24.963

Linguistic Phonetics

Basic Audition

Diagram of the inner ear removed due to copyright restrictions.

- Reading: Keating 1985
- 24.963 also read Flemming 2001
- Assignment 1 - basic acoustics. Due 9/22.

Audition

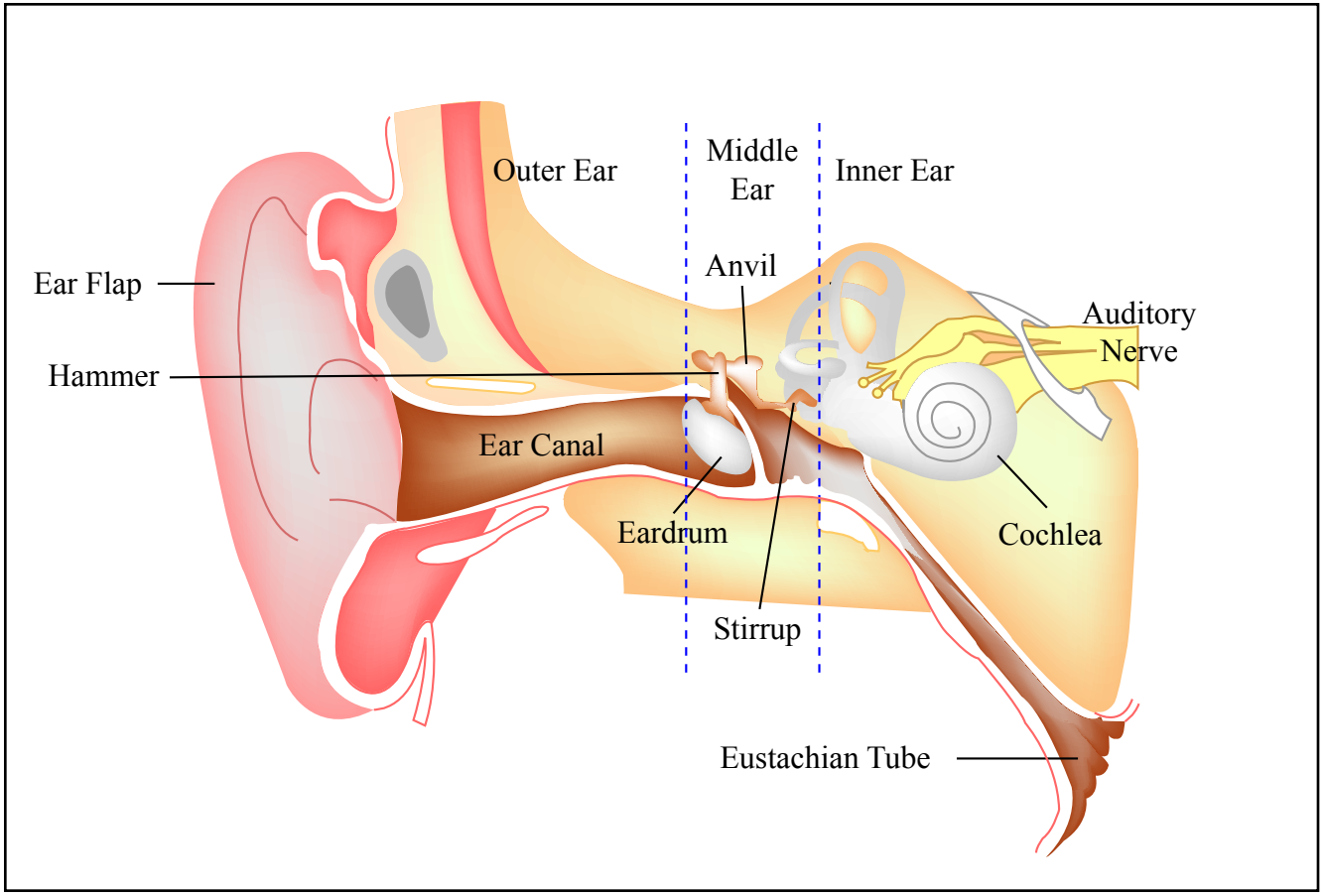


Image by MIT OCW.

Anatomy

Auditory 'spectrograms'

The auditory system performs a running frequency analysis of acoustic signals - cf. spectrogram.

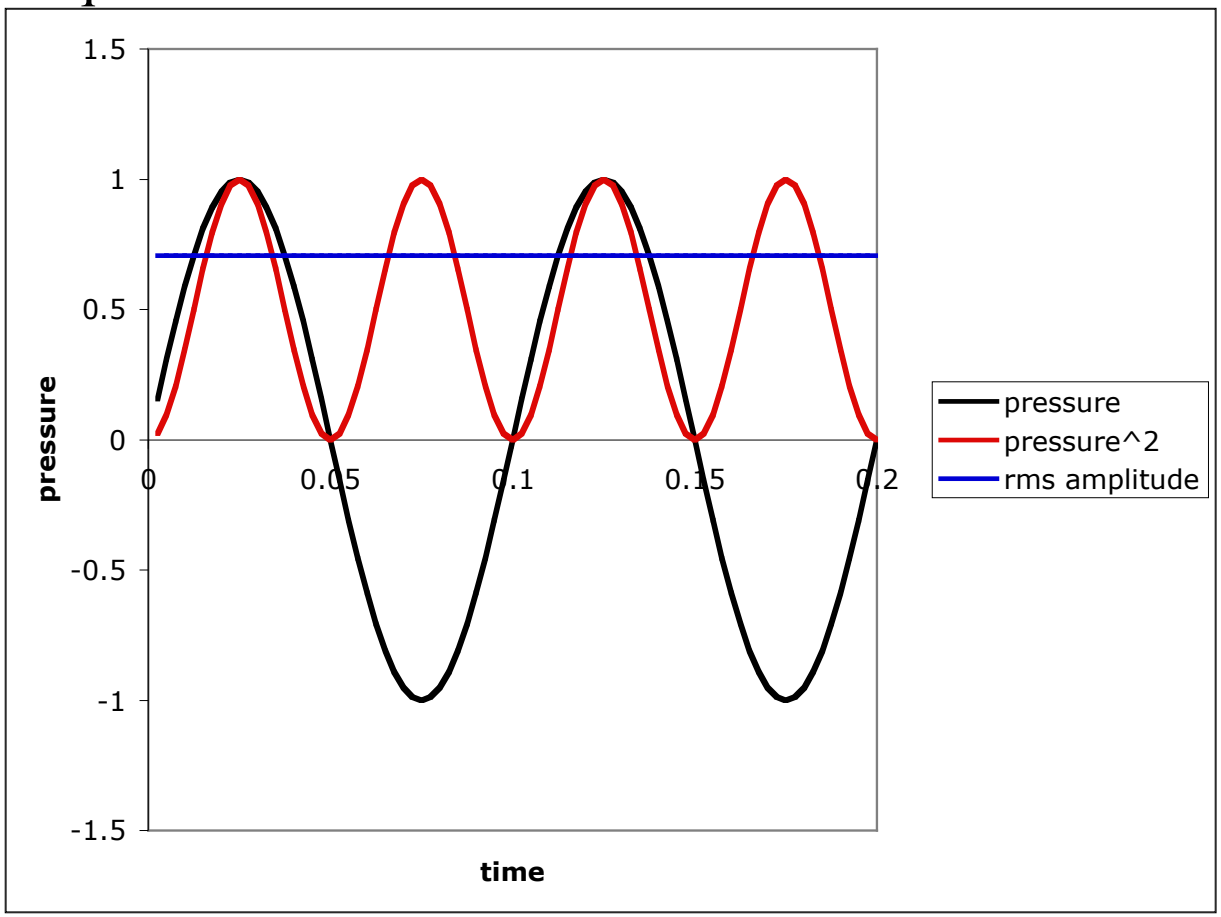
- But the auditory spectrogram differs from a regular spectrogram in significant ways, e.g.:
- Frequency - a regular spectrogram analyzes frequency bands of equal widths, but the peripheral auditory system analyzes frequency bands that are wider at higher frequencies.
- Loudness is non-linearly related to intensity
- Masking effects (simultaneous and non-simultaneous).
- It is useful to bear these differences in mind when looking at acoustic spectrograms.
- It is possible to generate 'auditory spectrograms' based on models of the auditory system.

