Applied Human Language Technology

Lecture 7
Voiceprints & Acoustic Processing of Speech

Irene Murtagh

This week:

- Introduction to the acoustic processing of speech signals
  (The basis of speech recognition by computers)
- Signal Analysis
- Feature Extraction
- Fourier Analysis and Linear Predictive Coding (LPC)
- Spectral Analysis and Spectra: **Human Voiceprints**
- Sound Waves
- Interpreting a waveform
- Some things to do

Acoustic Processing of Speech

This lecture presents a brief overview of the kind of acoustic processing commonly called **signal analysis** or **feature extraction**.

The term **features** refers to the vector of numbers which represents one time slice of a **speech signal**.

A number of kinds of features are commonly used,
e.g. **LPC** features.

These are **spectral features**, which means that they represent the waveform in terms of the distribution of different frequencies that make up the waveform.

Such a distribution of frequencies is called a **spectrum**.

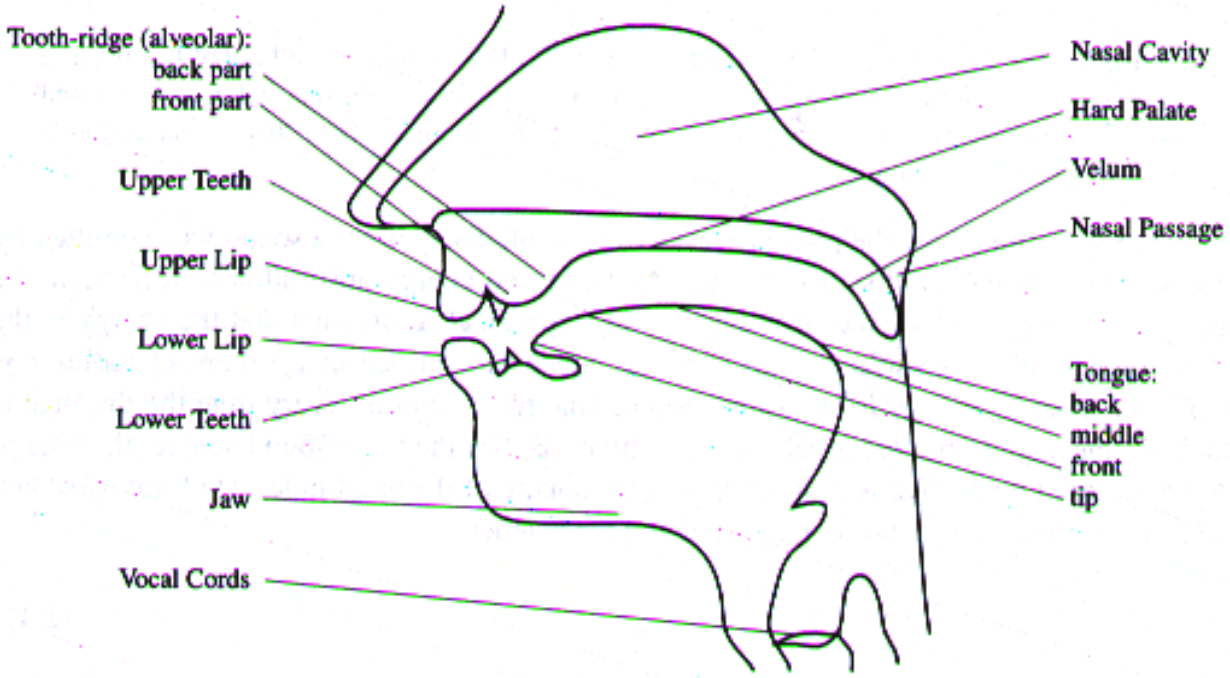
In this lecture we:

summarise the idea of frequency analysis and spectra, and
sketch out different kinds of extracted features.

Sound Waves

The input to a speech recogniser, like the input to the human ear, is a complex series of changes of air pressure.

These changes in air pressure originate with the speaker, and are caused by the specific way that air passes through something called the glottis and out the oral or nasal cavities.



A schematic diagram of the human speech production apparatus.

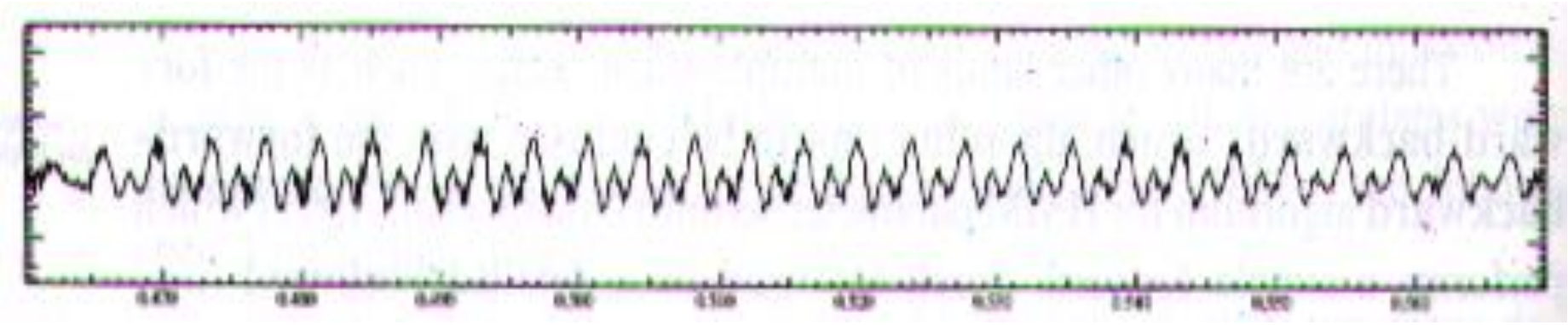
We represent sound waves by plotting the change in air pressure over time.

One way of visualising this is to imagine a graph plot of a vertical plate which is blocking the air pressure waves

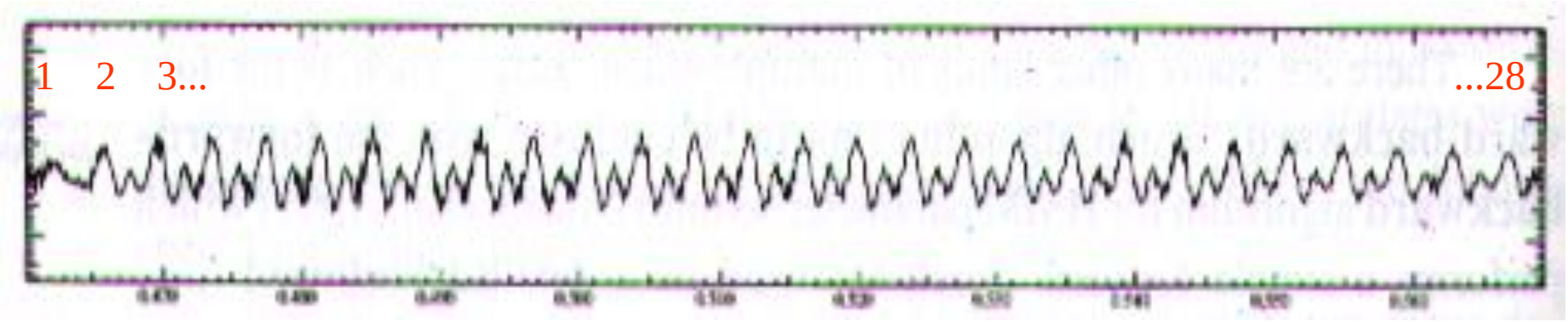
i.e. a microphone in front of the speaker, or the eardrum of the hearer.

The graph measures the amount of compression of the air molecules at this plate.

The diagram following, from the course textbook by Jurafsky and Martin, shows the waveform taken from a corpus of telephone speech of someone saying "*she just had a baby*".



