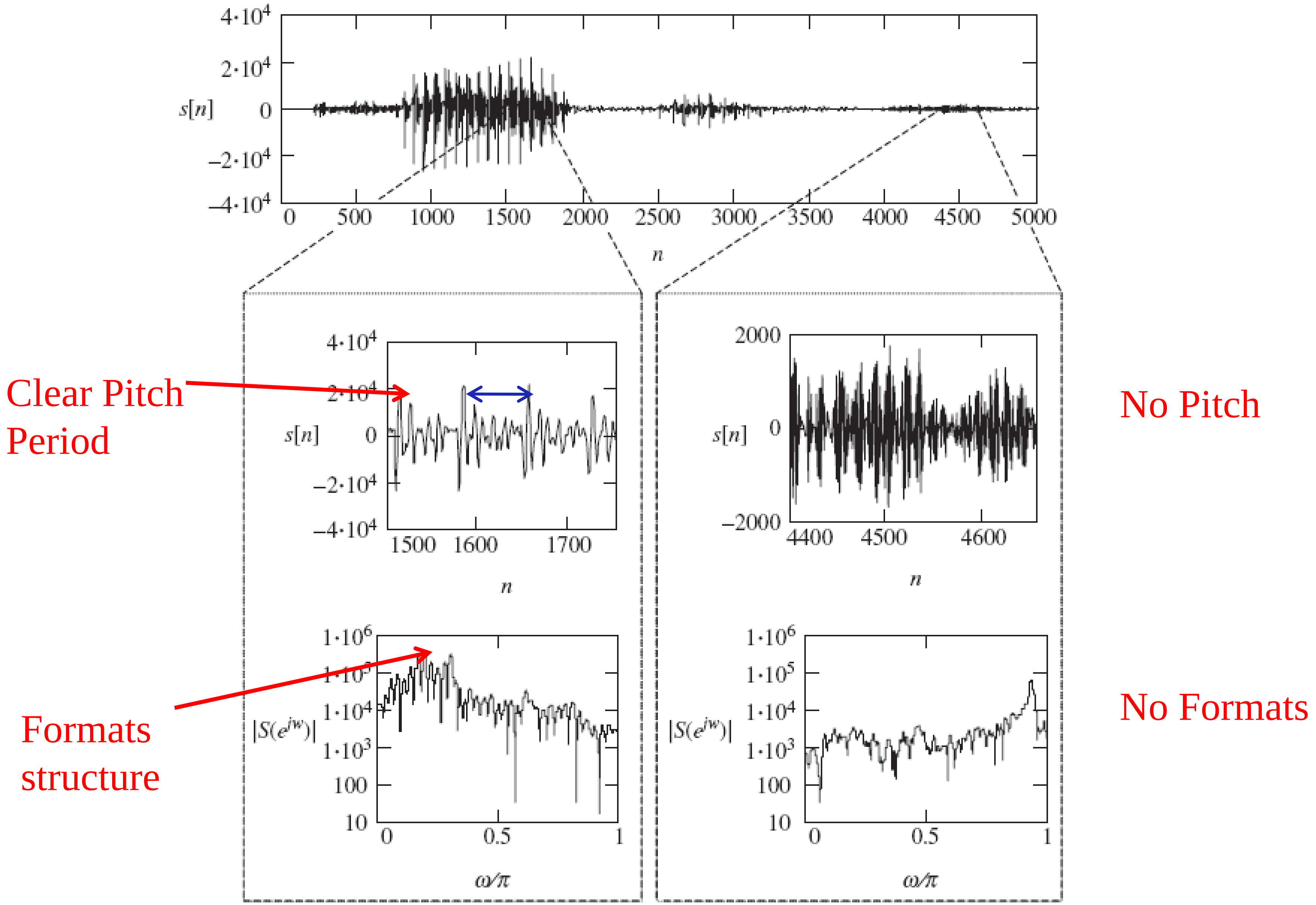
Plosive
(p,k,t)



האוויר משתחרר דרך הפה ולא דרך האף

Most common **Manner of articulation**

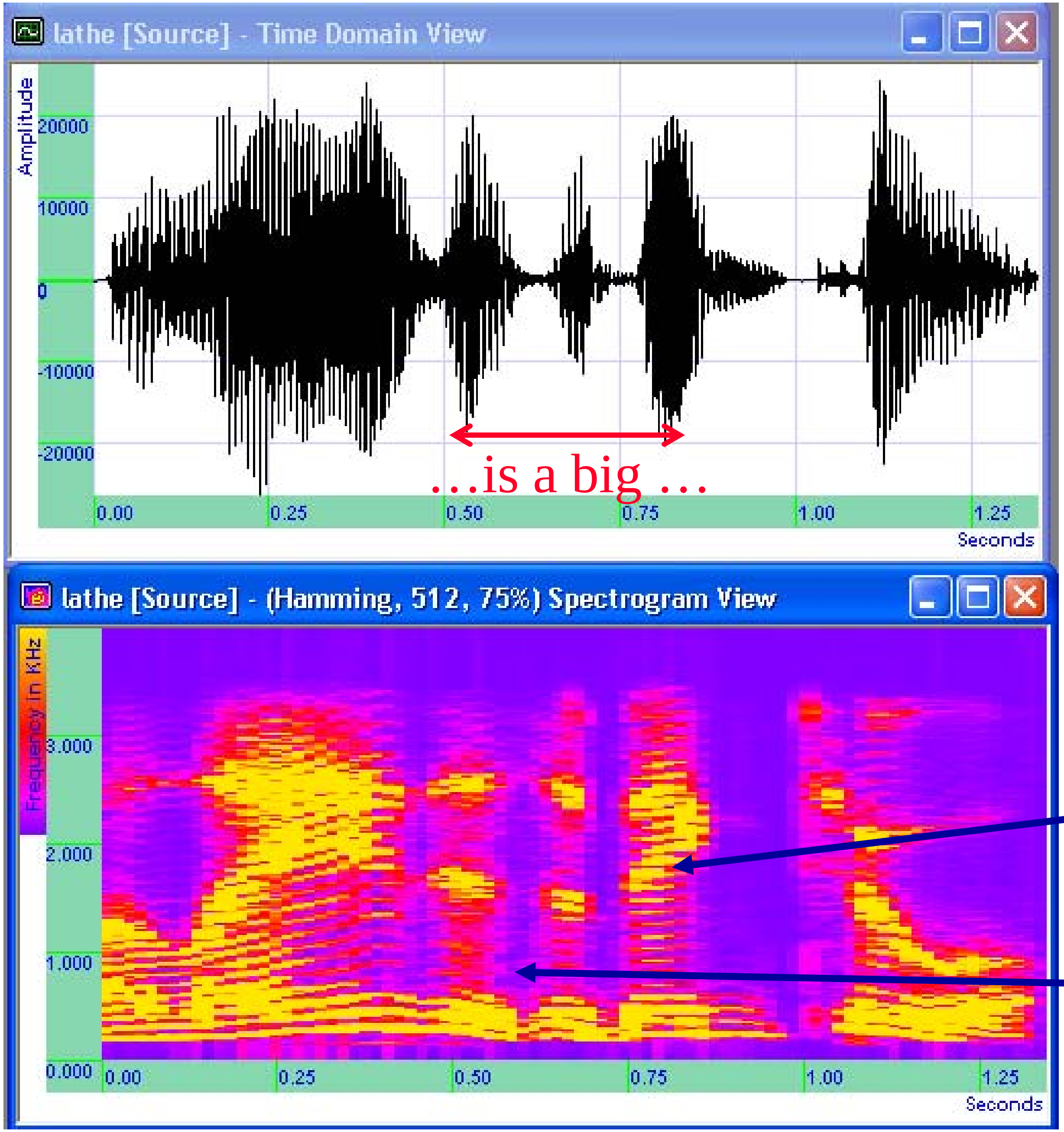
- **Plosive**, or **oral stop**, where there is complete *occlusion* (blockage) of both the oral and nasal cavities of the vocal tract, and therefore no air flow. Examples include English /p t k/ (voiceless) and /b d g/ (voiced). (פוצץ)
- **Nasal stop**, where there is complete occlusion of the oral cavity, and the air passes instead through the nose. The shape and position of the tongue determine the resonant cavity that gives different nasal stops their characteristic sounds. Examples include English /m, n/. (אפי)
- **Fricative**, sometimes called **spirant**, where there is continuous *frication* (turbulent and noisy airflow) at the place of articulation. Examples include English /f, s/ (voiceless), /v, z/ (voiced), etc. (חוכך א')
- **Sibilants** are a type of fricative where the airflow is guided by a groove in the tongue toward the teeth, creating a high-pitched and very distinctive sound. These are by far the most common fricatives. English sibilants include /s/ and /z/. (שורק)
- **Affricate**, which begins like a plosive, but this releases into a fricative rather than having a separate release of its own. The English letters "ch" and "j" represent affricates. (חוכך ב')
- **Trill**, in which the articulator (usually the tip of the tongue) is held in place, and the airstream causes it to vibrate. The double "r" of Spanish "perro" is a trill. (מסולסל)
- **Approximant**, where there is very little obstruction. Examples include English /w/ and /r/. **Lateral approximants**, usually shortened to **lateral**, are a type of approximant pronounced with the side of the tongue. English /l/ is a lateral.
- **And some more ….**



Example of speech waveform (male) of the word "problems."