Loudness

- The perceived loudness of a sound depends on the amplitude of the pressure fluctuations in the sound wave.
- Amplitude is usually measured in terms of root-mean-square (rms amplitude):
  - The square root of the mean of the squared amplitude over some time window.

rms amplitude

- Square each sample in the analysis window.
- Calculate the mean value of the squared waveform:
  - Sum the values of the samples and divide by the number of samples.
- Take the square root of the mean.



rms amplitude

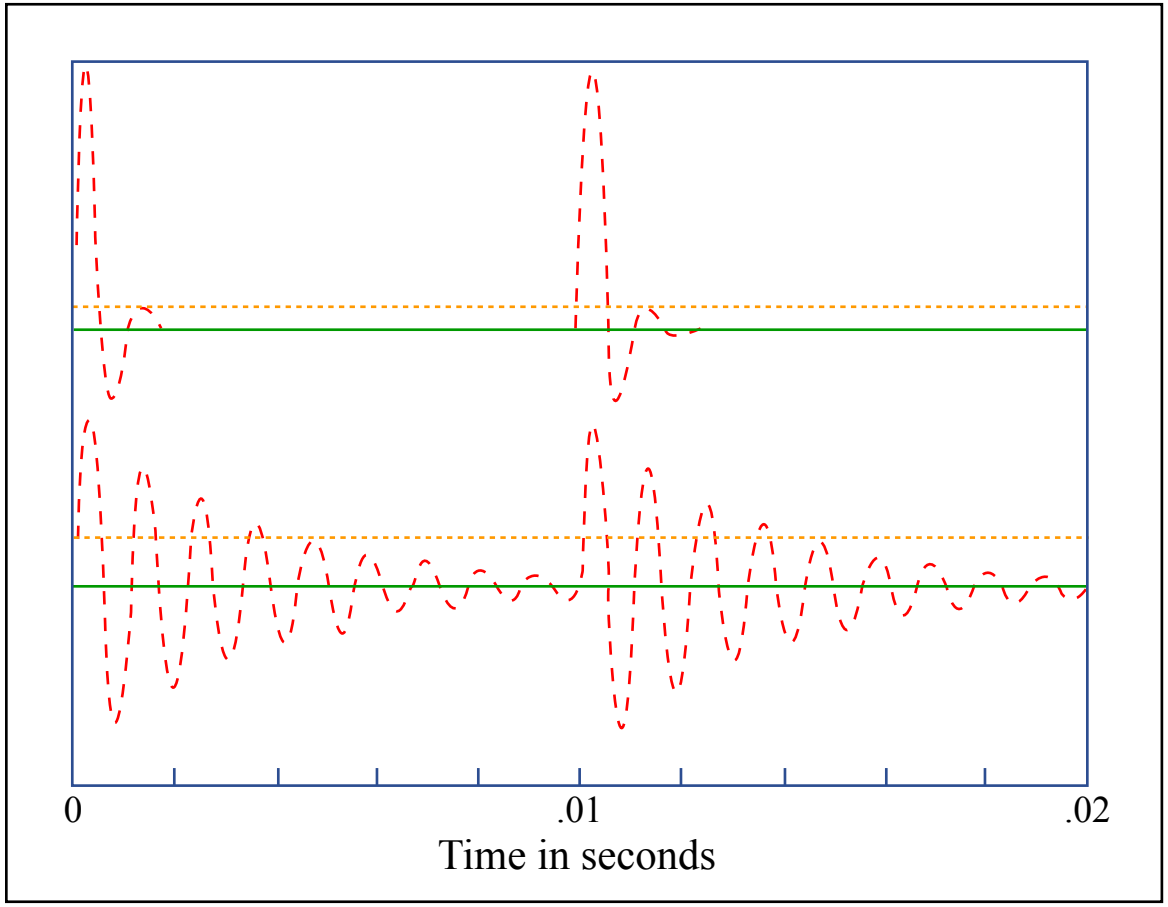


Image by MIT OCW.

Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics.* Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

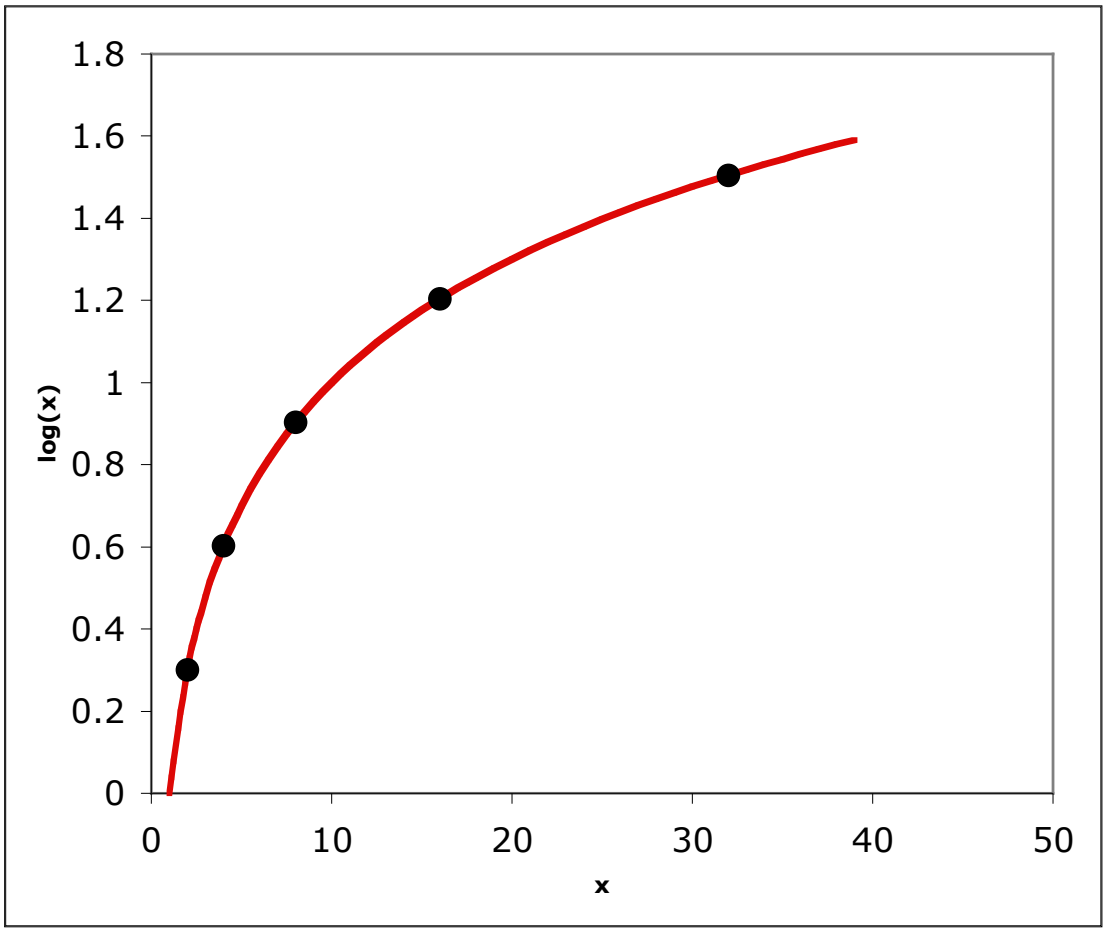
Intensity

- Perceived loudness is more closely related to intensity (power per unit area), which is proportional to the square of the amplitude.
- relative intensity in Bels = $\log_{10}(x^2/r^2)$
- relative intensity in dB $= 10 \log_{10}(x^2/r^2)$
  $= 20 \log_{10}(x/r)$

- In absolute intensity measurements, the comparison amplitude is usually 20μPa, the lowest audible pressure fluctuation of a 1000 Hz tone (dB SPL).

logarithmic scales

- $\log xy = \log x + \log y$



Loudness

- The relationship between intensity and perceived loudness of a pure tone is not exactly logarithmic.

– Loudness of a pure tone (> 40 dB) in Sones:

$$N = 2^{\frac{(dB-40)}{10}}$$

– Loudness is defined to be 1 sone for a 1000 Hz tone at 40 dB SPL

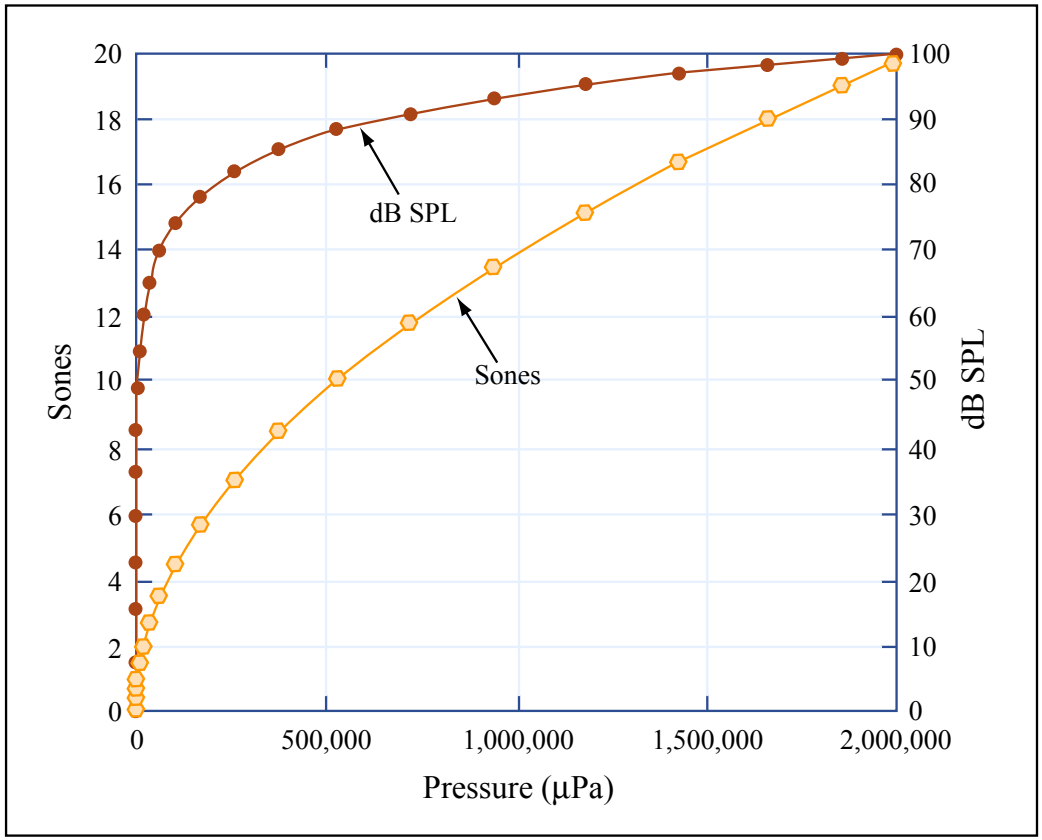
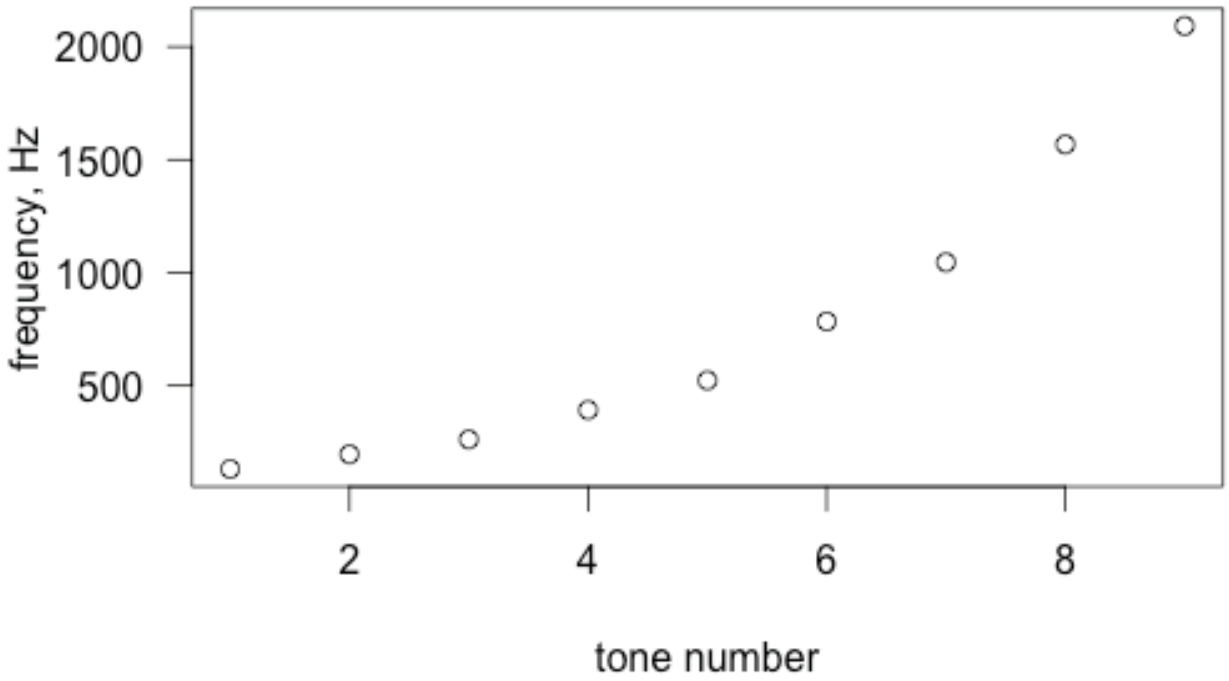