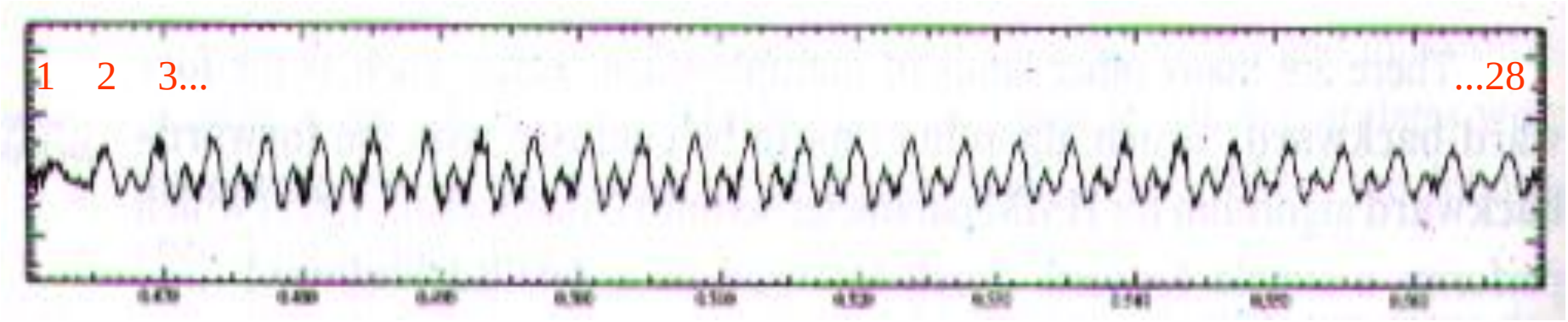
0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56



Two important characteristics of a wave are its:

1. frequency and
2. amplitude.

- The **frequency** is the number of times a second that a wave repeats itself, or more technically, cycles.

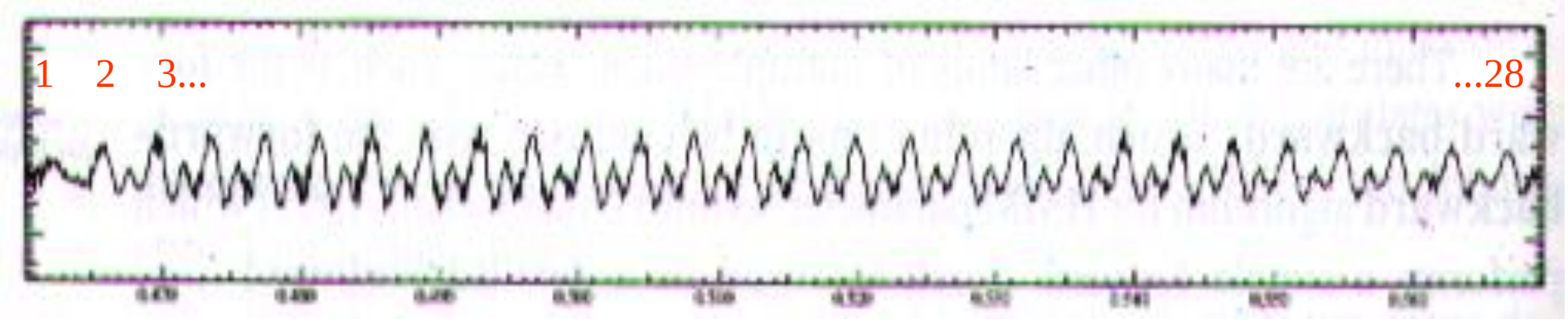


In the diagram there are **28 repetitions** of the wave in the **0.11 seconds** captured.

Therefore, the frequency of this segment of the wave is **28/0.11** or **255 cycles per second**

Frequency = repetitions/time-period

Cycles per second are called **Hertz** (Hz for short). The frequency in the diagram is **255Hz**.



The vertical axis in the diagram measures the amount of air pressure variation.

- A **high value** on the vertical axis (a high **amplitude**) indicates that there is more air pressure at that point in time,
- a **zero value** means that there is normal (atmospheric) pressure,
- a **negative value** means there is lower than normal air pressure.

Two important perceptual properties are related to frequency and amplitude.

The **pitch** of a sound is the perceptual correlate of frequency.

In general, if a **sound has a higher frequency we perceive it as having a higher pitch**, ... but the relationship is not linear since human hearing has different acuities for different frequencies.

Similarly, the loudness of a sound is the perceptual correlate of the power, which is related to the square of the amplitude.

Sounds with higher amplitude are perceived as louder, but again, the relationship is not linear.

How to interpret a waveform

Since humans (... and computers) can transcribe and understand speech just given the sound wave, the waveform must contain enough information to make this task possible.

In most cases, this information is hard to unlock just be looking at the waveform, but ... we can still learn many things by a visual inspection of the waveform.

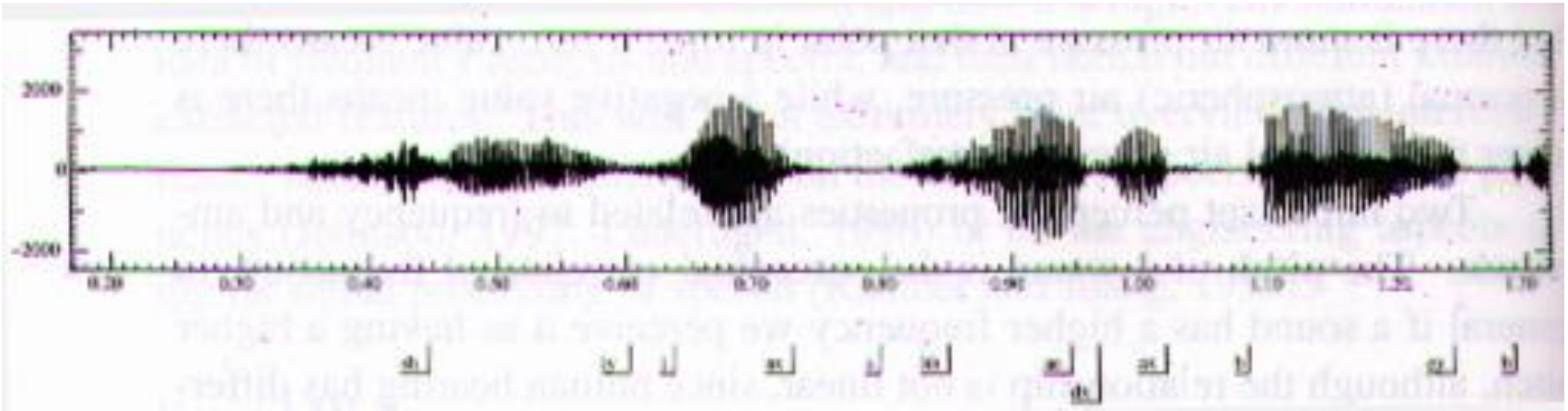
For example, the difference between vowels and consonants of spoken language is quite clear on a waveform

Vowels tend to be long and relatively loud.

Length in time manifests itself as high amplitude.

Fricatives [sh] can also be recognised in a waveform. They produce an intense irregular pattern

The diagram is a waveform by a 20 year female, speaking with an accent from the south midlands of the USA.