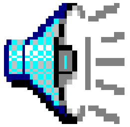
Spectrograms

- The time-varying spectral characteristics of the speech signal can be graphically displayed through the use of a tow-dimensional pattern.
- Vertical axis: frequency, Horizontal axis: time
- The pseudo-color of the pattern is proportional to signal **energy** (red: high energy)
- The resonance frequencies of the vocal tract show up as **“energy bands”**
- Voiced intervals characterized by striated appearance (periodically of the signal)
- Un-Voiced intervals are more solidly filled in



Time domain view

"A lathe is a big tool"



Spectrogram view

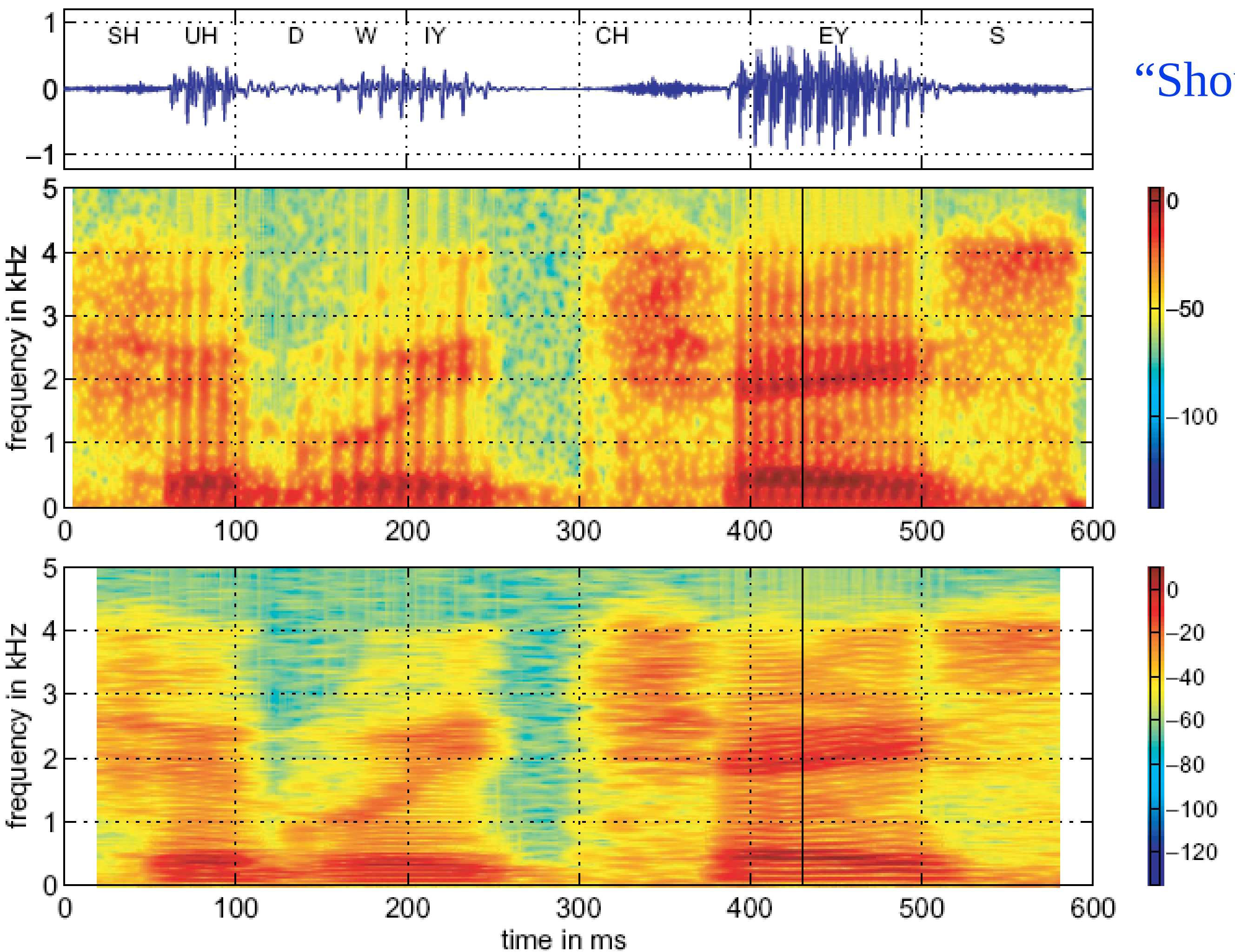
Voiced region

Un-Voiced region

Analysis Window

- **Wide-Band Spectrogram**: A narrow analysis window (narrower than the pitch period) – the narrow vertical lines match succeeding pitch periods
- **Narrow-Band Spectrogram**: A wide analysis window (includes several pitch periods) – the narrow horizontal lines are pitch harmonies
- The yellow bands describe the formants change in time (previous slide)

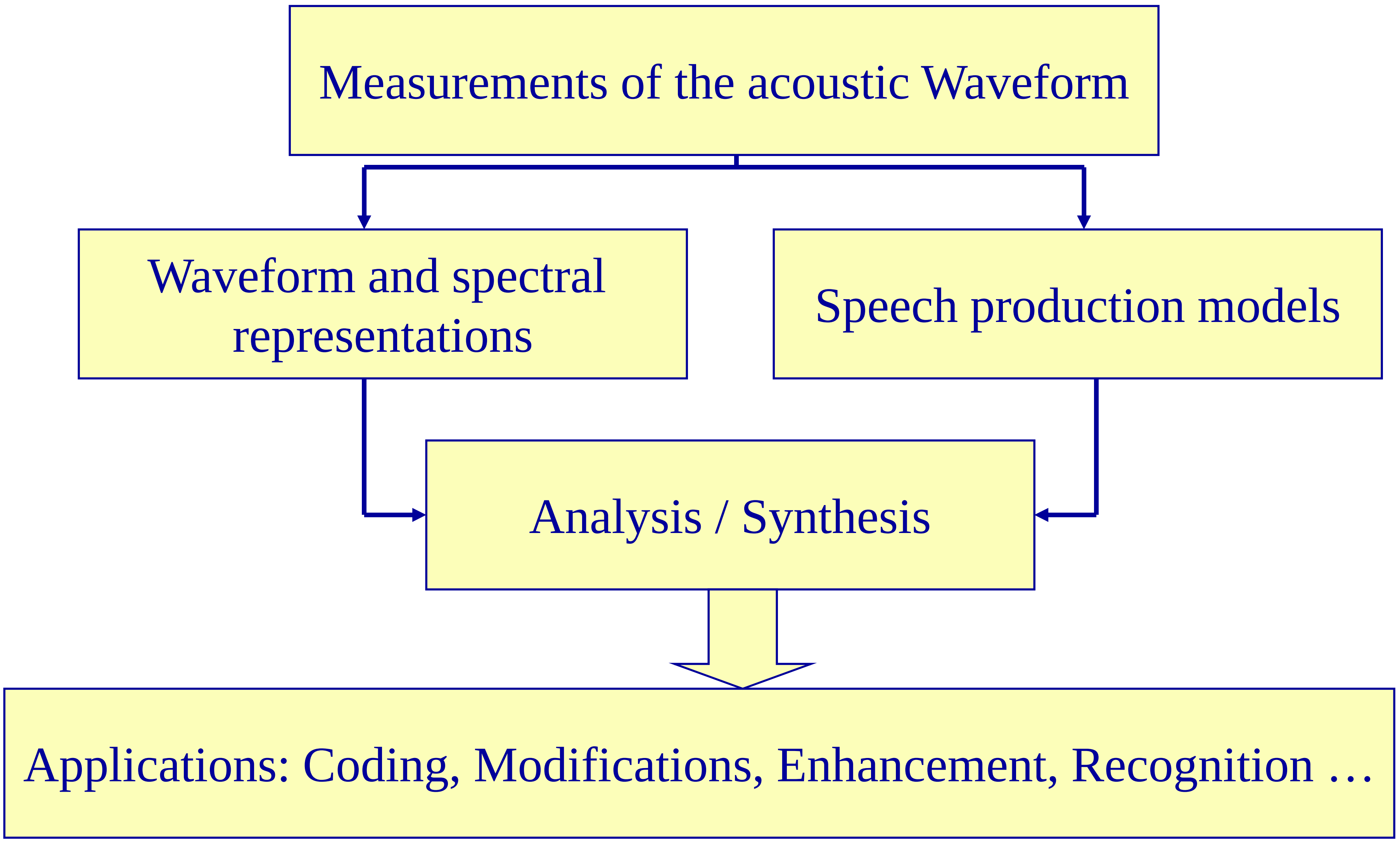
Spectrograms with different analysis Window



"Should we chase"

Speech Processing Applications

A Typical Speech Processing



Modifications

- The goal is to alter the speech signal to have some desired property: time-scale, pitch, and spectral changes.
- **Applications**: fitting radio and TV commercials into an allocated time slot, synchronization of audio and video presentations, etc.
- In addition, speeding up speech has use in message playback, voice mail, and reading machines and books for the blind, while slowing down speech has application to learning a foreign language.

Modification Demo

	Male Speaker	Female Speaker
Original	tfq.tea.org.10k	ln.swm.org.10k
Fast	tfq.tea.tsmtv0p8.10k	ln.swm.tsmtv0p8.10k
Faster	tfq.tea.tsmtv0p5.10k	ln.swm.tsmtv0p5.10k
Slow	tfq.tea.tsmtv1p2.10k	ln.swm.tsmtv1p2.10k
Slower	tfq.tea.tsmtv1p5.10k	ln.swm.tsmtv1p5.10k

Pitch and vocal tract length change - Sinewave-based modification

	Male Speaker	Female Speaker
Original	cp.seg.org.8k	glo.org.8k
Low pitch/Long vocal tract	cp.seg.PitSpec_low.8k	glo.PitSpec_low.8k
High pitch/Short vocal tract	cp.seg.PitSpec_high.8k	glo.PitSpec_high.8k

Some Wideband Audio Examples

TSM: Original (Depeche mode : Martyr)
Mono, 15 Sec
Fast – 50%
Slow – 200%

Automatic Transcription: Original

Polyphonic Wav-to-MIDI

Speech Enhancement

- The goal: to improve the quality of degraded speech.
- One approach is to *pre-process* the (analog) speech waveform before it is degraded.

- Another is *post-processing* : enhancement after the signal is degraded:
  - Increasing the transmission power, e.g.: automatic gain control (AGC) in a noisy environment.
  - Reduction of additive noise in digital telephony, and vehicle communication and aircraft communication.
  - Reduction of interfering backgrounds and speakers for the hearing-impaired.

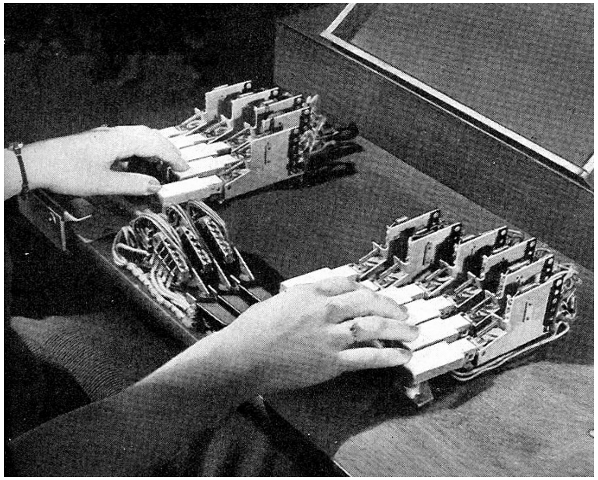
Enhancement demo:

Noise reduction adaptive Wiener filter with adaptivity based on spectral change

	Original	Enhanced
Cellular Telephone Noise	s4141t03.org.10k	s4141t03.enh.10k
Cocktail Party Noise	party.org_modsnr.10k	party.enh_modsnr.10k
Automobile Noise	auto.org_lowsnr.10k	auto.enh_lowsnr.10k

Speech Synthesis Demo

- Voder
- 1939 New York Worlds Fair: First speech synthesizer
- H. Dudley, R.R. Reisz, and S.S.A. Watkins,
- “A synthetic speaker”, Journal of the Franklin Institute, vol. 227, pp. ,1939.



http://davidszondy.com/future/robot/voder.htm

And Some Modern TTS Machines

AT&T

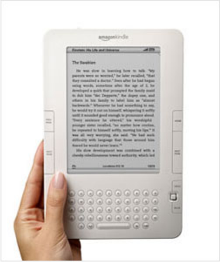
- http://public.research.att.com/~ttsweb/tts/demo.html