Image by MIT OCW.

Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics.* Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

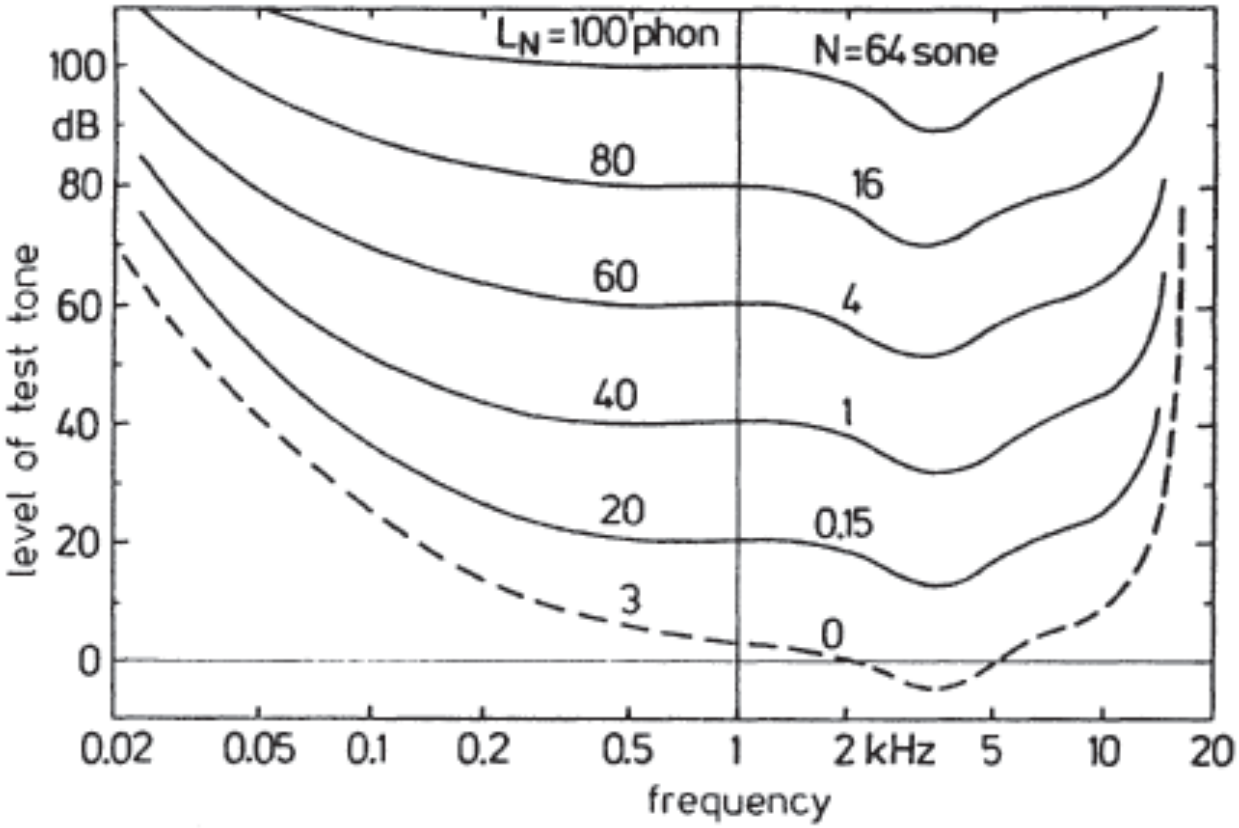
Loudness

- Pure tones: 131-2092 Hz
- Which sounds loudest?



Loudness

- Loudness also depends on frequency.
- equal loudness contours for pure tones:



Loudness

- At short durations, loudness also depends on duration.
- Temporal integration: loudness depends on energy in the signal, integrated over a time window.
- Duration of integration is often said to be about 200ms, i.e. relevant to the perceived loudness of vowels.

Pitch

- Perceived pitch is approximately linear with respect to frequency from 100-1000 Hz, between 1000-10,000 Hz the relationship is approximately logarithmic.

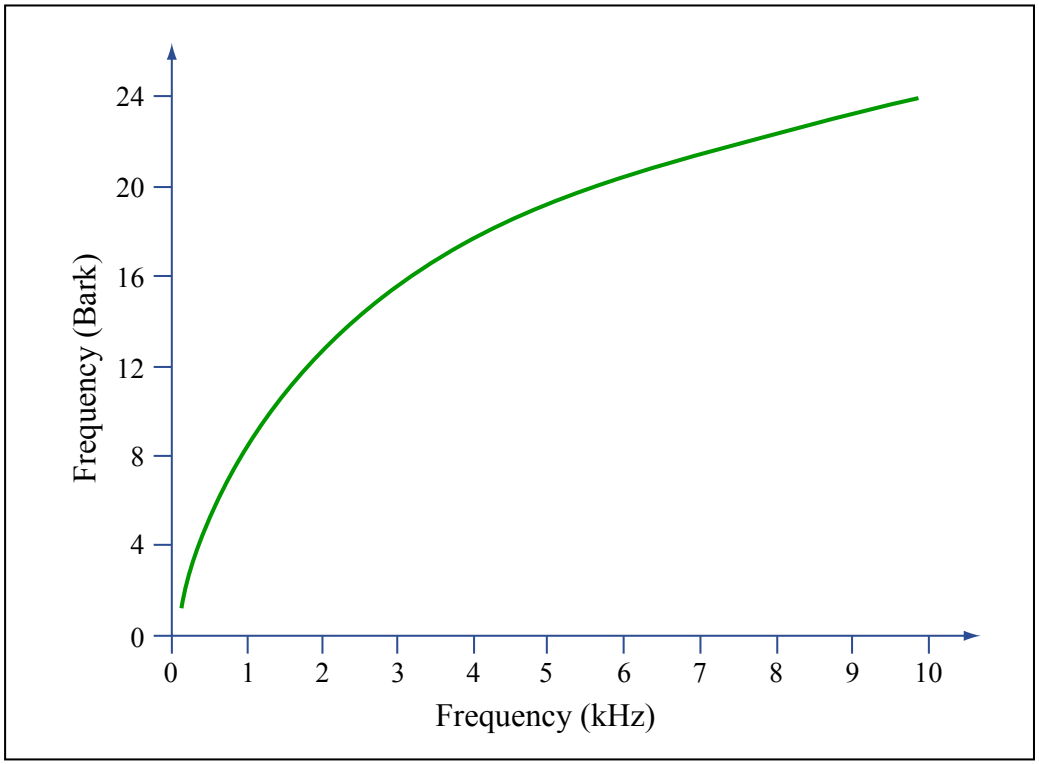


Image by MIT OCW.

Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics.* Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

Pitch

- The non-linear frequency response of the auditory system is related to the physical structure of the basilar membrane.
- basilar membrane ‘uncoiled’:

Diagram of the inner ear removed due to copyright restrictions.