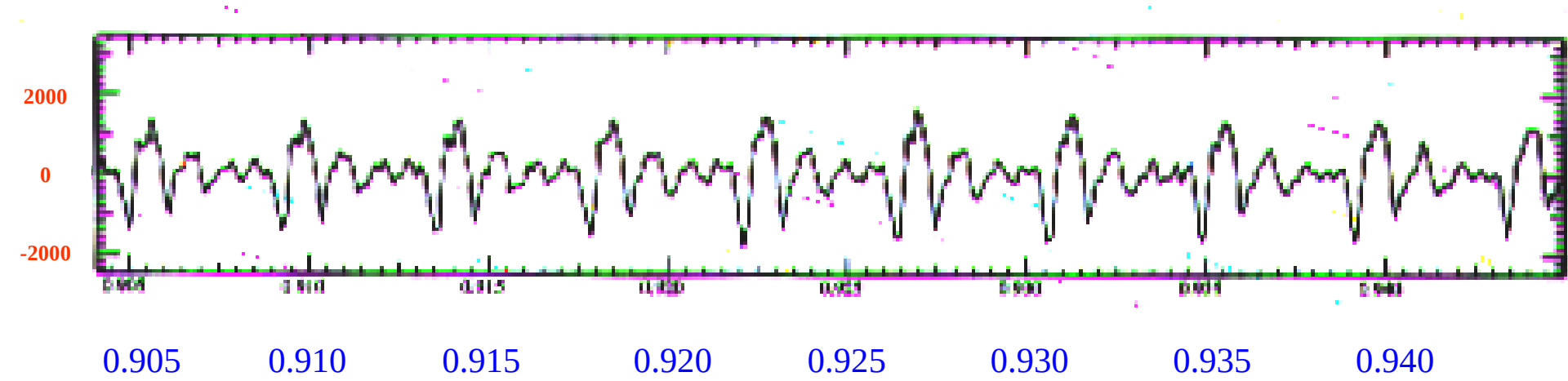


Spectra

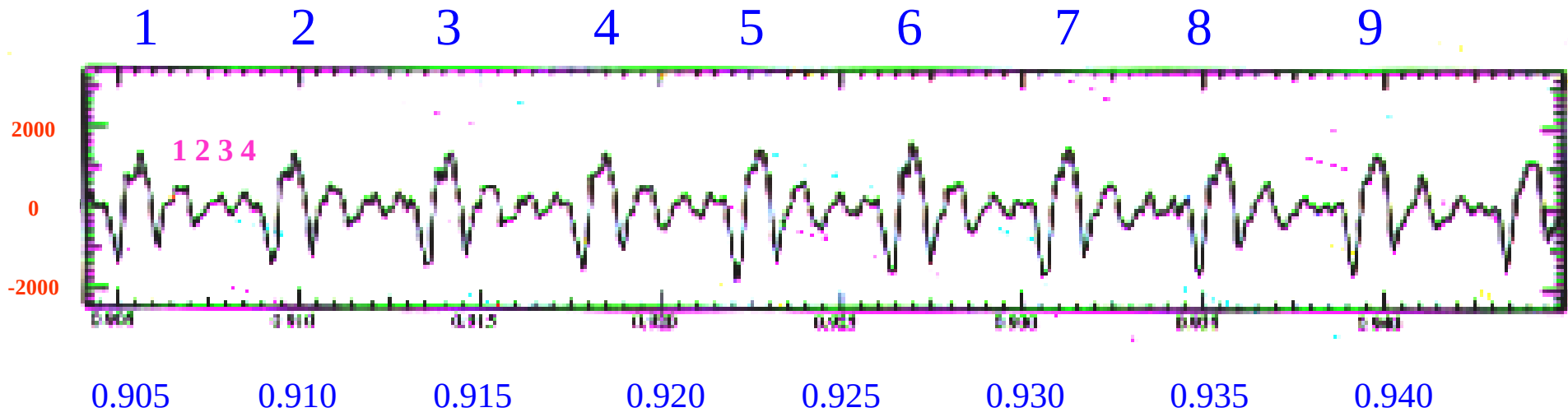
While some broad **phonetic** features can be interpreted from a waveform, more detailed classification requires a different representation of the input in terms of spectral features.

Spectral features are based on the insight of Fourier that every complex wave can be interpreted as a sum of many simple waves of different frequencies.

A musical analogy is a chord. Just as a chord is composed of multiple notes, any waveform is composed of the waves corresponding to its individual “notes”



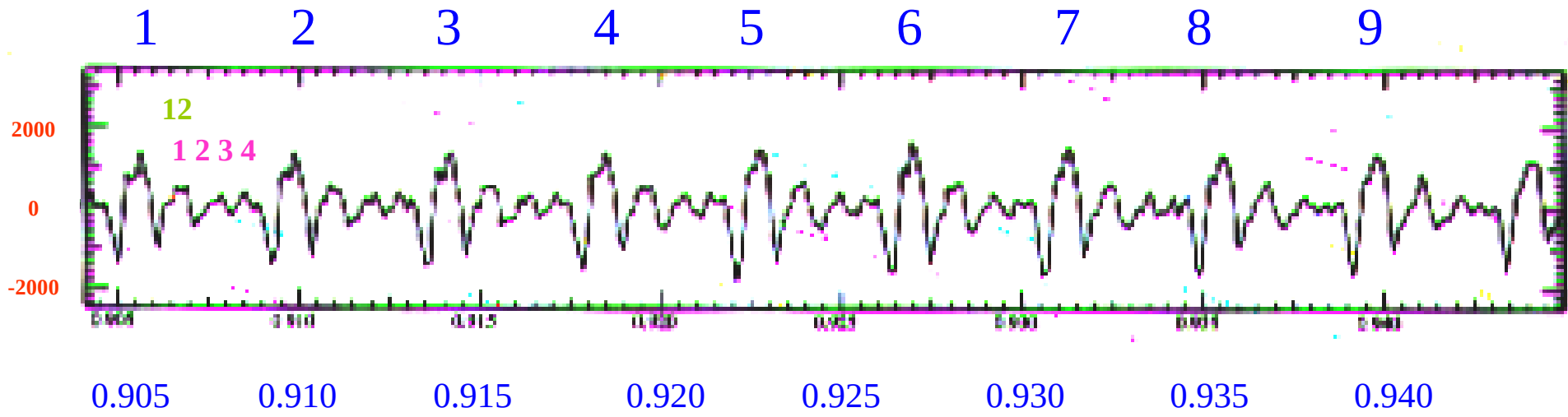
The diagram shows part of the waveform for the vowel [Q] ...in American speech... of the word **had** at second 0.9 of the sentence ÒShe had a babyÓ.



Note however, that there is a complex wave which repeats about 9 times in the diagram.

Note that there is also a **smaller repeated wave which repeats 4 times** for every larger pattern (... look at the 4 small peaks inside every repeated wave).

- The complex wave has a frequency of about 250 Hz.
- We can figure this out since it repeats about 9 times in 0.036 seconds giving 9 cycles/0.036 = **250 Hz**.



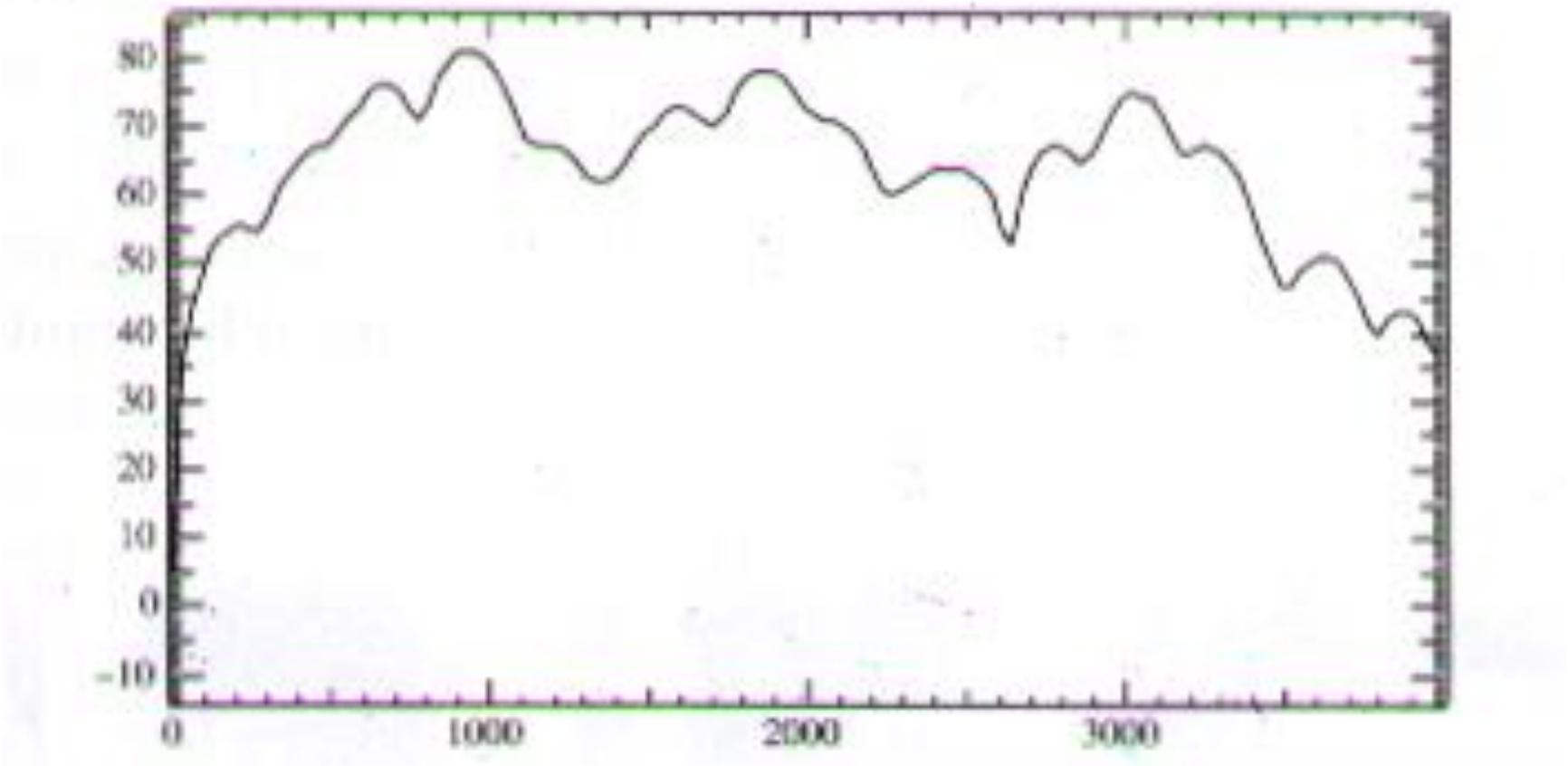
The smaller wave should then have a frequency of approx. 4 times the frequency of the larger wave, say **1000Hz**.