Oddcast

- http://vhost.oddcast.com/vhost_minisite/demos/tts/tts_example.html

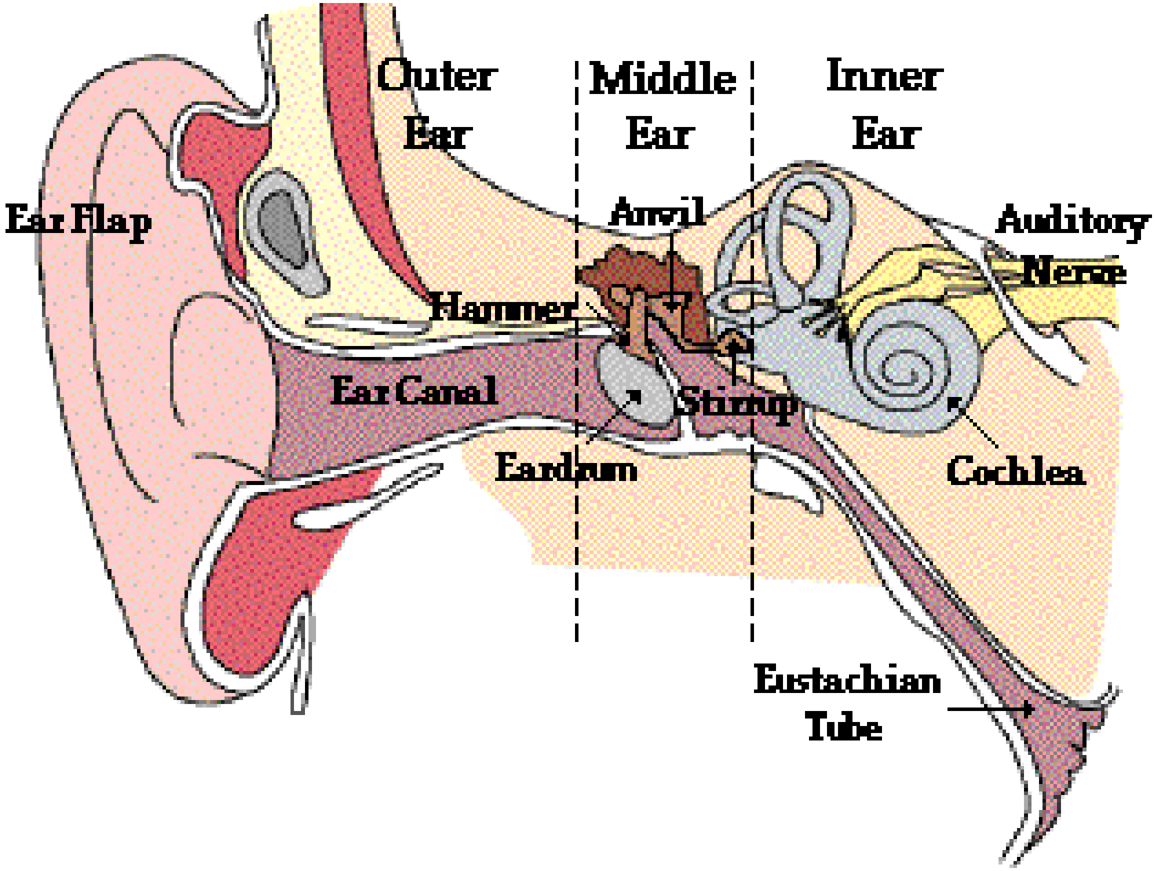
- A modern book reader: Kindle by Amazon

  And some Criticism …

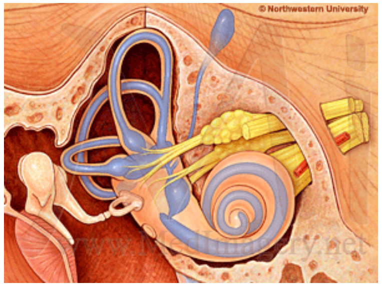


The Human Auditory System

The Human Hearing System



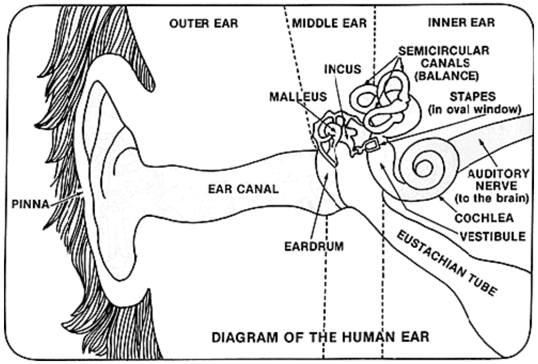
Hearing System



- The ear performs frequency analysis of the received signal, and allows the listener to discriminate small differences in time and frequency found in the sound

- Hearing range: ~ 16Hz - 18KHz
- Most sensitive to the range: 2KHz - 4KHz
- Dynamic range (quietest to loudest) is about 96 dB
-
- Normal human voice range is about 500 Hz to 2 kHz

How does it work ?



1. **The Outer Ear: Catch the Wave**

- Called the **pinna** or **auricle**
- After waves enter the outer ear, they travel through the **ear canal** and make their way toward the middle ear.
- The outer ear canal's other job is to protect the ear by making earwax. That special wax contains chemicals that fight off infections that could hurt the skin inside the ear canal. It also collects dirt to help keep the ear canal clean.

The Eardrum

- The eardrum is a piece of thin skin stretched tight.
- It is attached to the first ossicle, a small bone called the **malleous** (hammer) which is attached to another tiny bone called the **incus** (anvil), which is attached to the smallest bone in the body, called the stirrup (**stapes)**
- When sound waves travel into the ear and reach the eardrum, they cause the eardrum to vibrate. These sound vibrations are carried to the three tiny bones of the middle ear.

The Middle Ear

- The middle ear's main job is to take the sound waves, turn them into vibrations, and deliver them to the inner ear.

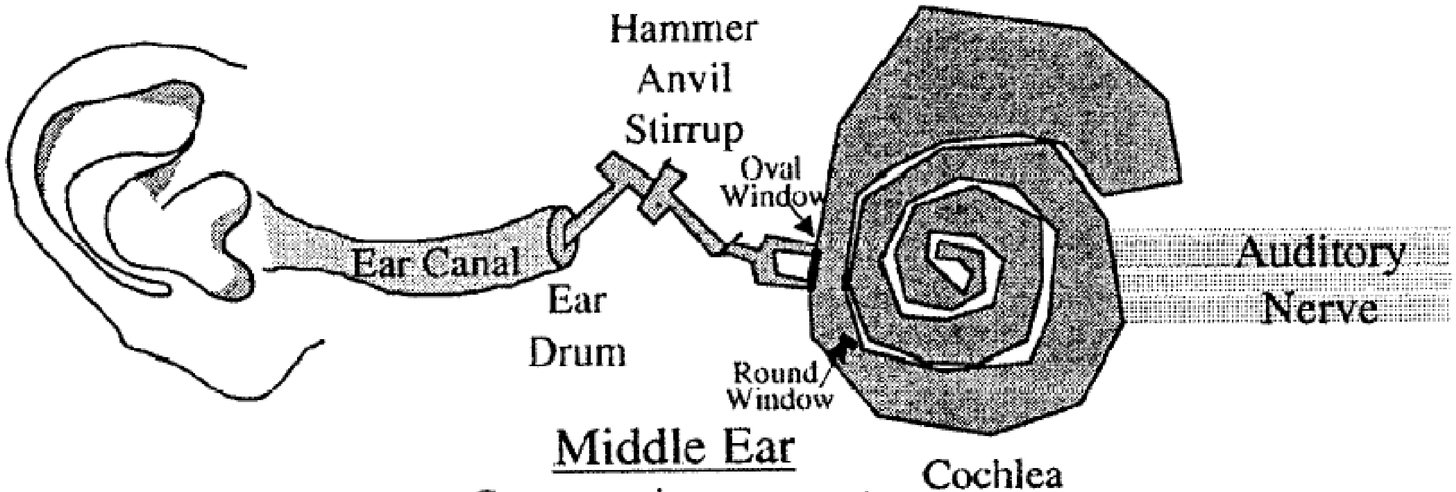
- It also helps the eardrum handle the pressure:
  - The middle ear is connected to the back of the nose by a narrow tube called the **eustachian** tube. Together the eustachian tube and the middle ear keep the air pressure equal on both sides of the eardrum.

Keeping the air pressure equal is important so the eardrum can work properly and not get injured.

The Inner Ear

- The vibrations in the **inner ear** go into the **cochlea** - a small, curled tube in the inner ear.
- The cochlea is filled with liquid and lined with cells that have thousands of tiny microscopic hairs (15,000 to 20,000) on their surface.
- When the sound vibrations hit the liquid in the cochlea, the liquid begins to vibrate. Different kinds of sounds will make different patterns of vibrations.
- The vibrations cause the sensory hairs in the cochlea to move - sound vibrations are transformed into nerve signals and delivered to the brain via the hearing nerve
  - Also called the "eighth nerve"

Outer, Middle and Inner Ear



Middle Ear

Converts air movement in ear canal to fluid movement in cochlea.

Outer Ear

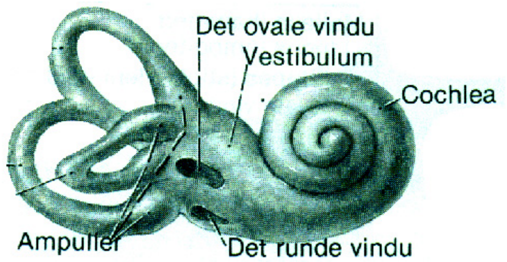
Collects sound and funnels it down to ear drum. Physical size tuned to sounds around 4 kHz.