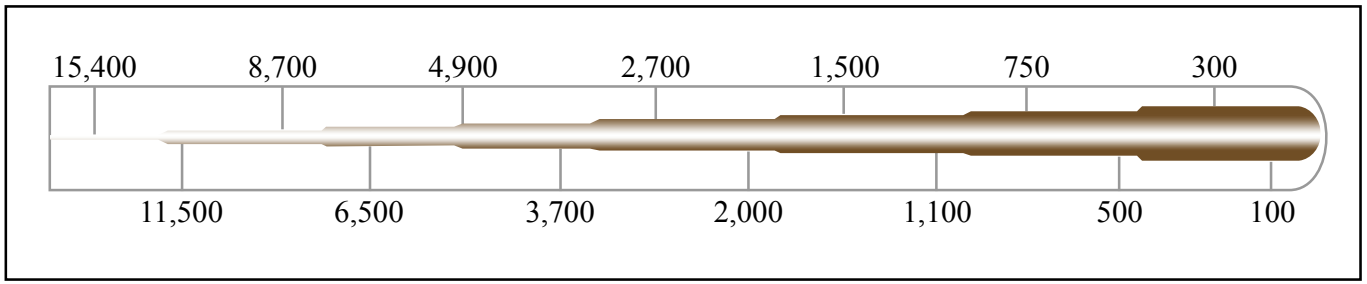


Image by MIT OCW.

Frequency resolution

- In the same way, the structure of the basilar membrane affects the frequency resolution of the auditory 'spectrogram'.
- An acoustic spectrogram represents the variation in intensity over time in a set of frequency bands.
  - E.g. a standard broad-band spectrogram might use frequency bands of 0-200 Hz, 200-400 Hz, 400-600 Hz, etc.
- The ear represents loudness in frequency bands that are narrower at low frequencies and wider at high frequencies.
  - ?-100Hz, 100-200 Hz, …, 400-510 Hz, …1080-1270 Hz, … 12000-15500.
  - These 'critical bands' correspond to a length of about 1.3mm along the basilar membrane (Fastl & Zwicker 2007: 162)

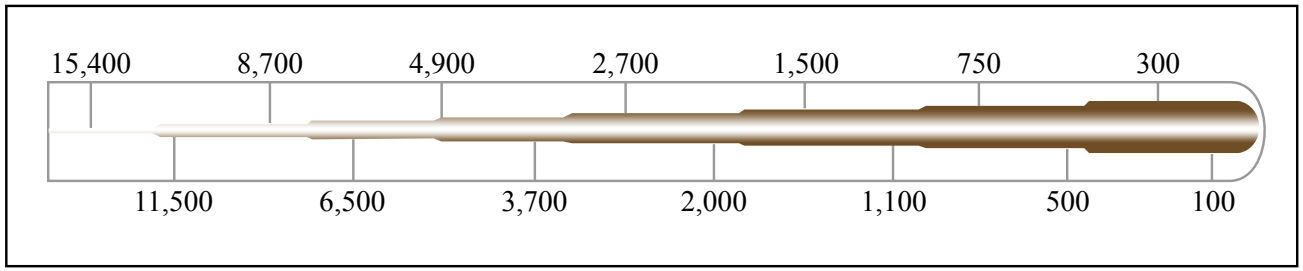


Image by MIT OCW.

Auditory spectra

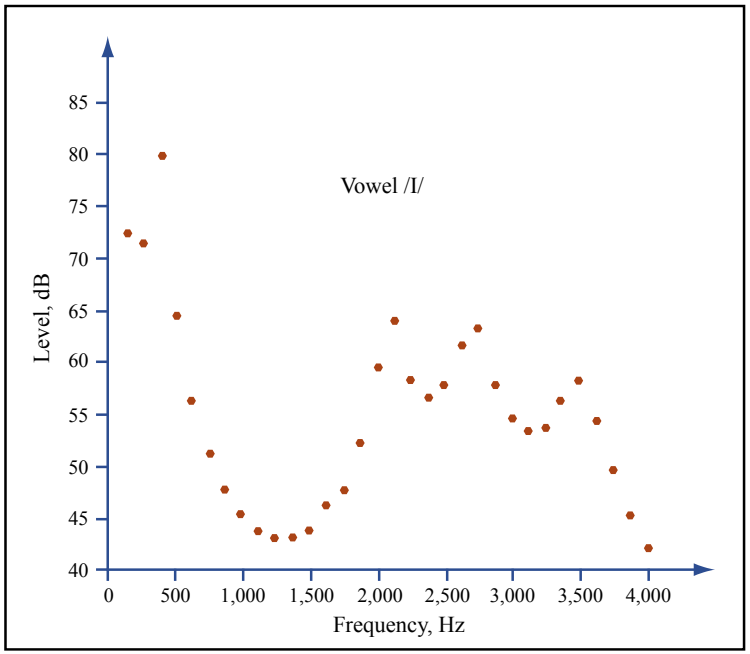


Image by MIT OCW.

Adapted from Moore, Brian. *The Handbook of Phonetic Science.* Edited by William J. Hardcastle and John Laver. Malden, MA: Blackwell, 1997. ISBN: 9780631188483.

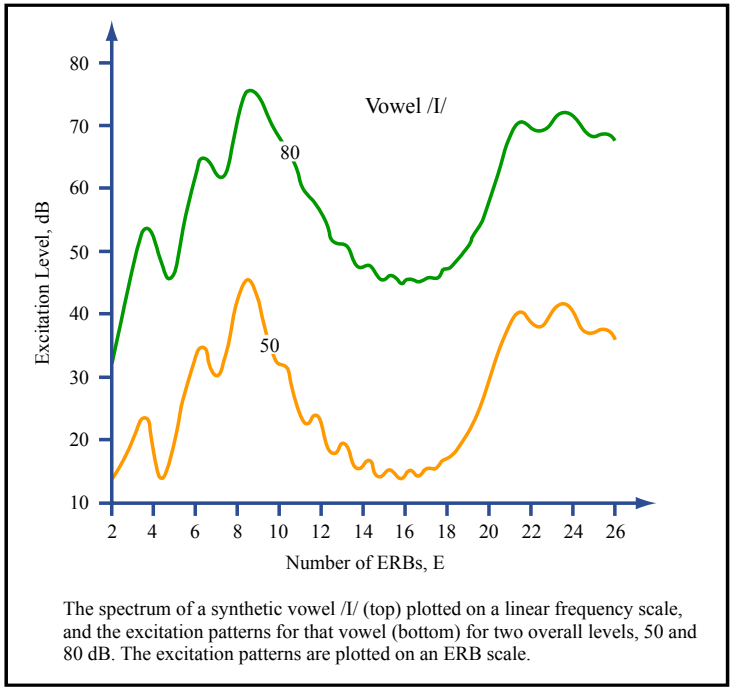
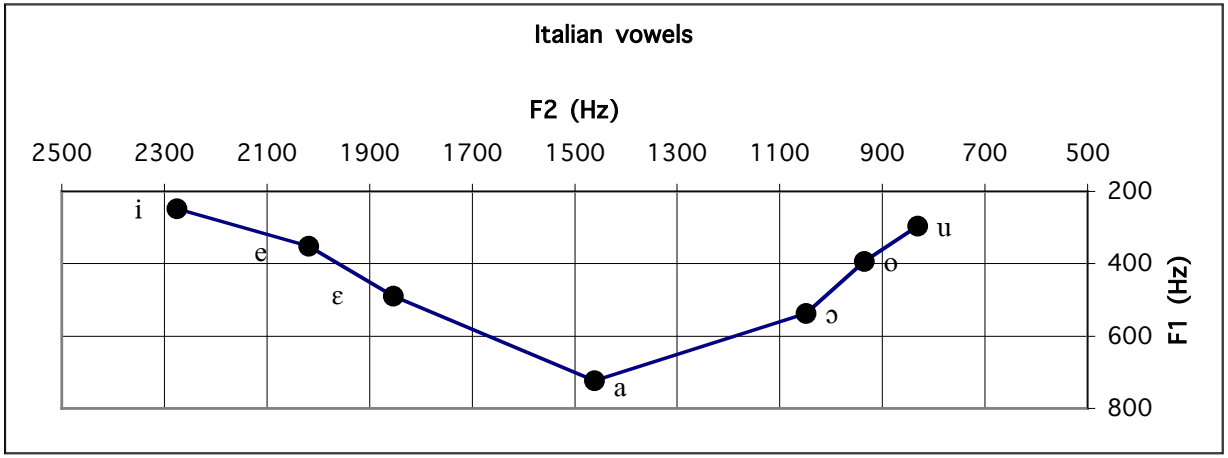
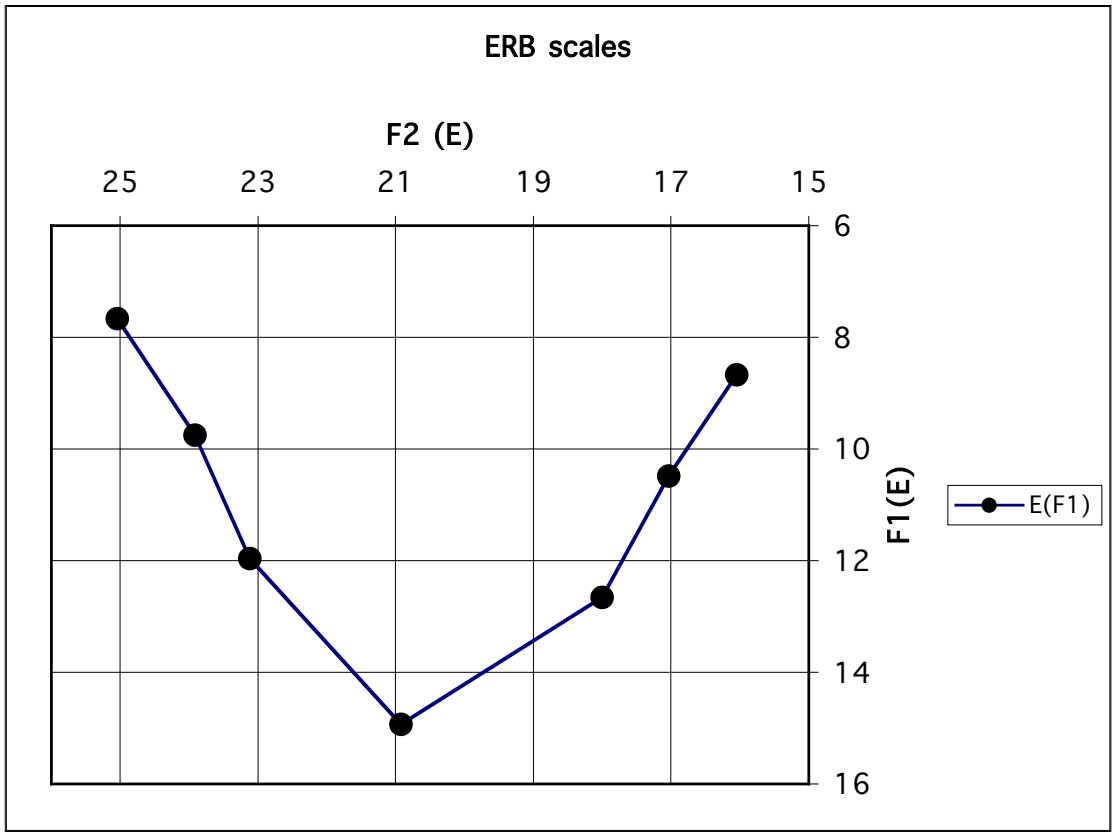