- Then, if you look very closely, you can see two little lines on the peak of many of the 1000Hz waves.
- The frequency of this tiny wave must be approx. 2 times that of the 1000Hz wave = about **2000Hz**.

A spectrum is a representation of these different frequency components of a wave.

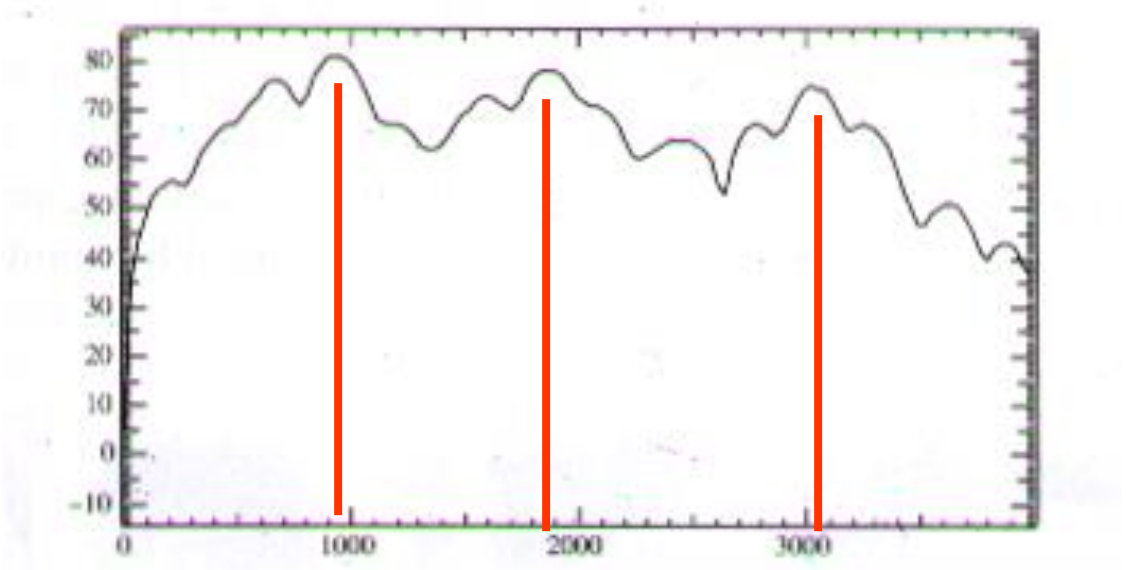
- It can be computed by a Fourier transform, a mathematical procedure which separates out each of the frequency components of a wave.

- Many speech applications use an LPC (Linear Predictive Coding) spectrum because this makes it easier to see where the peaks are.



An LPC spectrum for the vowel [Q] waveform of ÒShe just had a babyÓ at the point in time shown in the previous diagram.

LPCs makes it easier to see **formants**.



The x-axis of a spectrum shows **frequency** while the y-axis shows some measure of the **magnitude** of each frequency component (in decibels **dB**, a logarithmic measure of amplitude).

The diagram shows that there are important frequency components at 930Hz, 1860Hz, and 3020Hz, along with many other lower-magnitude frequency components.

These components at approx. 1000Hz and 2000Hz are what we predicted earlier.

Why is a Spectrum Useful?

It turns out that these spectral peaks that are easily visible in a spectrum are very characteristic of different sounds.

Phones have characteristic spectral “signatures”.

For example;
Chemical elements give off different wavelengths of light when they burn, allowing scientists to detect elements in stars that are light-years away by looking at the spectrum of the light.

Similarly, by looking at the spectrum of a waveform, we can detect the characteristic signature of the different phones that are present.

This use of spectral information is essential to both **human** and **machine speech recognition**.