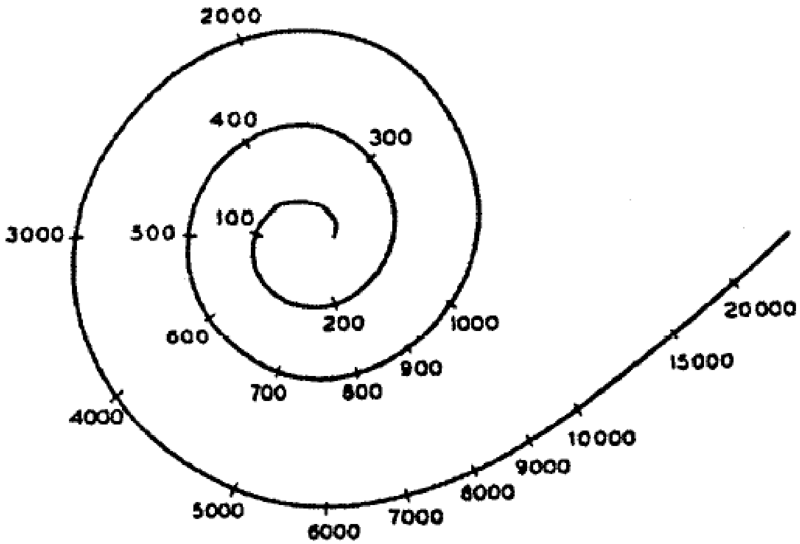
Inner Ear

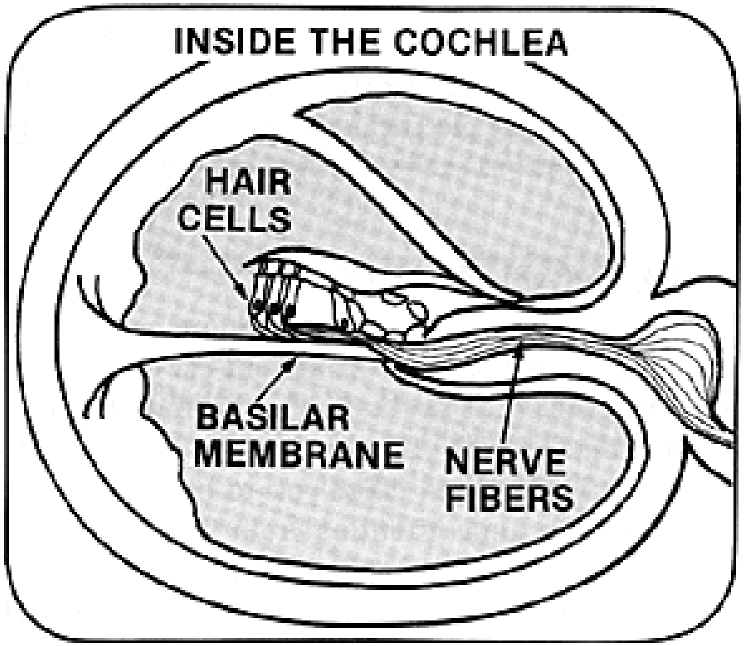
Cochlea separates sounds by frequency. Hair cells convert fluid motion into electrical impulses in auditory nerve.

(From: Intro. to Digital Audio Coding and Standards, Bosi & Goldberg)

Inside The Cochlea



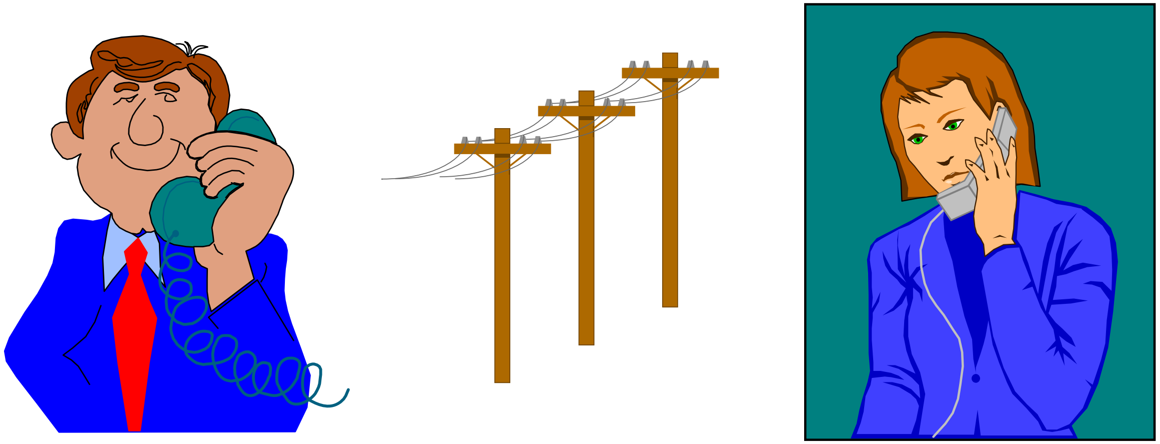




Frequency Sensitivity along the **Basilar Membrane**
American Physical Society, 1940

Hearing: Thresholds

- *Hearing Threshold*: Minimum intensity at which sounds can be perceived
- The ear is less sensitive to the pitch and first formant (F1) than to the higher formants, in the sense of intelligibility

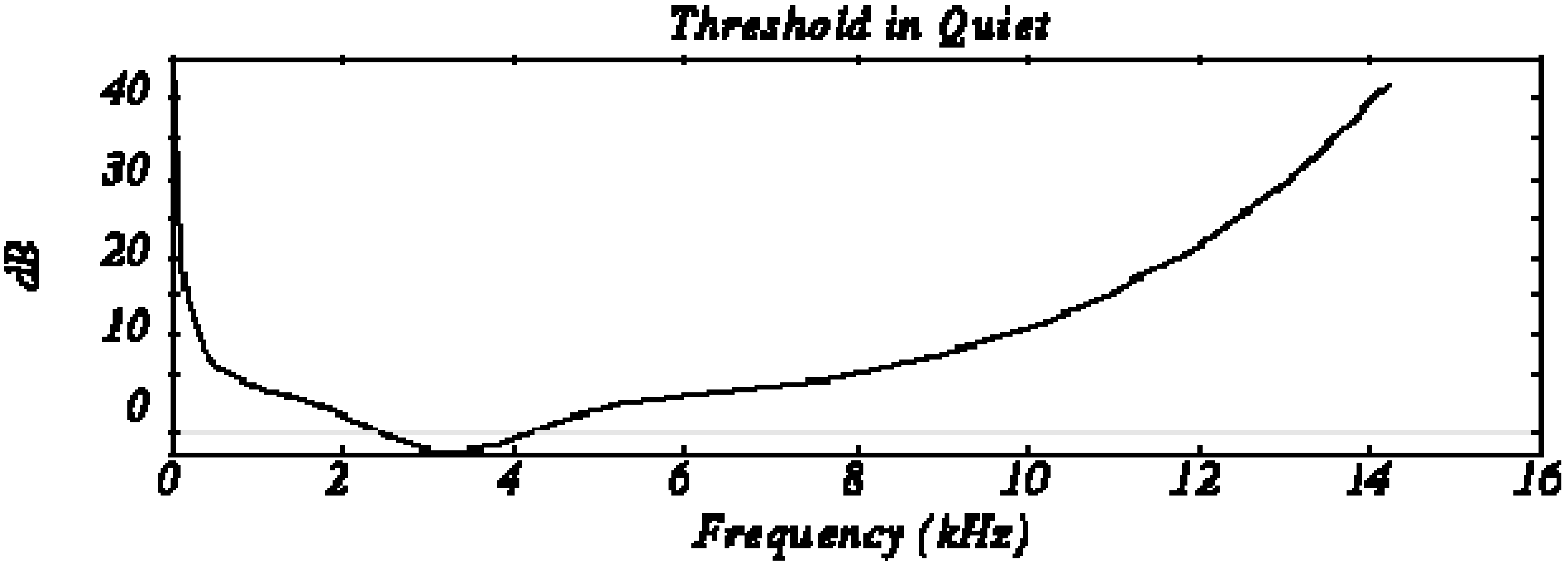


Hearing: Masking

- *Masking*: One sound is obscured in the presence of another: presence of one raises the threshold for another one
- Lower frequencies usually mask higher frequencies, with largest effect near the harmonics of the masker
- A wider band signal masks narrower band signal

How sensitive is human hearing?

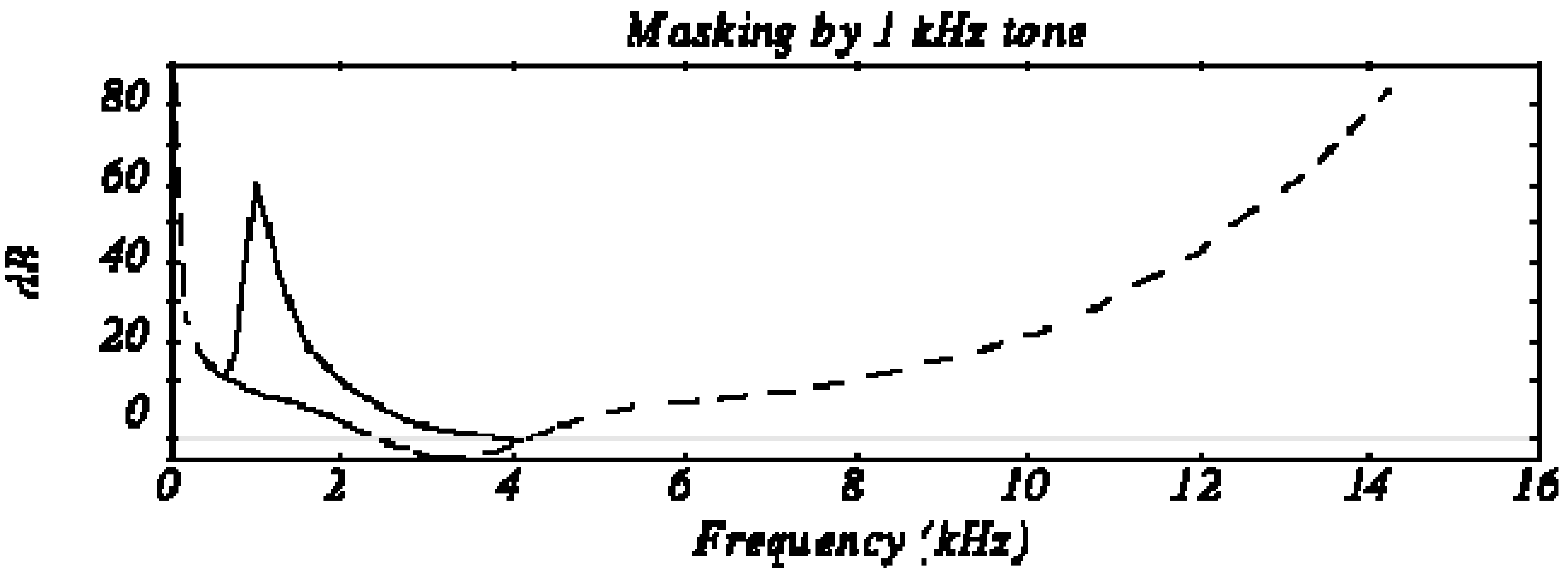
- Put a person in a quiet room. Raise level of 1 kHz tone until just barely audible.
- Vary the frequency and plot.