Image by MIT OCW.

Adapted from Moore, Brian. *The Handbook of Phonetic Science.* Edited by William J. Hardcastle and John Laver. Malden, MA: Blackwell, 1997. ISBN: 9780631188483.


Italian vowels
F2 (Hz)
2500
2300
2100
1900
1700
1500
1300
1100
900
700
500
200
400
600
800
F1 (Hz)
i
e
ɛ
a
ɔ
o
u


ERB scales
F2 (E)
25
23
21
19
17
15
6
8
10
12
14
16
F1(E)
E(F1)

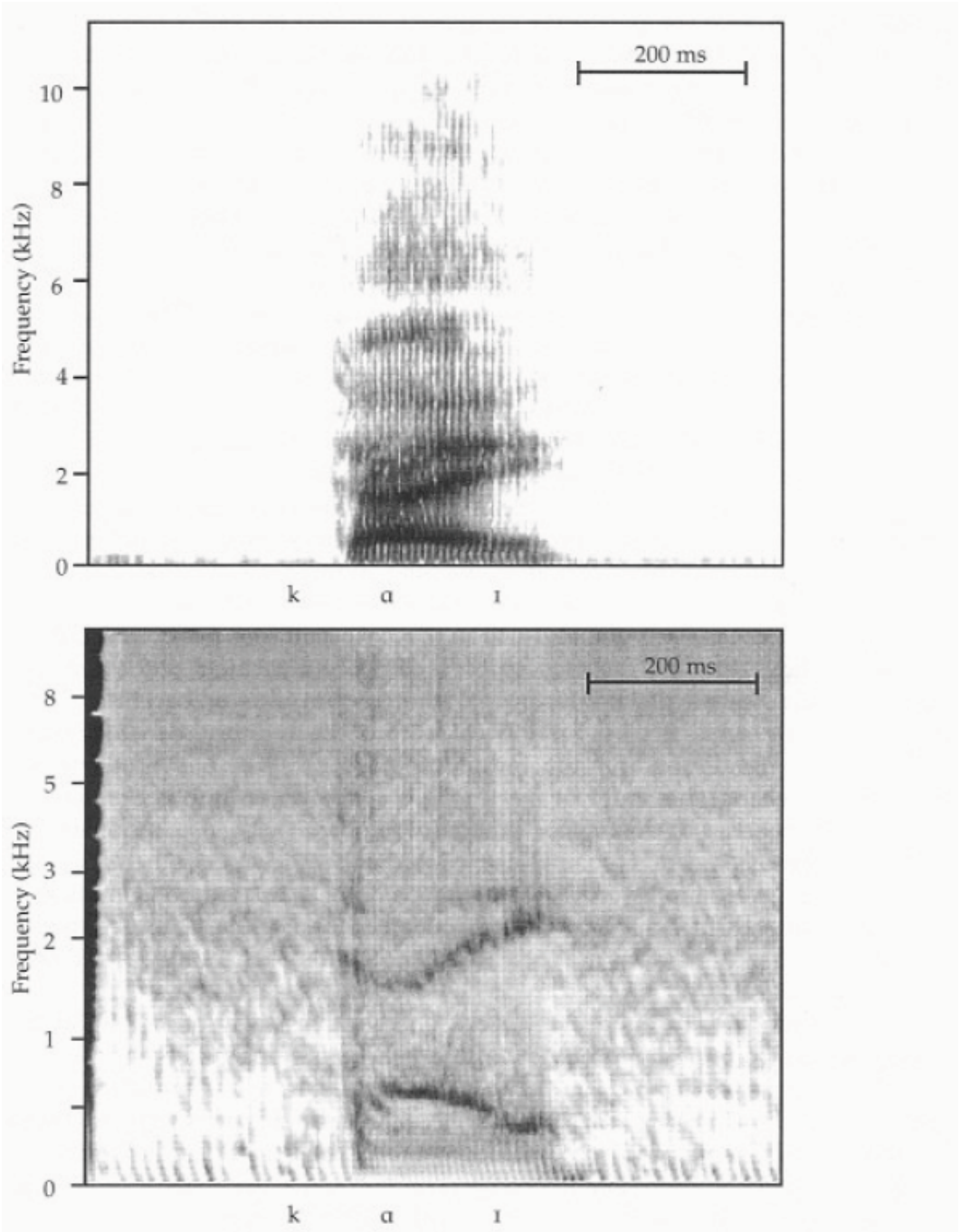


Figure 3.8 Comparison of a normal acoustic spectrogram (top), and an auditory spectrogram, or cochleagram (bottom), of the Cantonese word [kɑˡ] "chicken." The cochleagram was produced by Lyons's (1982) cochlear model.

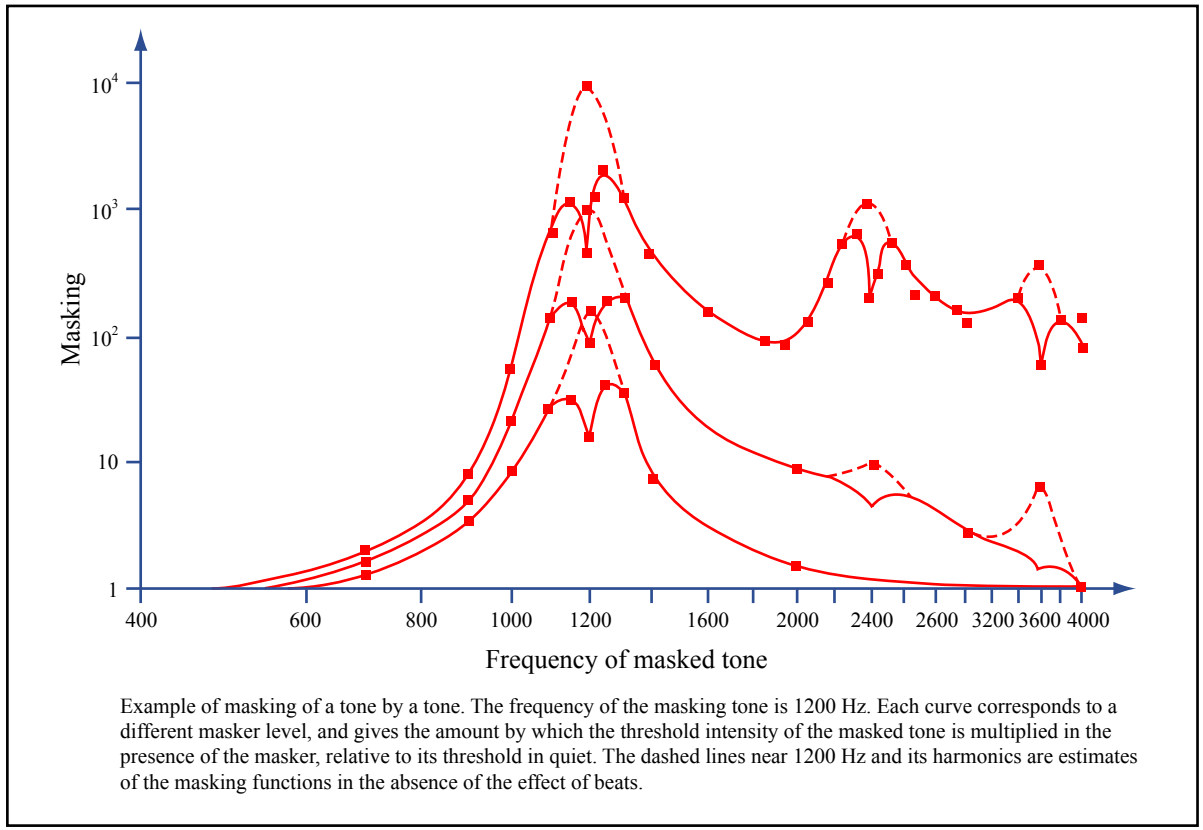
Source: Johnson, K. (2011) "Acoustic and Auditory Phonetics, 3rd ed. Wiley-Blackwell, Oxford.

Masking - simultaneous

- Energy at one frequency can reduce audibility of simultaneous energy at another frequency (masking).
- Single tone, followed by the same tone and a higher-frequency tone, repeated with progressive reduction in the intensity of the higher tone
  - For how many steps can you hear two tones?
- Same pattern, same amplitude tones, but greater difference in the frequencies of the tones
  - For how many steps can you hear two tones?