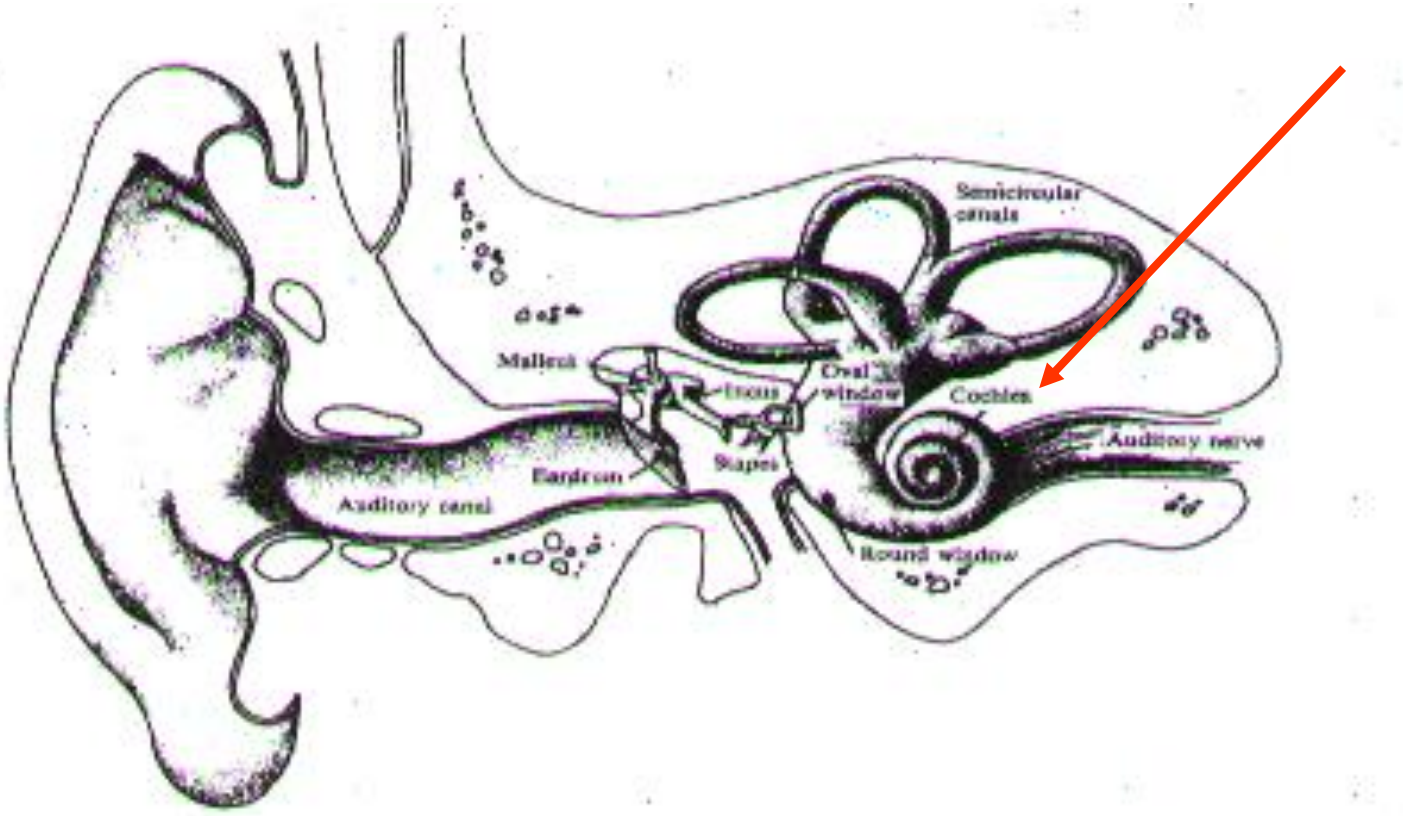
Spectrogram

While a **spectrum** shows the frequency components of a wave at one point in time, ...

a **spectrogram** is a way of envisioning how the different frequencies which make up a waveform change over time.

In human audition, the function of the **cochlea** or **inner ear** is to compute a spectrum of the incoming waveform.

Similarly, the features that are input to **HMMs in speech recognition** are all representations of spectra.

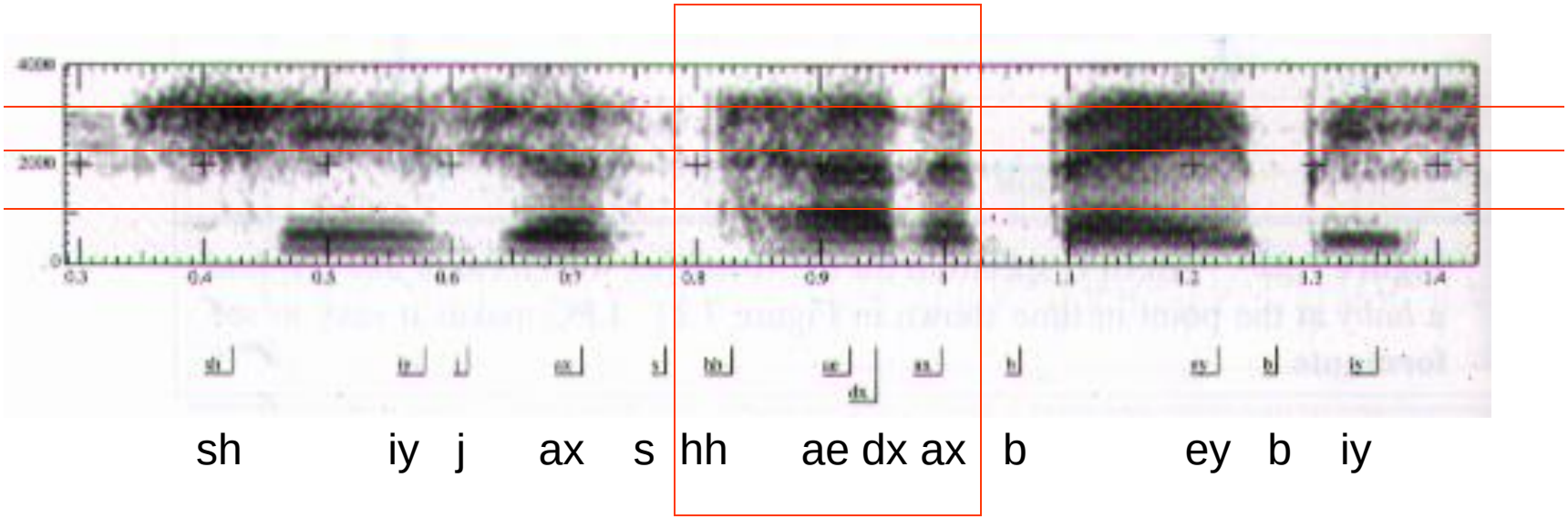


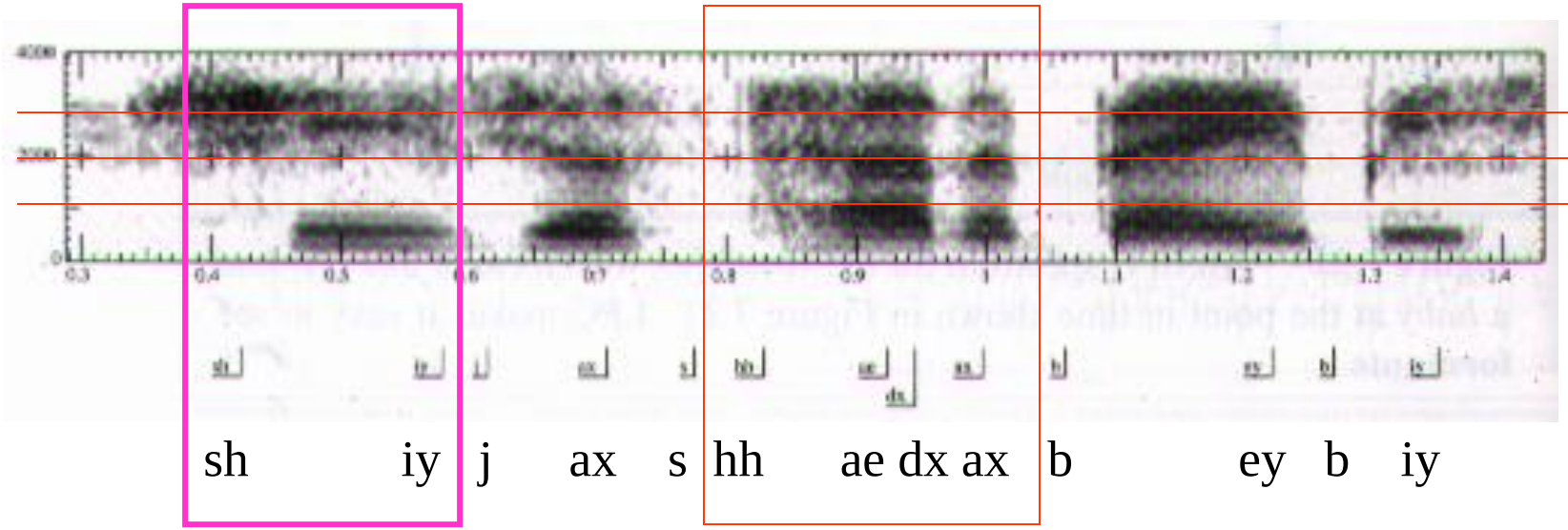
The structure of the peripheral auditory system with the outer, middle, and inner ear.

The x-axis shows time, as it did for the waveform, but the y-axis now shows frequencies in Hertz (Hz).

The darkness of a point on a spectrogram corresponds to the amplitude of the frequency of the component.

For example, in the diagram at the point in time of second 0.9, notice the dark bar at around 1000Hz.





For example, in the diagram at the point in time of second 0.9, notice the dark bar at around 1000Hz.

This means that the vowel [iy] of the word “she” has an important component around 1000Hz.