Frequency Masking

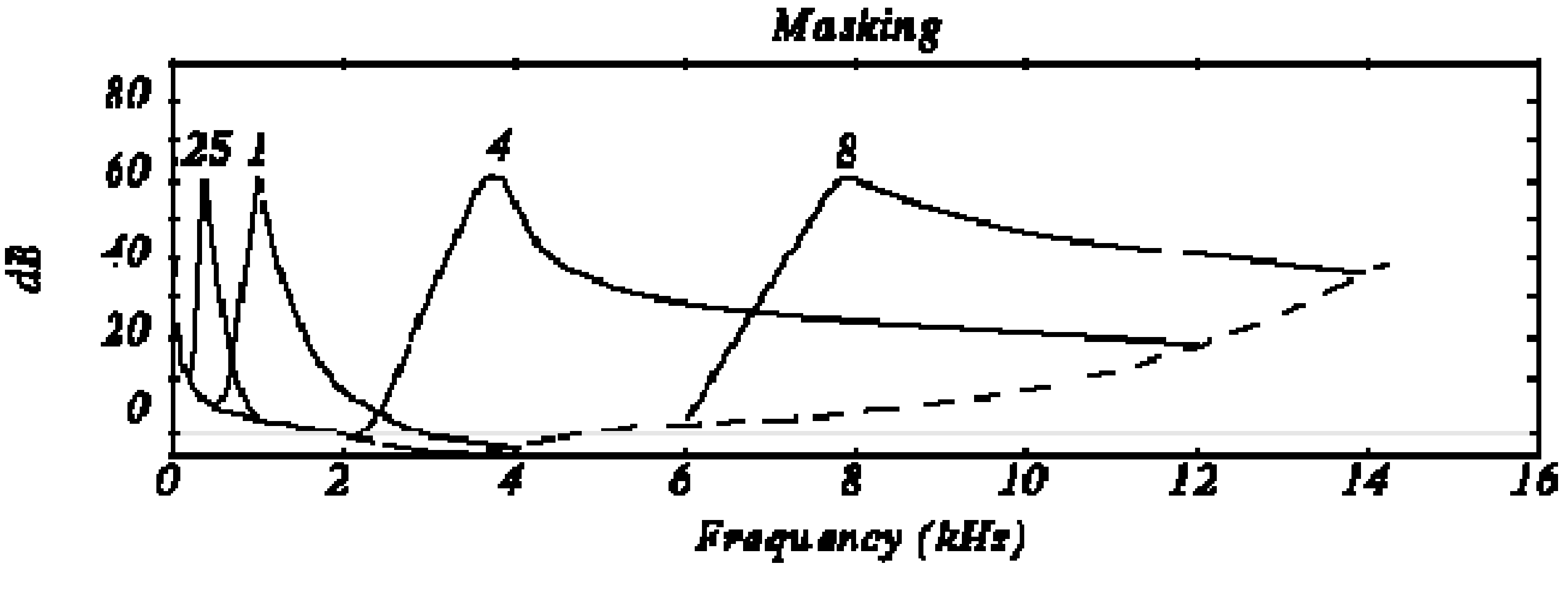
Play 1 kHz tone (masking tone) at fixed level (60dB). Play test tone at a different level (e.g., 1.1kHz), and raise level until just distinguishable.

Vary the frequency of the test tone and plot the threshold when it becomes audible:



Frequency Masking (Cont'd)

Repeat for various frequencies of masking tones:



Critical Bands

- Perceptually uniform measure of frequency, non-proportional to width of masking curve,

About 100 Hz for masking frequency < 500 Hz, grow larger and larger above 500 Hz.

- The width is called the size of the critical band

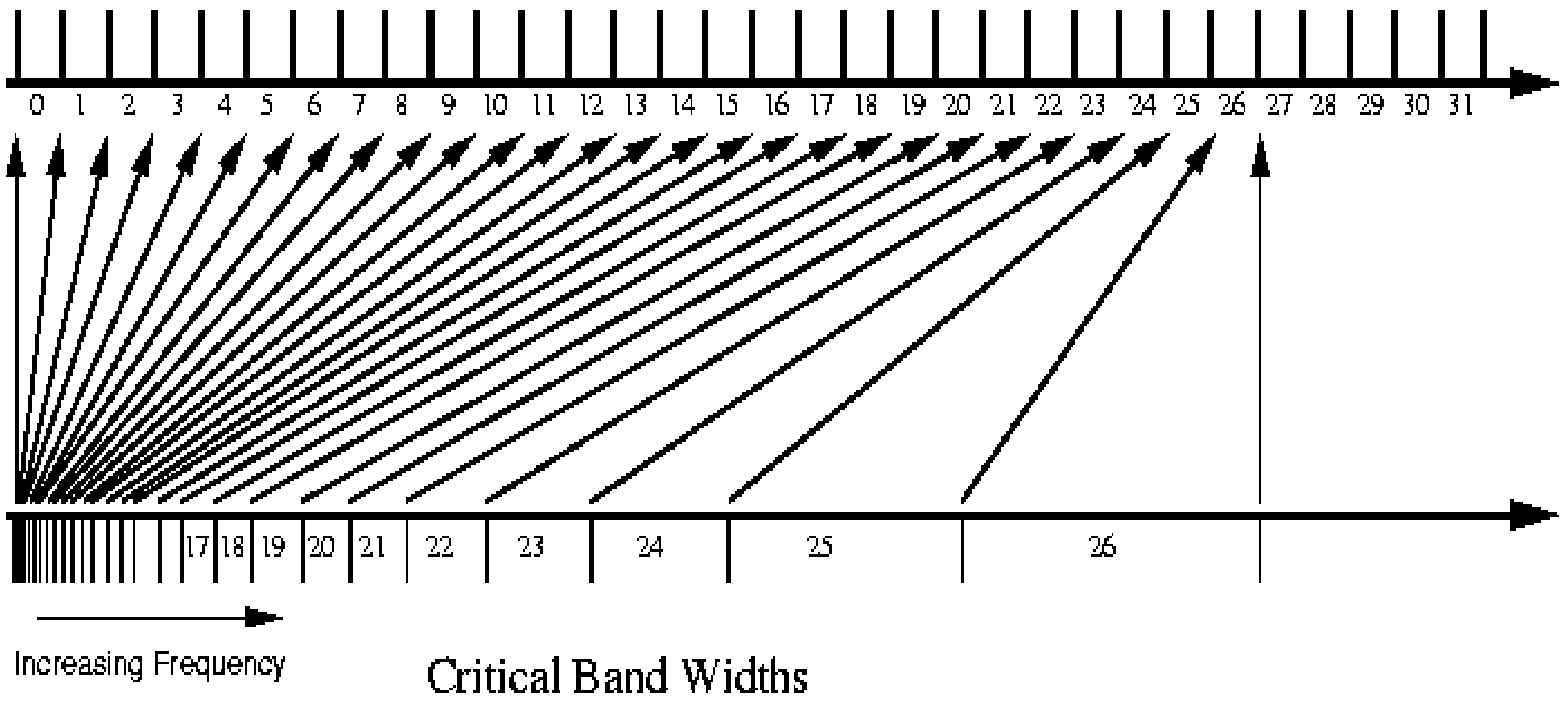
Barks

- Introduce new unit for frequency called a bark (after Barkhausen)
- 1 Bark = width of one critical band

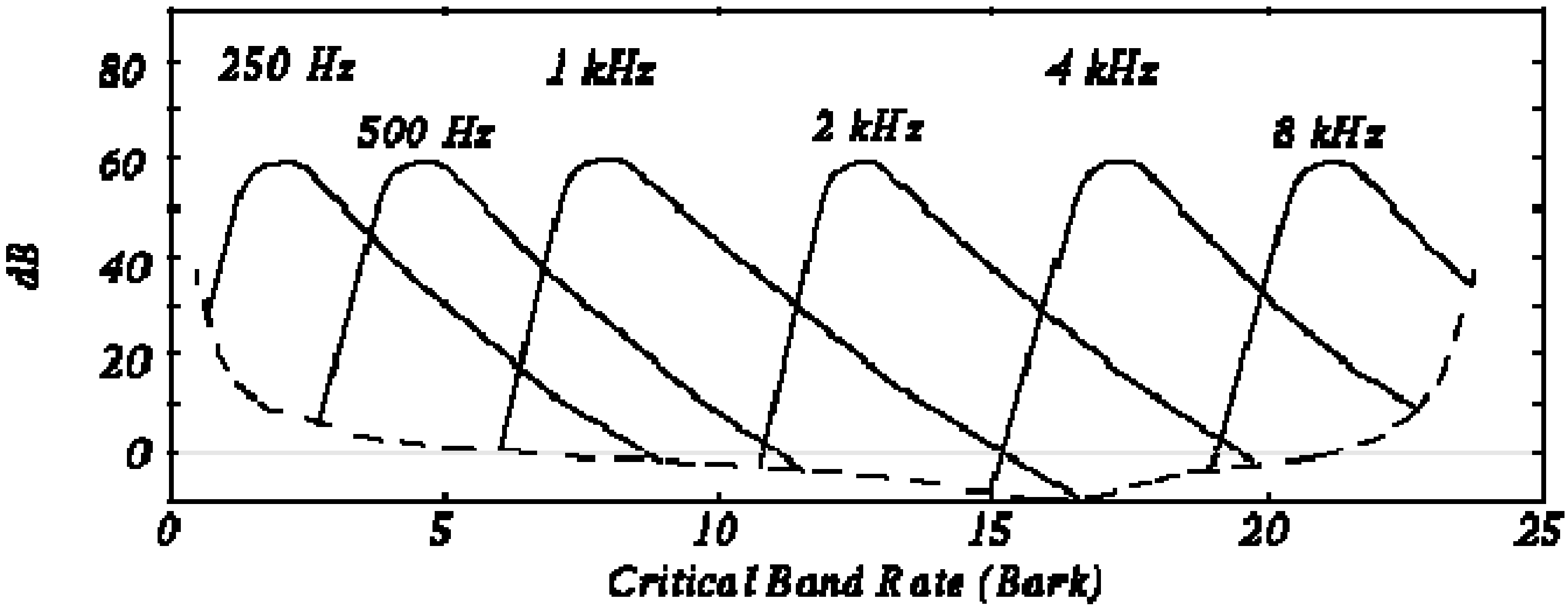
- For frequency < 500 Hz, 1 Bark $\simeq$ Freq/100
- For frequency > 500 Hz,