Masking - simultaneous

- Energy at one frequency can reduce audibility of simultaneous energy at another frequency (masking).



Example of masking of a tone by a tone. The frequency of the masking tone is 1200 Hz. Each curve corresponds to a different masker level, and gives the amount by which the threshold intensity of the masked tone is multiplied in the presence of the masker, relative to its threshold in quiet. The dashed lines near 1200 Hz and its harmonics are estimates of the masking functions in the absence of the effect of beats.

Stevens (1999)

Image by MIT OCW.

Adapted from Stevens, Kenneth N. *Acoustic Phonetics.* Cambridge, MA: MIT Press, 1999. ISBN; 9780262194044.

Time course of auditory nerve response

Response to a noise burst:

- Strong initial response
- Rapid adaptation (~5 ms)
- Slow adaptation (>100ms)
- After tone offset, firing rate only gradually returns to spontaneous level.

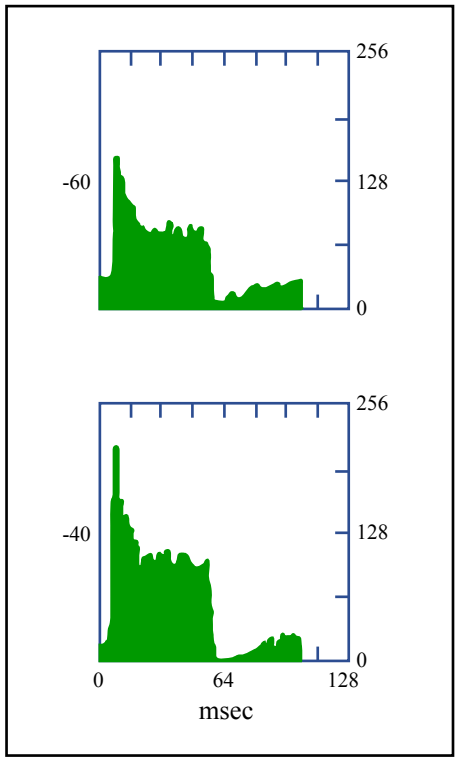


Image by MIT OCW.
Adapted from Kiang et al. (1965)

Kiang et al (1965)

Interactions between sequential sounds

- A preceding sound can affect the auditory nerve response to a following tone (Delgutte 1980).

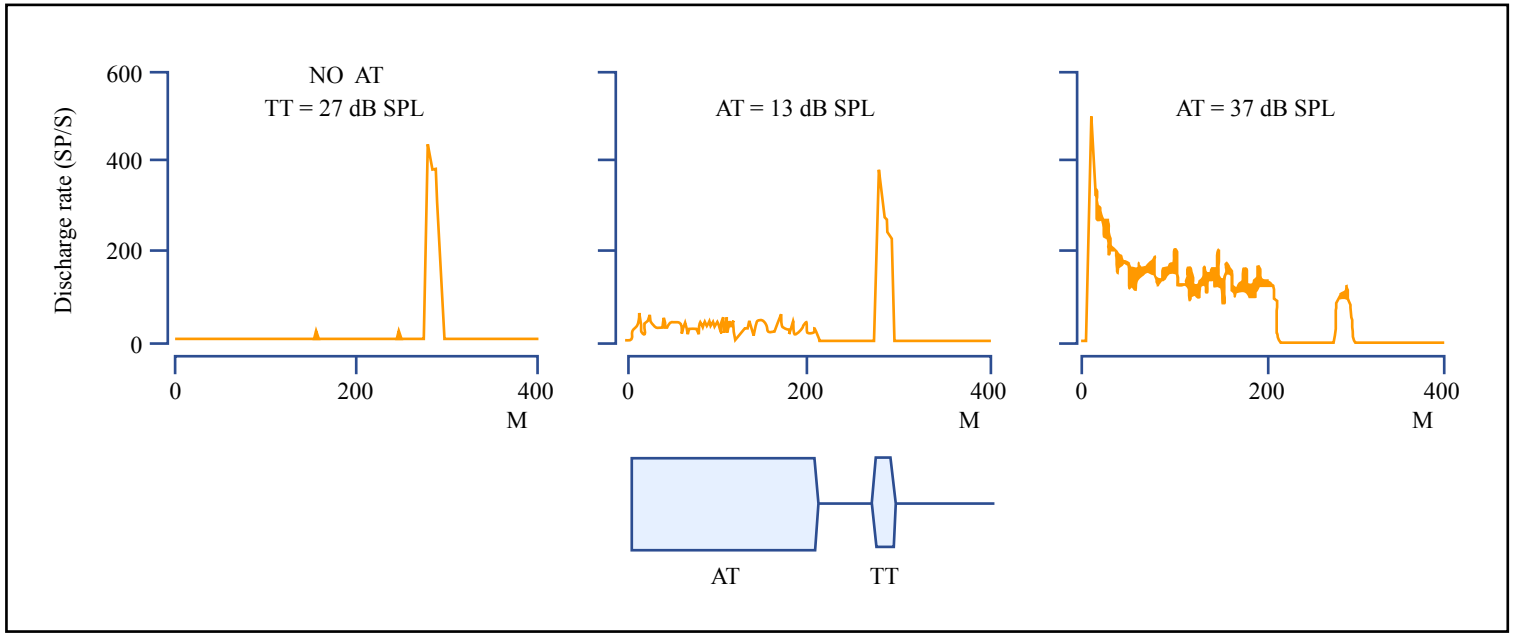
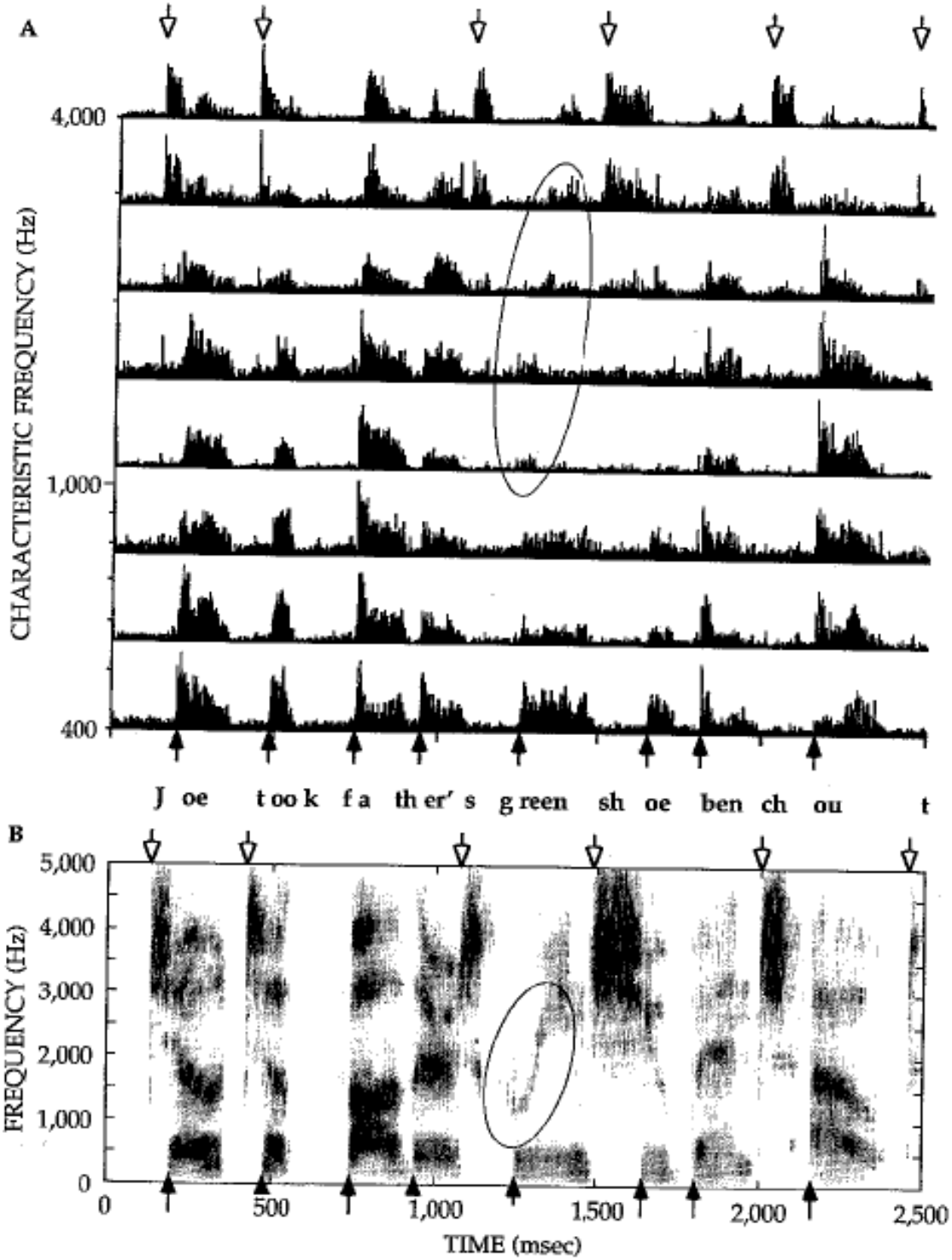


Image by MIT OCW.

Adapted from Stevens, Kenneth N. *Acoustic Phonetics.* Cambridge, MA: MIT Press, 1999. ISBN; 9780262194044, after Delgutte, B. "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers." *Journal of the Acoustical Society of America* 68, no. 3 (1980): 843-857.



Courtesy of the MIT Press from Schnupp, Jan, Israel Nelken, and Andrew King. Auditory neuroscience: Making sense of sound. MIT press, 2011. Used with permission.
Source: Schnupp, Nelken & King (2011) "Auditory Neuroscience: Making Sense of Sound. MIT Press.