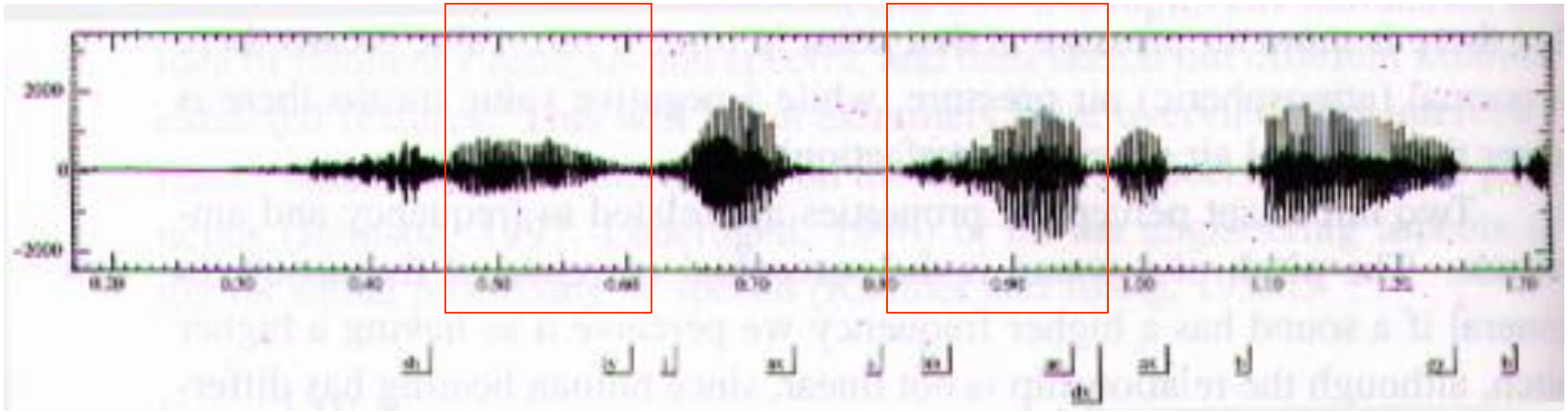
The dark horizontal bars on a spectrogram, representing spectral peaks, usually of vowels, are called **formants**.

What specific clues can spectral representations give for phone identification?

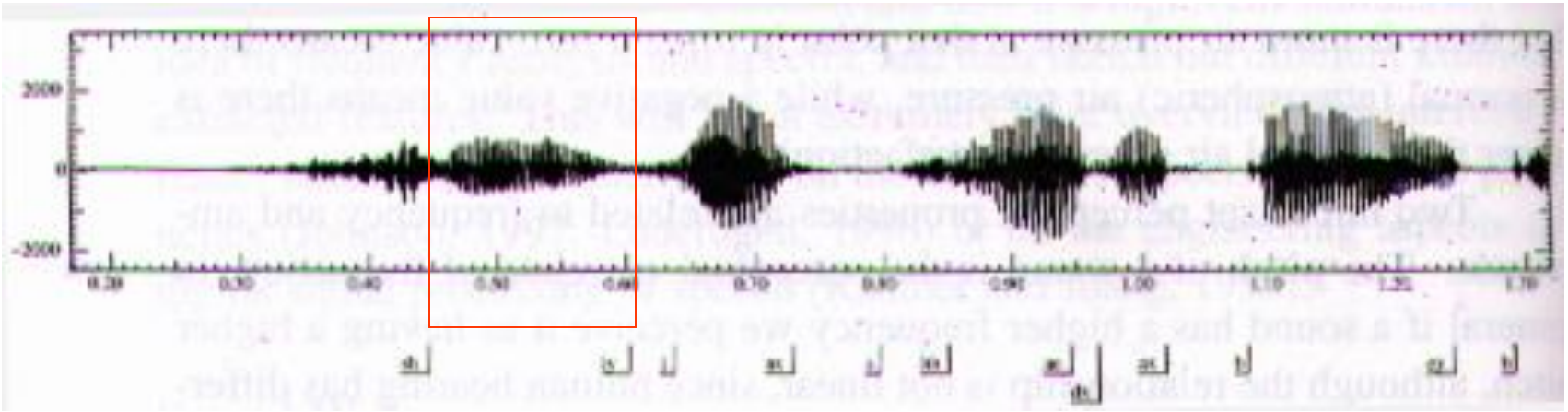
First, different vowels have their formants at characteristic places.

We have seen that the vowel [Q] in the simple waveform had formants at 930Hz, 1860Hz and 3020Hz.

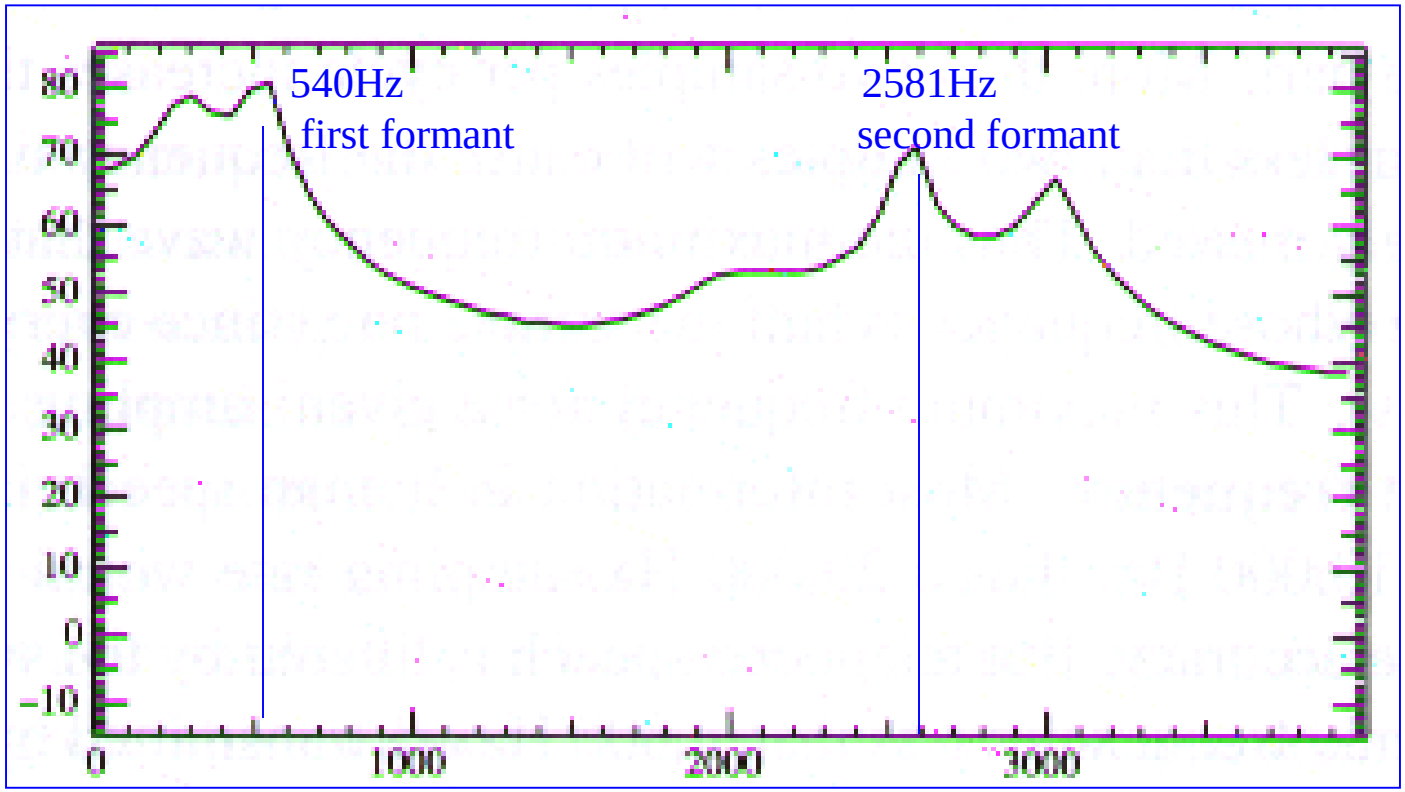
Consider the vowel [iy], at the beginning of the utterance in the first diagram following of Òshe had a babyÓ.



sh iy j ax s hh ae dx ax b ey b



sh iy j ax s hh ae dx ax b ey b



F1 (formant 1) and F2 (formant 2) play a large role in determining vowel identity,

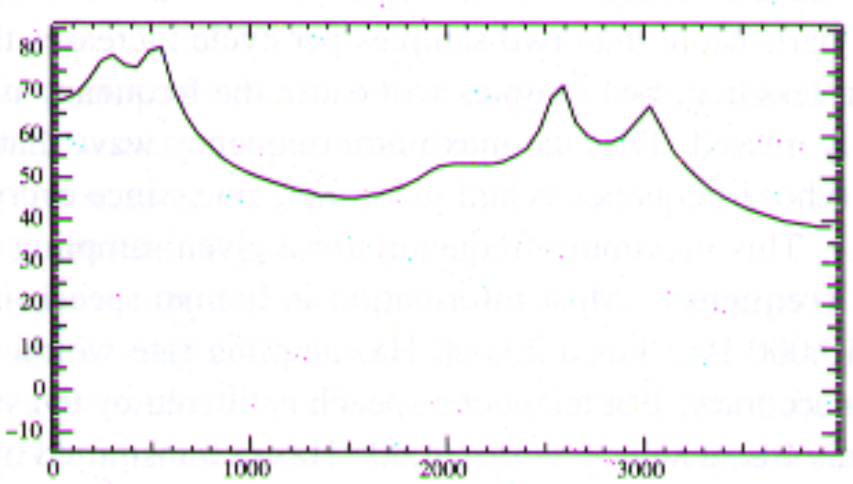
although ....

the formants still differ from speaker to speaker.