$$1 \text{ Bark} \simeq 9 + 4\log(\text{Freq}/1000)$$

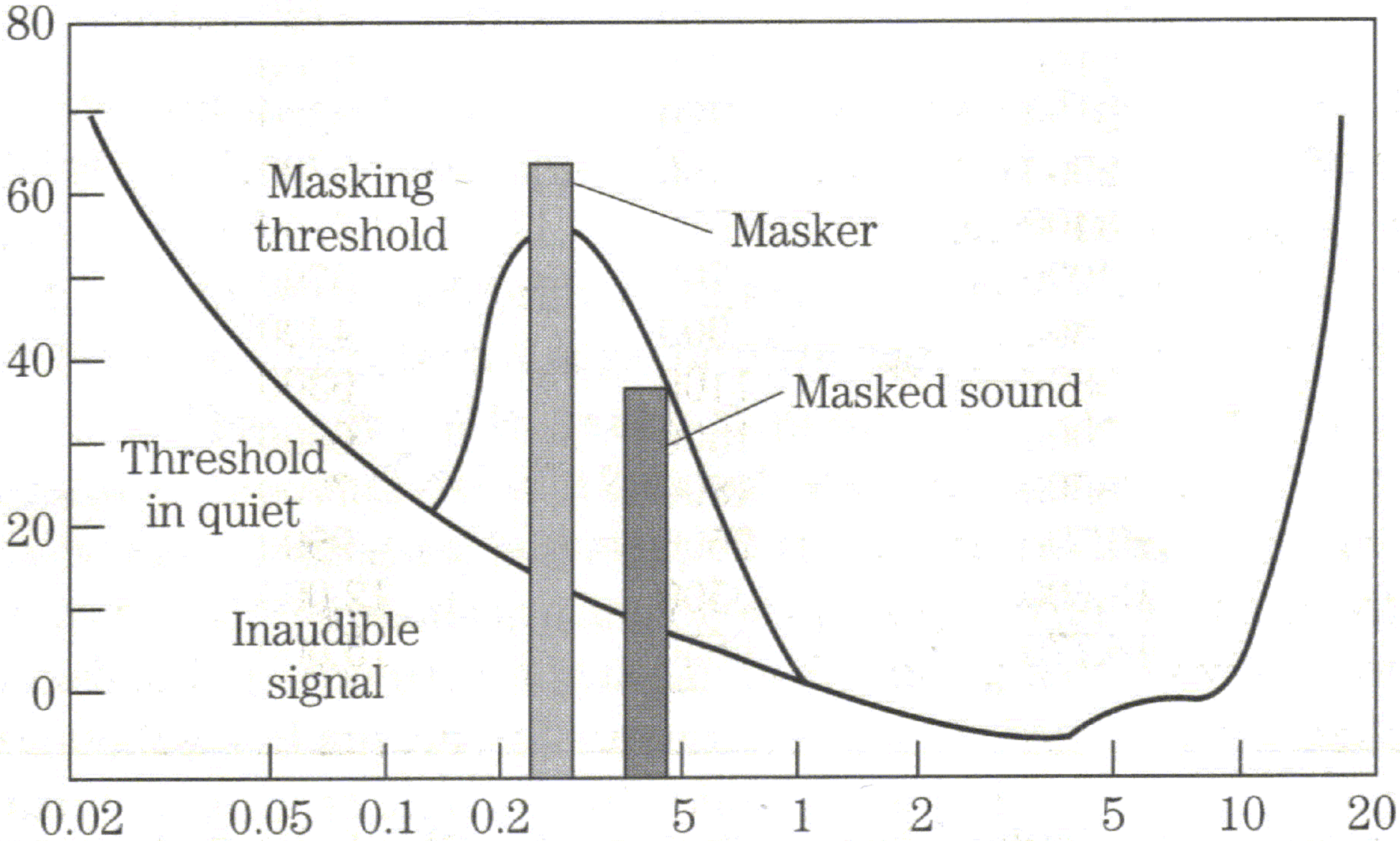
Frequency partitioning into Barks



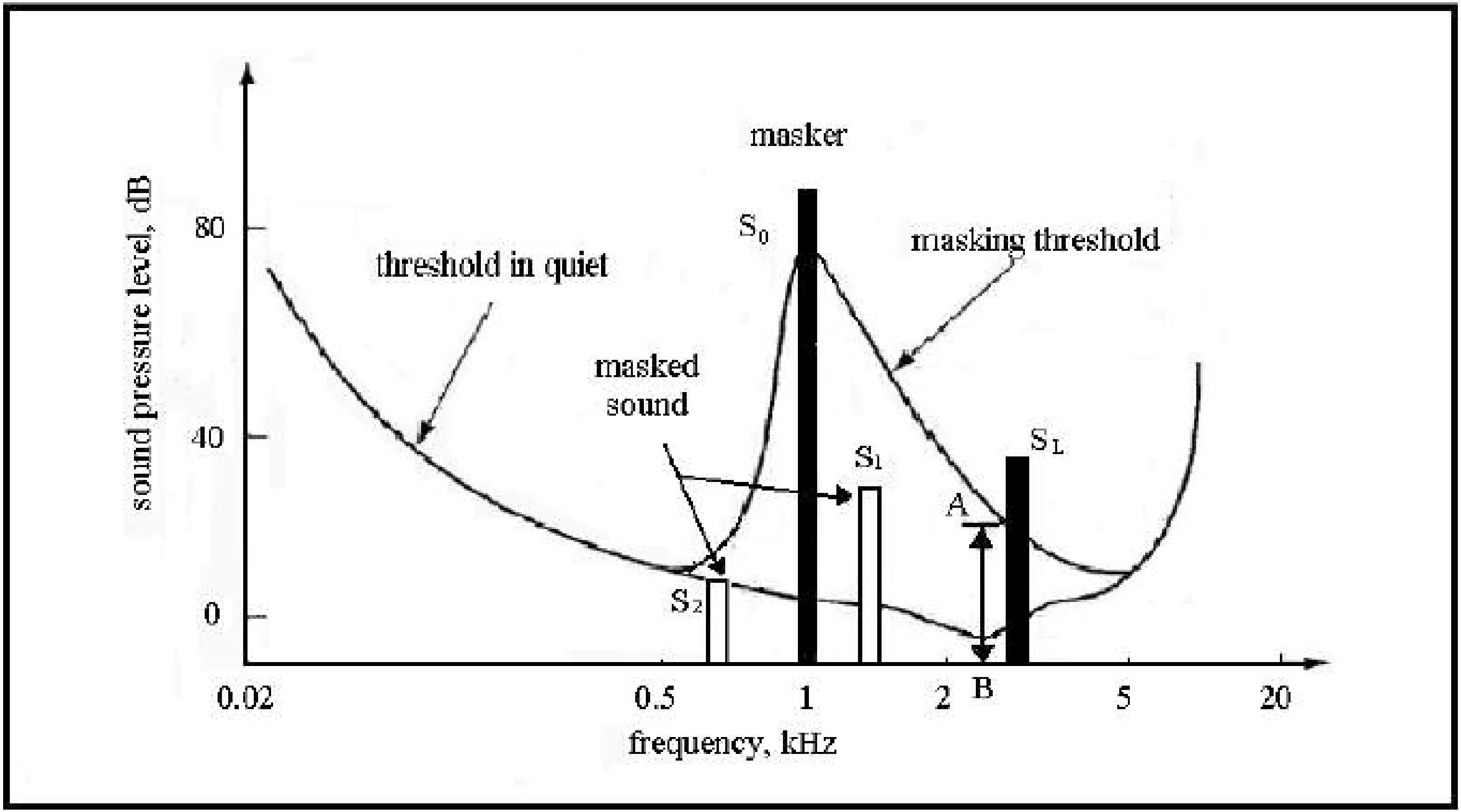
Masking Thresholds on critical band scale:



Masking effect



Masked and non-masked tones



Temporal masking

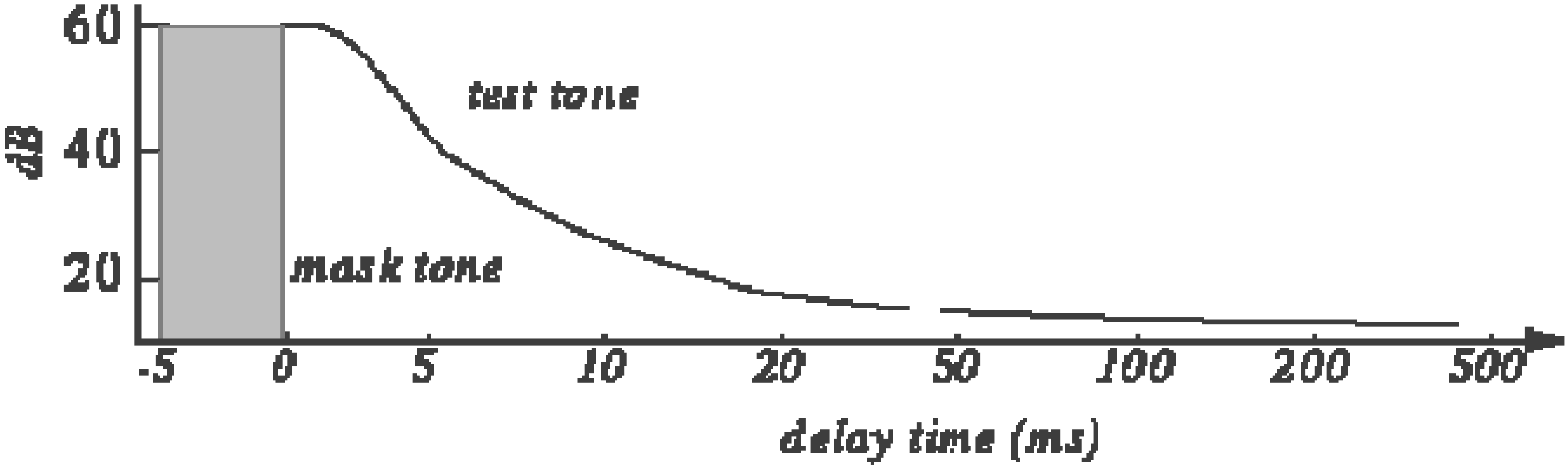
- If we hear a loud sound, then it stops, it takes a little while until we can hear a soft tone nearby
- Question: how to quantify?

Temporal masking (Cont’d)

- Experiment: Play 1 kHz masking tone at 60 dB, plus a test tone at 1.1 kHz at 40 dB:
  - Test tone can't be heard (it's masked).
- Stop masking tone, then stop test tone after a short delay.
- Adjust delay time to the shortest time that test tone can be heard (e.g., 5 ms).