24.963

Linguistic Phonetics

Analog-to-digital conversion of speech signals

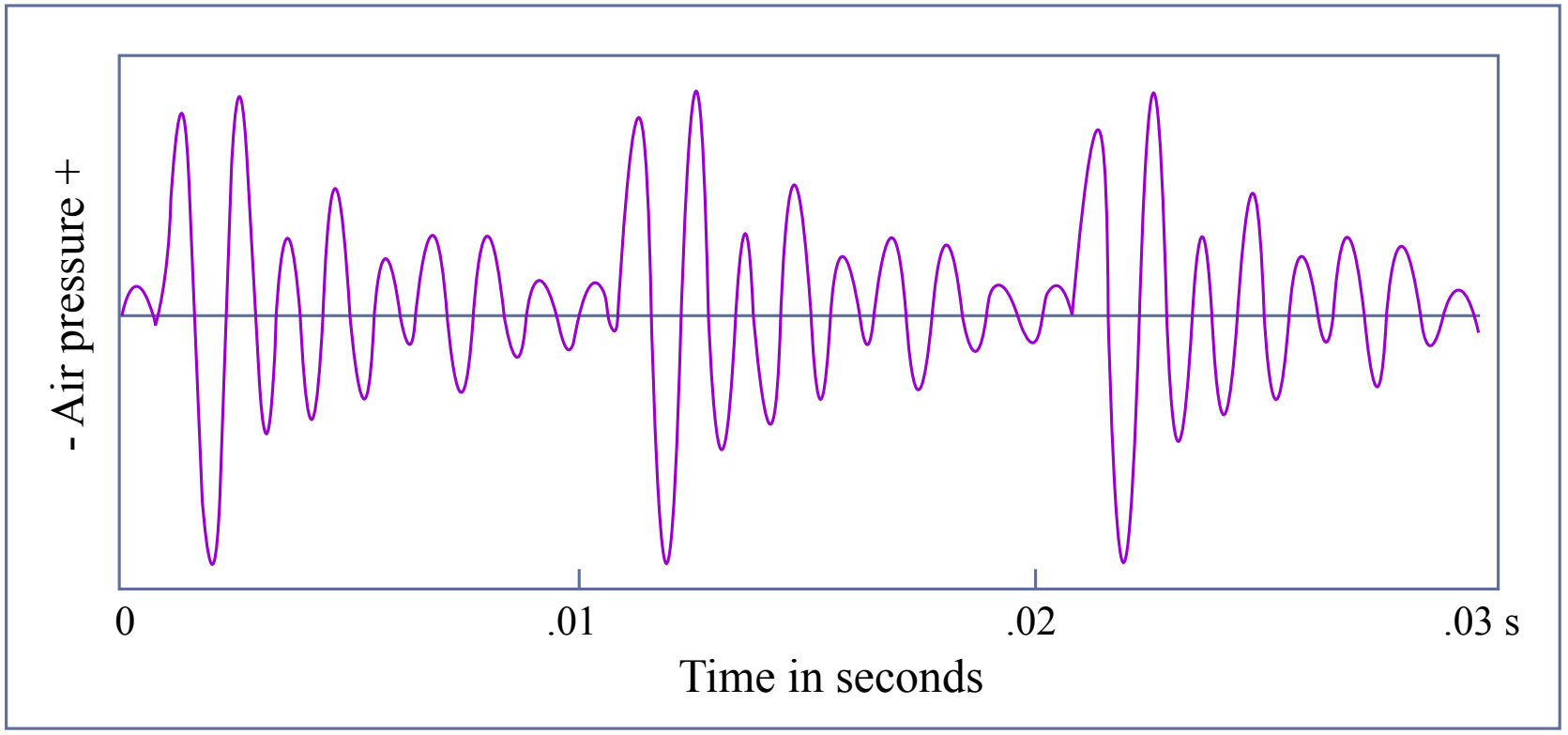


Image by MIT OCW.

Analog-to-digital conversion

- Almost all acoustic analysis is now computer-based.
- Sound waves are analog (or continuous) signals, but digital computers require a digital representation - i.e. a series of numbers, each with a finite number of digits.
- There are two continuous scales that must be divided into discrete steps in analog-to-digital conversion of speech: time and pressure (or voltage).
  - Dividing time into discrete chunks is called **sampling**.
  - Dividing the amplitude scale into discrete steps is called **quantization**.

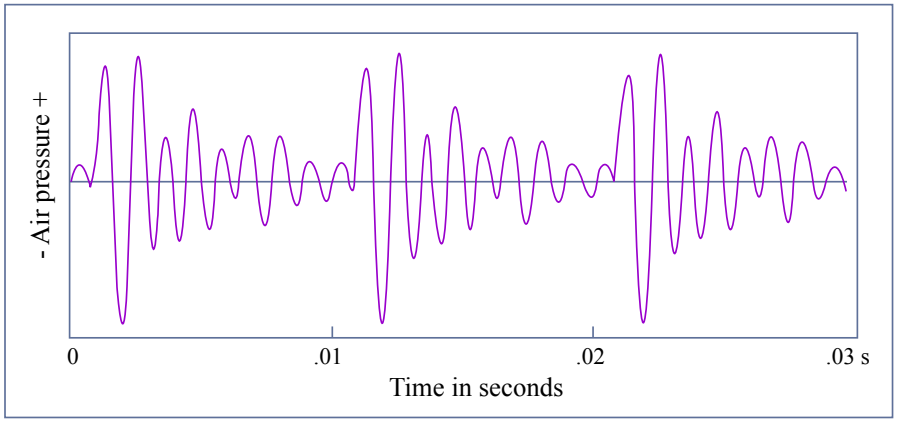


Image by MIT OCW.

Sampling

- The amplitude of the analog signal is sampled at regular intervals.
- The sampling rate is measured in Hz (samples per second).
- The higher the sampling rate, the more accurate the digital representation will be.

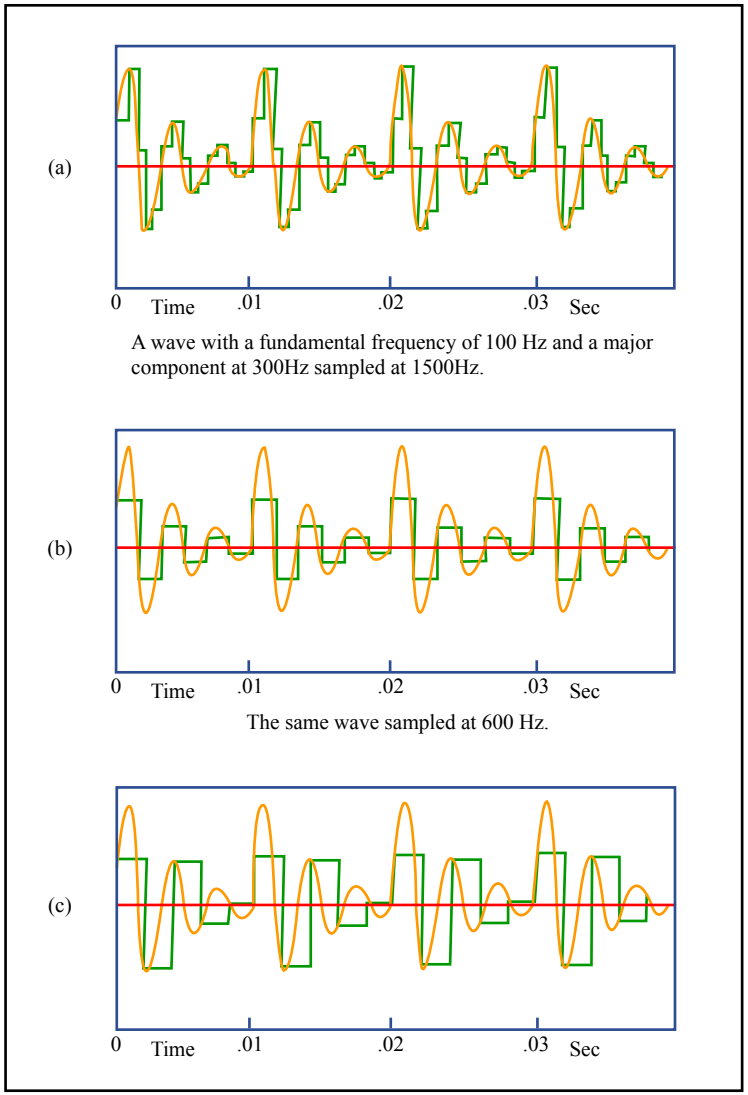


Image by MIT OCW.
Adapted from Ladefoged, Peter. L104/204 Phonetic Theory lecture notes, University of California, Los Angeles.

Sampling

- In order to represent a wave component of a given frequency, it is necessary to sample the signal with at least twice that frequency (the Nyquist Theorem).
- The highest frequency that can be represented at a given sampling rate is called the Nyquist frequency.