Formants can also be used to identify:

the nasal phones **[n] etc..**

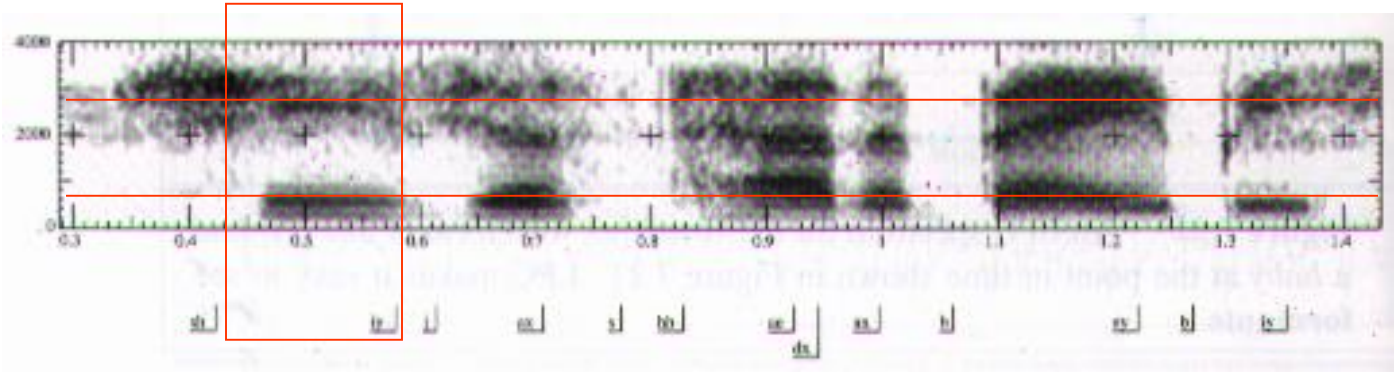
the lateral phone [**l**] etc..



The spectrum for this particular vowel is shown next.

The first formant of [iy] is 540 Hz, much lower than the first formant for [**ae**], while the second formant at 2581 Hz is much higher that the second formant for [**ae**].

We can see these as dark bars on the **spectrogram** diagram

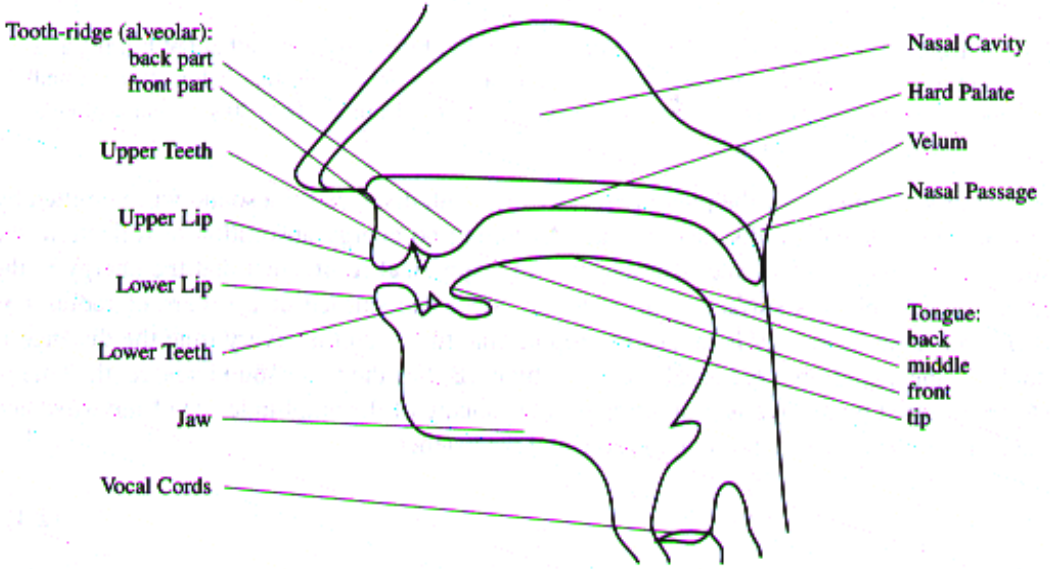


Why do different vowels have different spectra?

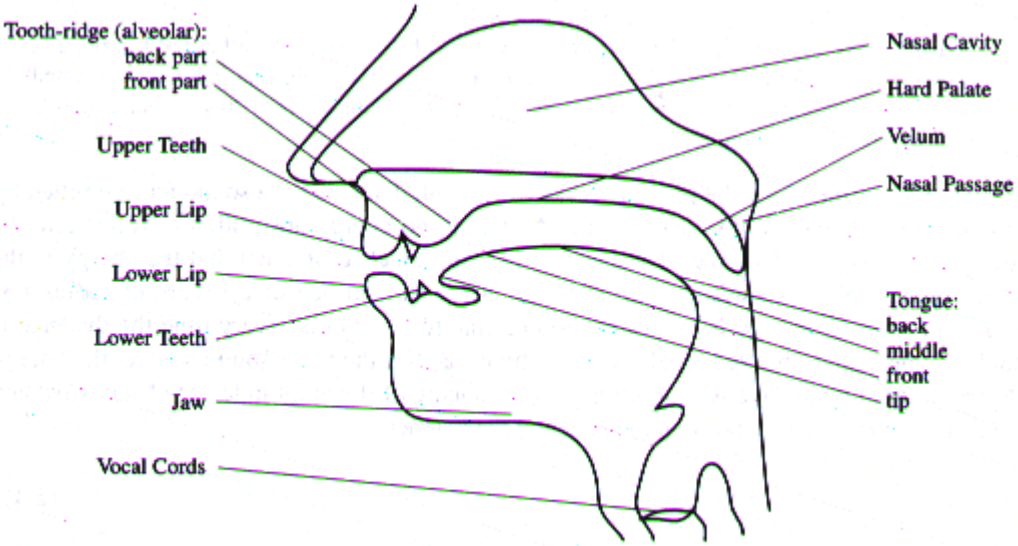
The formants are caused by the resonant cavities of the mouth.

The oral cavity can be thought of as a filter which selectively passes through some of the harmonics of the vocal cord vibrations.