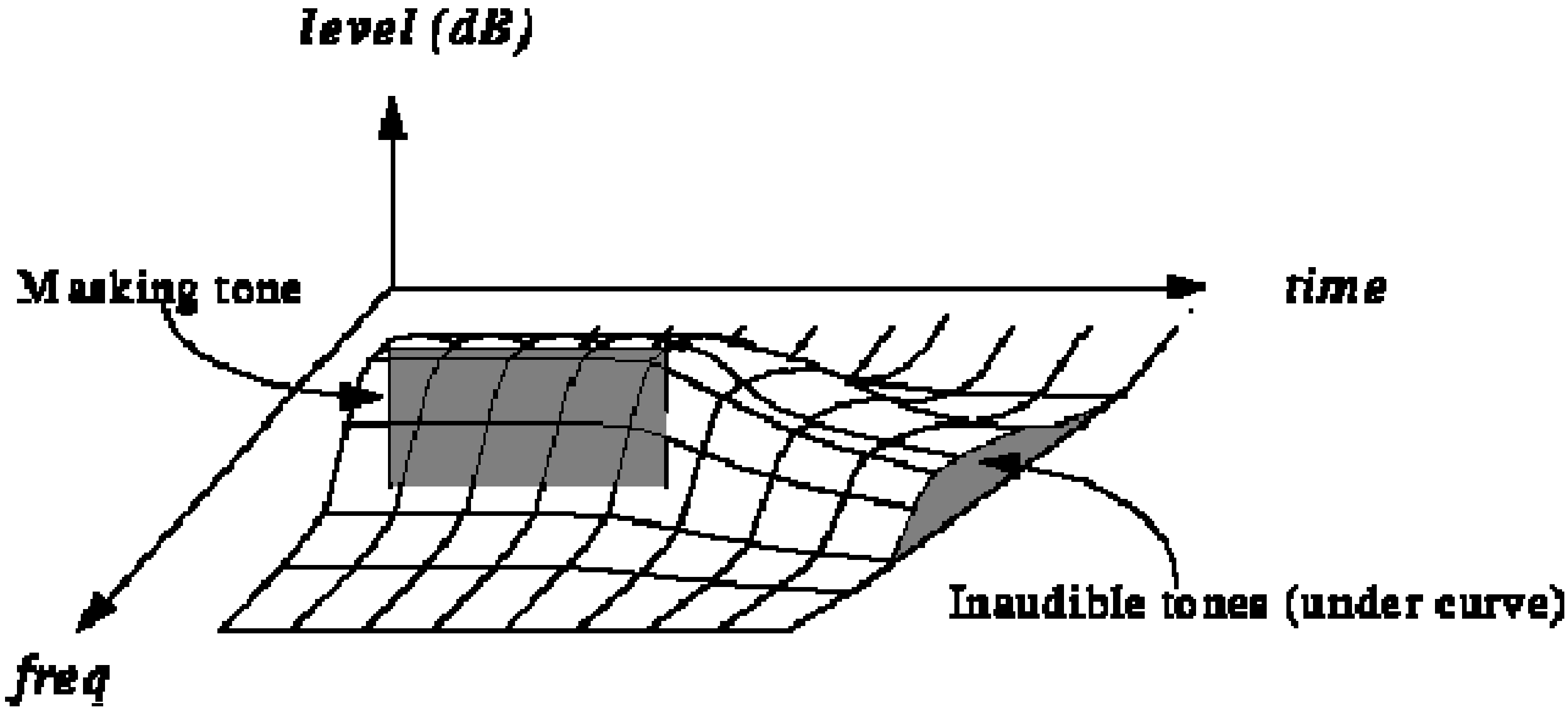
Temporal masking (Cont'd)

Repeat with different level of the test tone and plot:

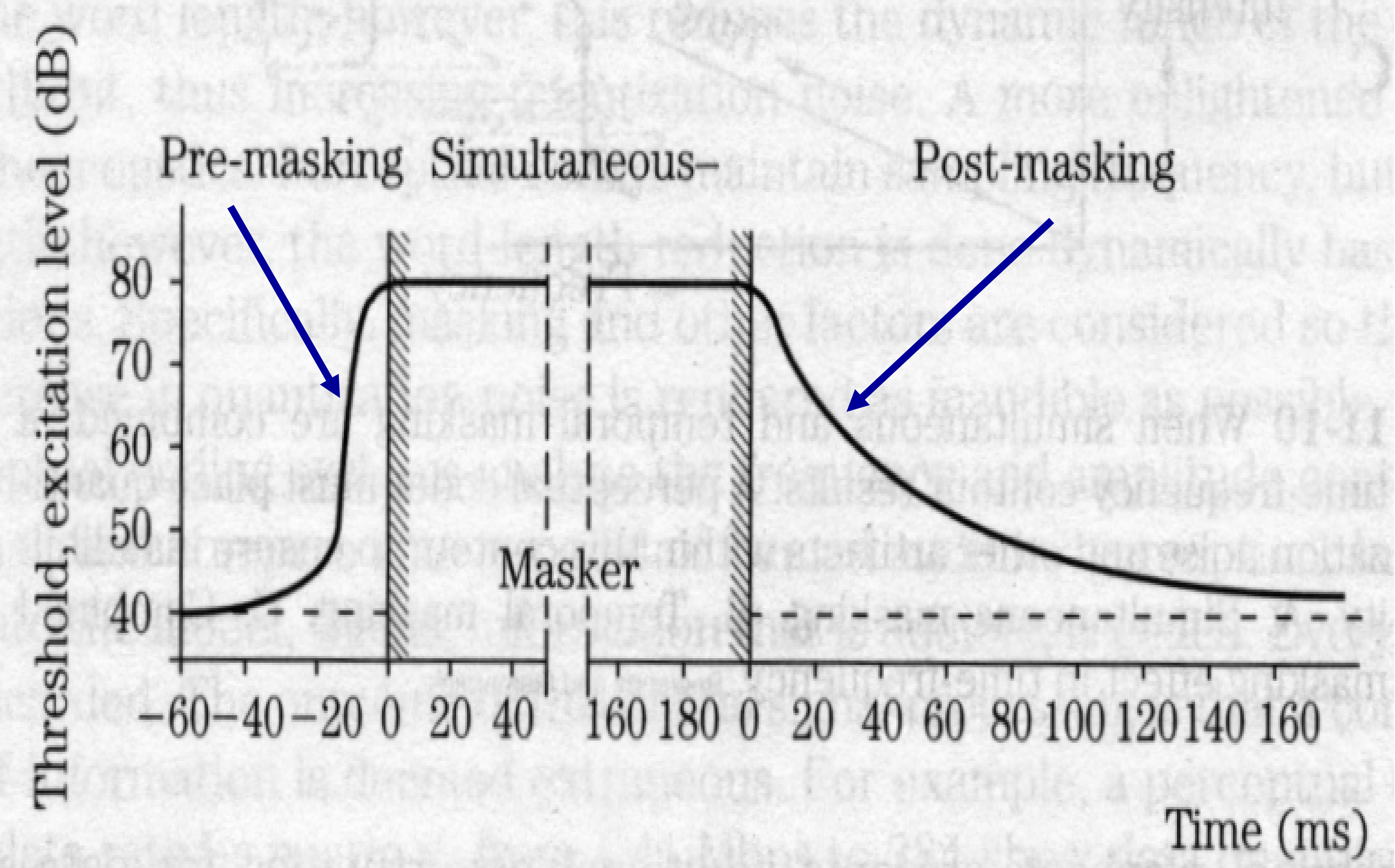


Temporal masking (Cont'd)

Try other frequencies for test tone (masking tone duration constant). Total effect of masking:



Masking effect in time



Voiced and Unvoiced Noise Thresholds

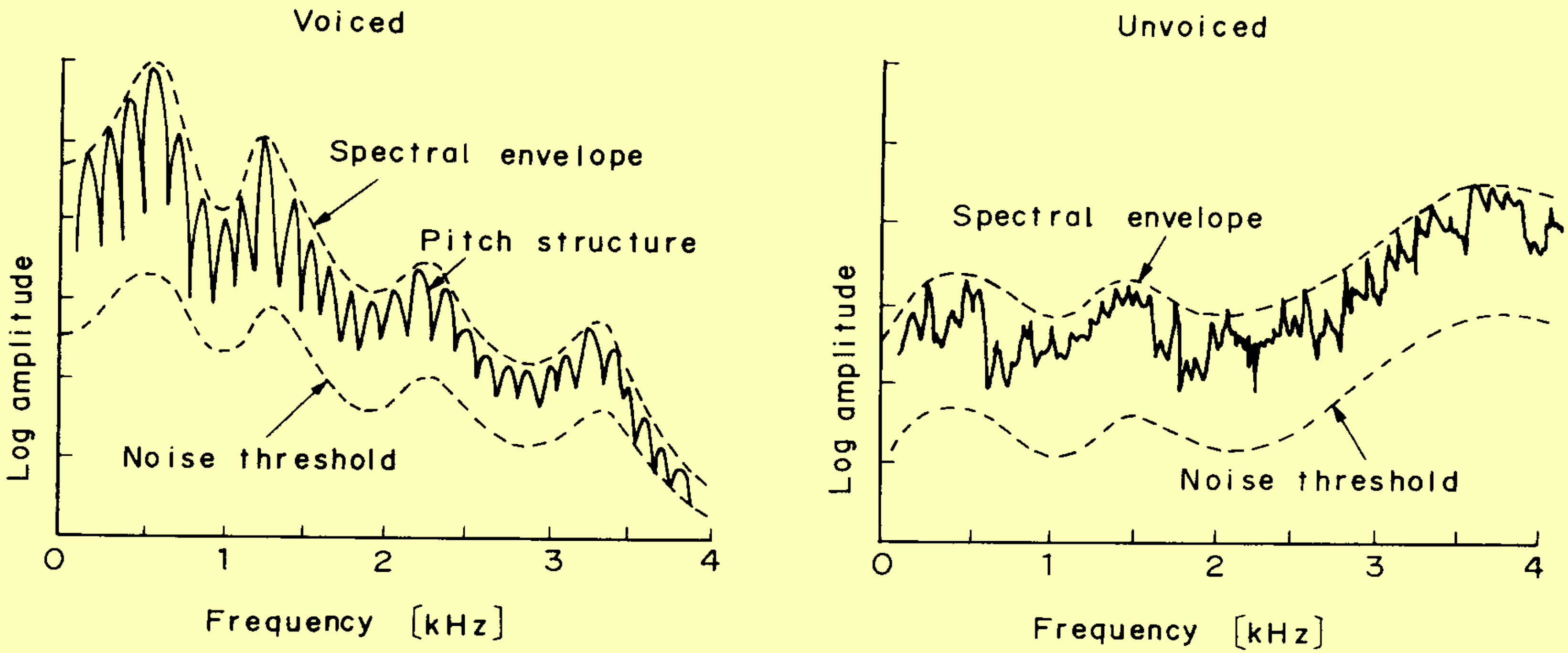
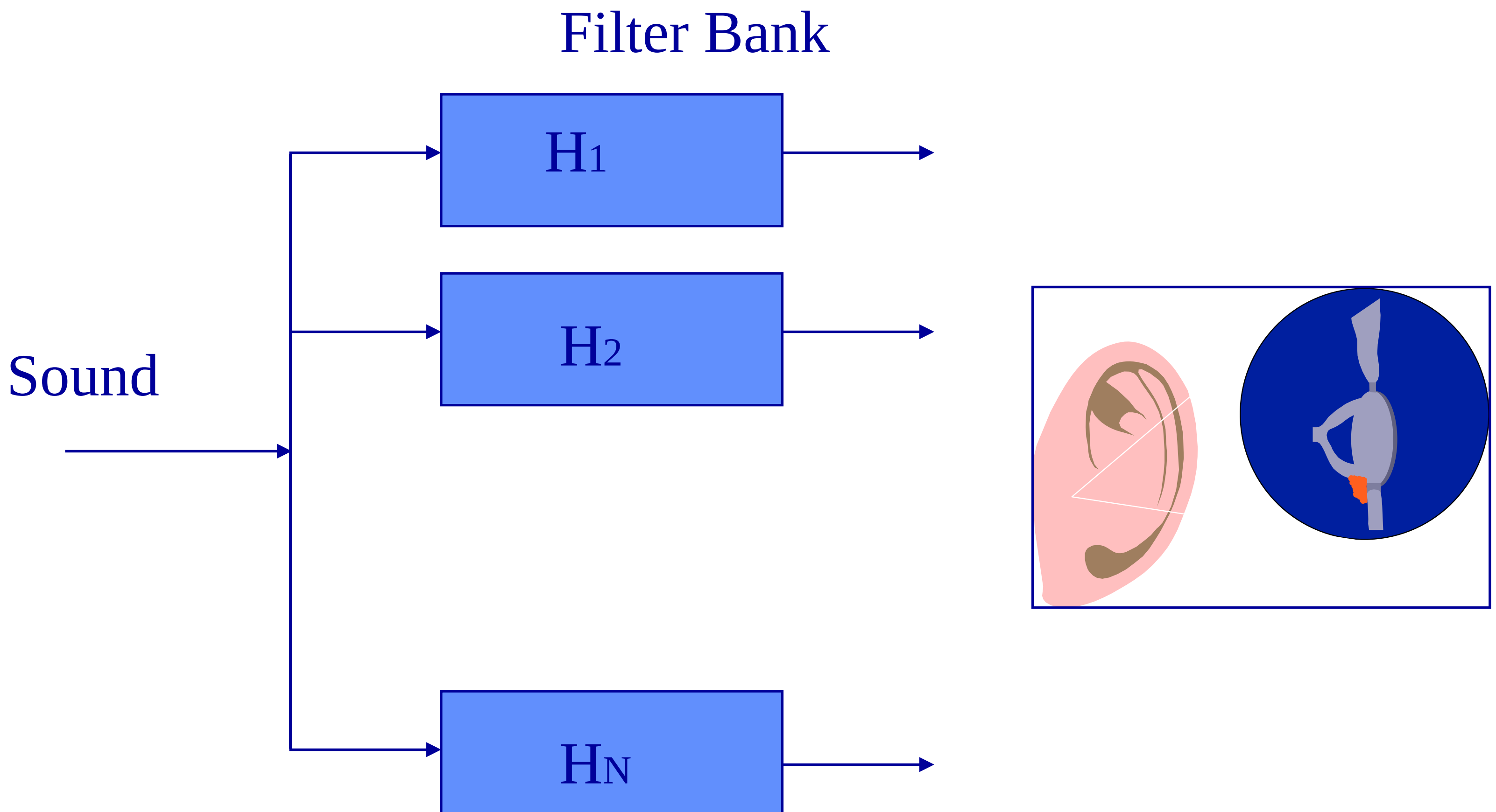