- The wave at right has a significant harmonic at 300 Hz
  - (a) sampling rate 1500 Hz
  - (b) sampling rate 600 Hz
  - (c) sampling rate 500 Hz

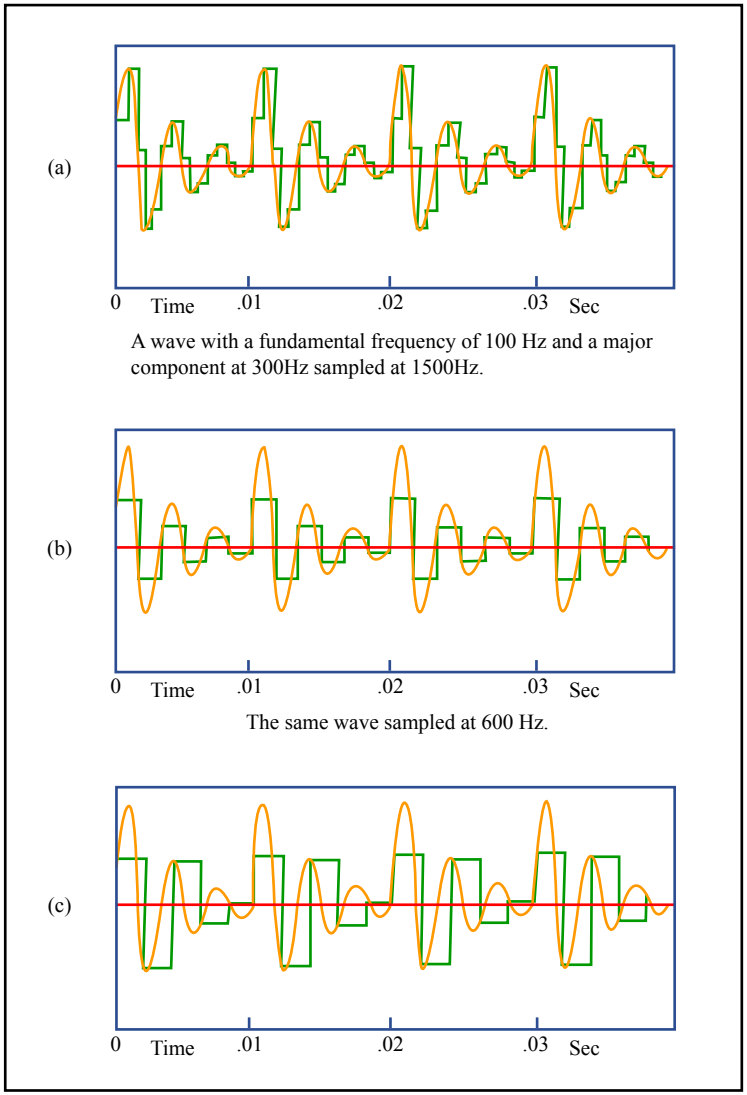


Image by MIT OCW.
Adapted from Ladefoged, Peter. L104/204 Phonetic Theory lecture notes, University of California, Los Angeles.

What sampling rate should you use?

- The highest frequency that (young, undamaged) ears can perceive is about 20 kHz, so to ensure that all audible frequencies are represented we must sample at 2×20 kHz = 40 kHz.
- The ear is relatively insensitive to frequencies above 10 kHz, and almost all of the information relevant to speech sounds is below 10 kHz, so high quality sound is still obtained at a sampling rate of 20 kHz.
- There is a practical trade-off between fidelity of the signal and memory, but memory is getting cheaper all the time.

What sampling rate should you use?

- For some purposes (e.g. measuring vowel formants), a high sampling rate can be a liability, but it is always possible to **downsample** before performing an analysis.

- Audio CD uses a sampling rate of 44.1 kHz.
- Many A-to-D systems only operate at fractions of this rate (44100 Hz, 22050 Hz, 11025 Hz).
- For most purposes, use a sampling rate of 44.1 kHz.

Aliasing

- Components of a signal which are above the Nyquist frequency are misrepresented as lower frequency components (**aliasing**).
- To avoid aliasing, a signal must be filtered to eliminate frequencies above the Nyquist frequency.
- Since practical filters are not infinitely sharp, this will attenuate energy near to the Nyquist frequency also.

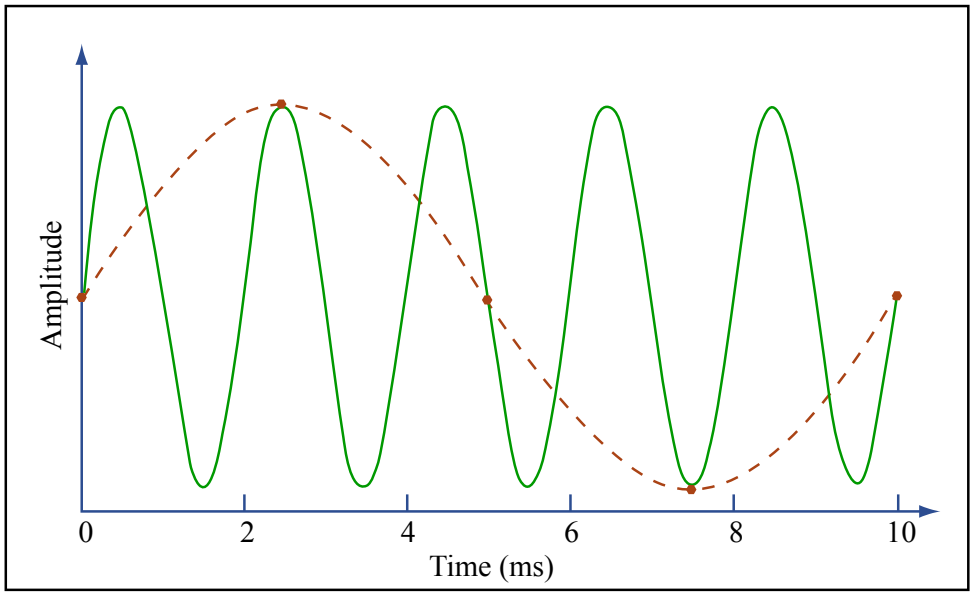


Image by MIT OCW.

Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics.* Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

Quantization

- The amplitude of the signal at each sampling point must be specified digitally - quantization.
- Divide the continuous amplitude scale into a finite number of steps. The more levels we use, the more accurately we approximate the analog signal.

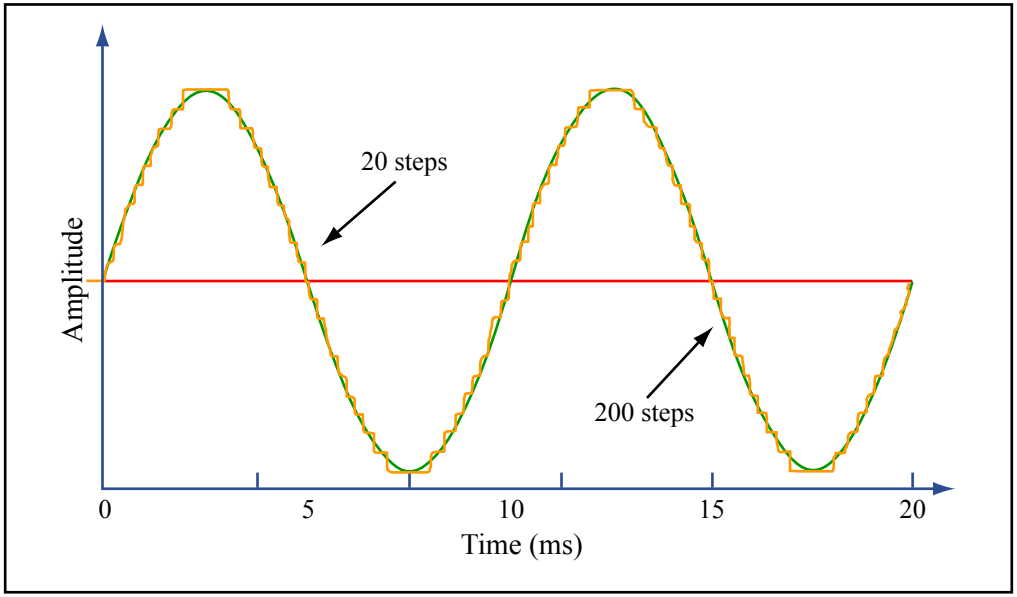


Image by MIT OCW.
Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics.*
Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

Quantization

- The number of levels is specified in terms of the number of bits used to encode the amplitude at each sample.
  - Using n bits we can distinguish 2^n levels of amplitude.
  - e.g. 8 bits, 256 levels.
  - 16 bits, 65536 levels.
- Now that memory is cheap, speech is almost always digitized at 16 bits (the CD standard).

Quantization

- Quantizing an analog signal necessarily introduces quantization errors.
- If the signal level is lower, the degradation in signal-to-noise ratio introduced by quantization noise will be greater, so digitize recordings at as high a level as possible without exceeding the maximum amplitude that can be represented (clipping).
- On the other hand, it is essential to avoid clipping.

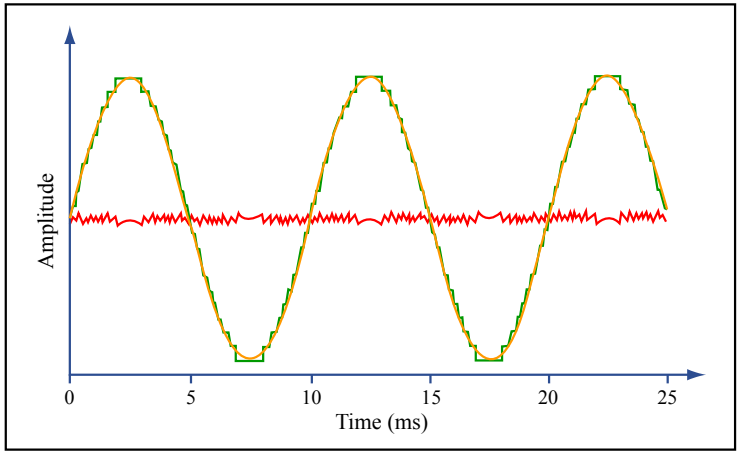


Image by MIT OCW.
Adapted from Johnson, Keith. *Acoustic and Auditory Phonetics.* Malden, MA: Blackwell Publishers, 1997. ISBN: 9780631188483.

24.915 / 24.963 Linguistic Phonetics

Fall 2015