Fig. 4.12 Noise threshold for auditory masking associated with spectral envelope.

Hearing Model



Speech Quality

- Includes:
  - Intelligibility (Phone)
  - Naturalness
  - Speaker Identify
  - Perceptual convenience (?)
  - More….

How to quantify the degradation ?

- SNR, MSE etc. are not perceptual measures

- <u>**Subjective criteria**</u> (*Listening tests: MOS*) are expensive and time consuming

- <u>PESQ</u>: an objective used standard

SP DEMO Speech and Audio Processing Learning Tool

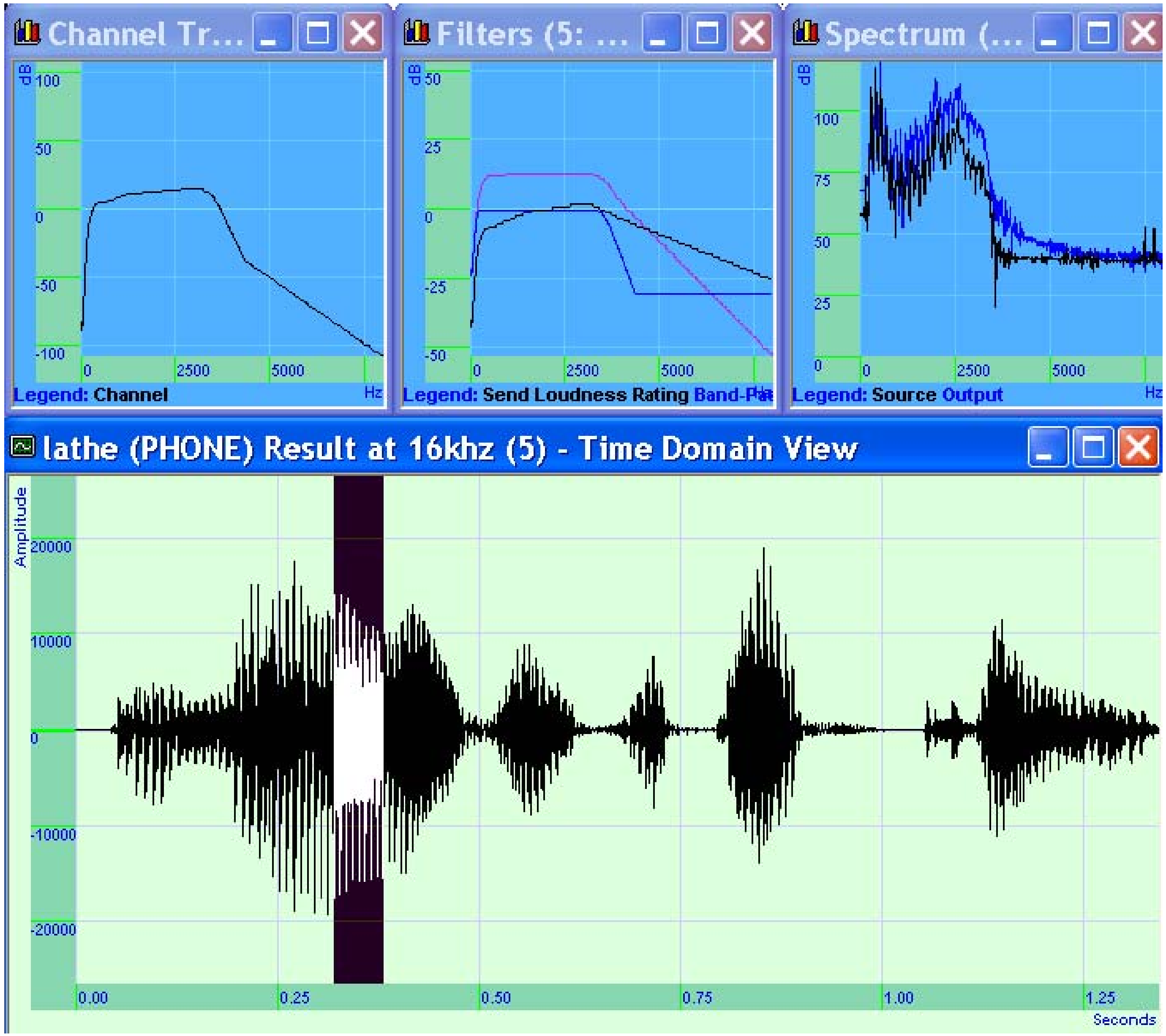
What about the Phone Line ?

- Band limited: ~300-3.3 KHz
- Distortions, Echo, Noise etc.

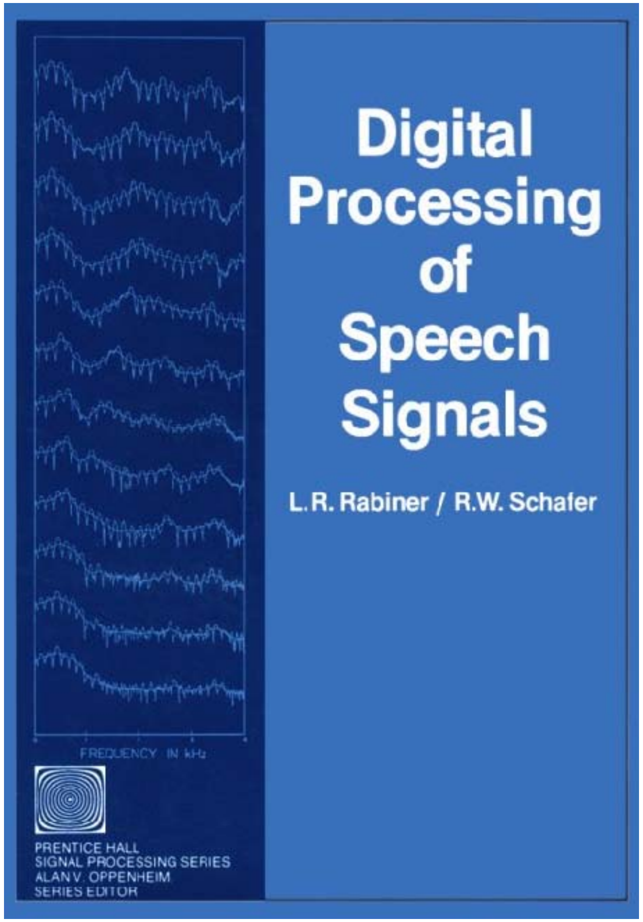
SP DEMO Speech and Audio Processing Learning Tool

Original

Band limited
(Phone-line)

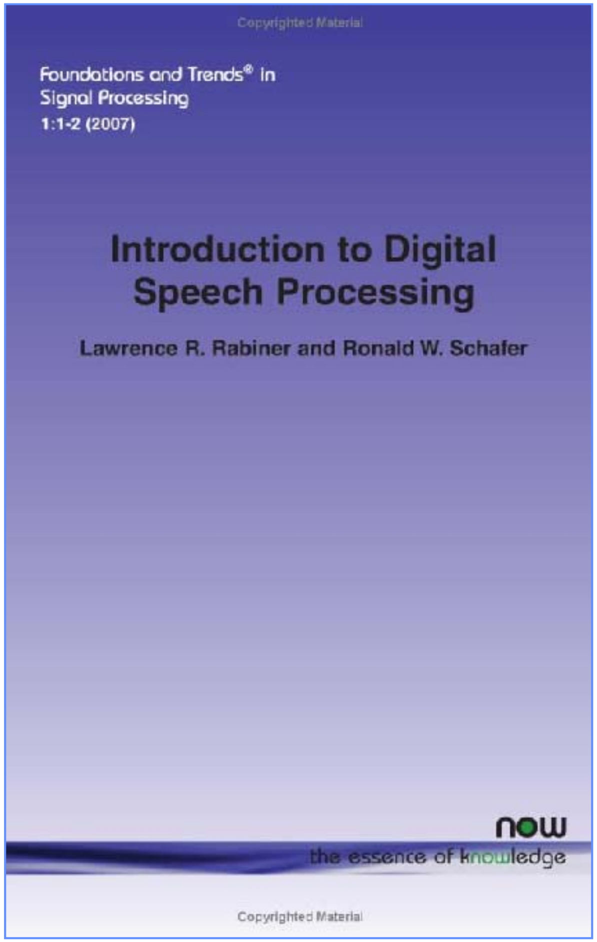


(A very good) Reference book:



Introduction to Digital Speech Processing (Foundations and Trends in Signal Processing)

by Lawrence R. Rabiner (Author), Ronald W. Schafer (Author)



First edition, 1978

Second edition, 2007