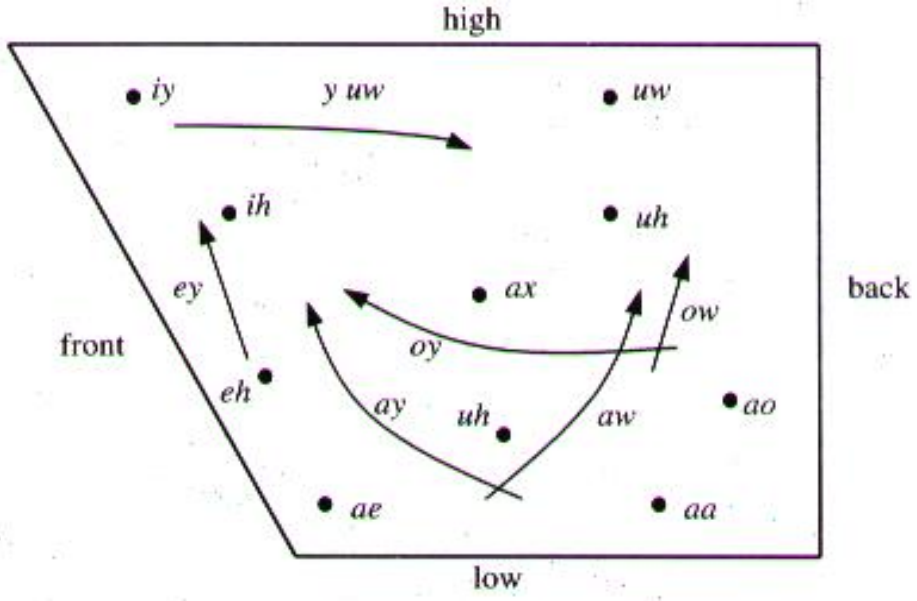
Moving the tongue creates spaces of different size the mouth which selectively amplify waves of the appropriate wavelength, hence amplifying different frequency bands.



A schematic diagram of the human speech production apparatus.



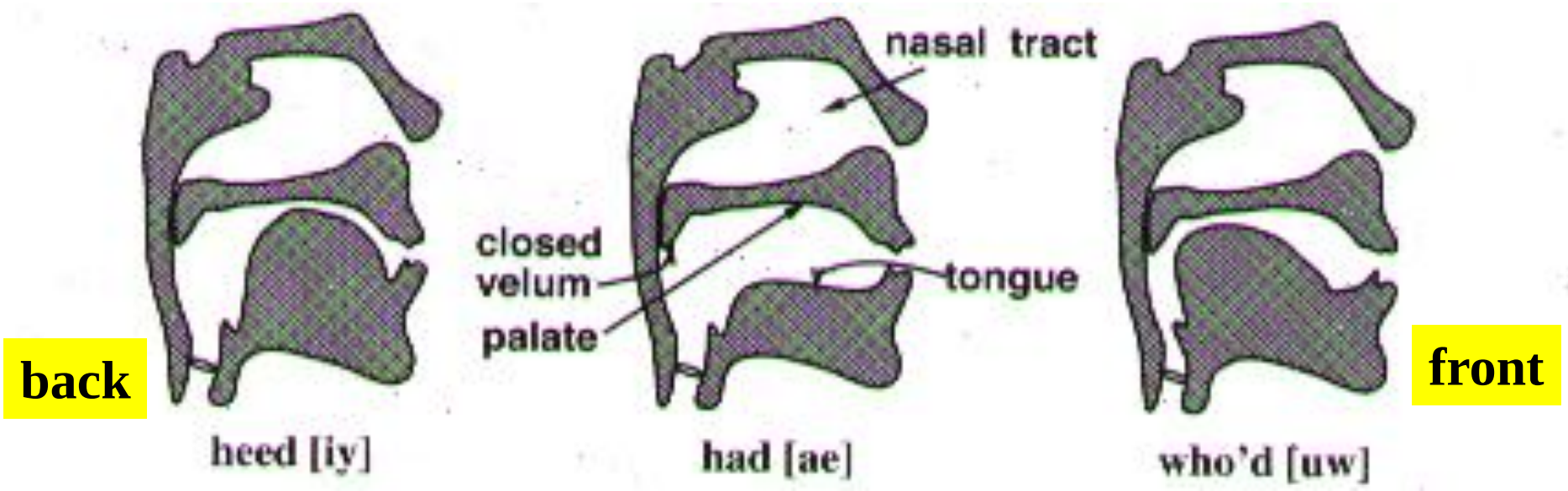
A schematic diagram of the human speech production apparatus.



Relative tongue positions of English vowels

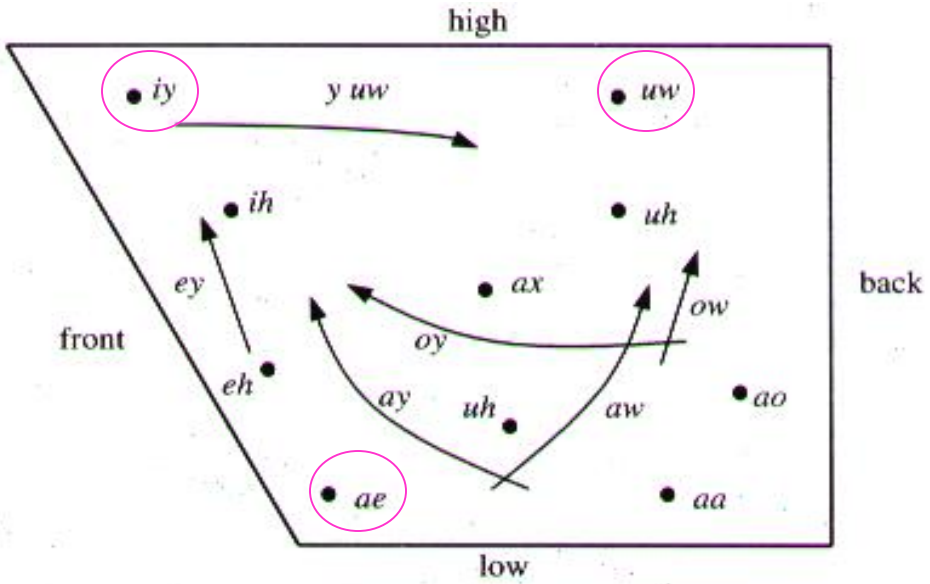
Phonological (abstract) feature decomposition of basic English vowels.

owel	high	low	front	back	round	tense
iy	+	-	+	-	-	+
ih	+	-	+	-	-	-
ae	-	+	+	-	-	+
aa	-	+	-	-	-	+
ah	-	-	-	-	-	+
ao	-	+	-	+	+	+
ax	-	-	-	-	-	-
eh	-	-	+	-	-	-
ow	-	-	-	+	+	+
uh	+	-	-	+	-	-
uw	+	-	-	+	-	+



Positions of the tongue for three English vowels...

high front [iy]
low front [ae] and
high back [uw]



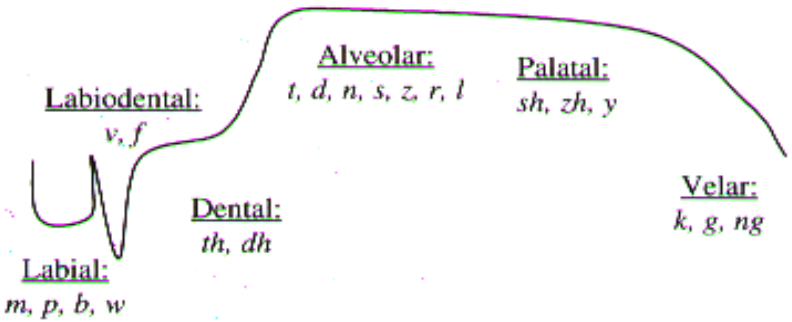
Relative tongue positions of English vowels

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[i]	[iy]	lily	['lɪli]	[l ih l iy]
[ɪ]	[ih]	lily	['lɪli]	[l ih l iy]
[eɪ]	[ey]	daisy	['deɪzi]	[d ey z i]
[ɛ]	[eh]	poinsettia	[pɔɪn'sɛɾiə]	[p oy n s eh dx iy ax]
[æ]	[ae]	aster	['æstɚ]	[ae s t axr]
[ɑ]	[aa]	poppy	['papi]	[p aa p i]
[ɔ]	[ao]	orchid	['ɔrkɨd]	[ao r k ix d]
[ʊ]	[uh]	woodruff	['wʊdrʌf]	[w uh d r ah f]
[oʊ]	[ow]	lotus	['loʊɾəs]	[l ow dx ax s]
[u]	[uw]	tulip	['tulip]	[t uw l ix p]
[ʌ]	[uh]	buttercup	['bʌɾɚ kʌp]	[b uh dx axr k uh p]
[ɝ]	[er]	bird	['bɝd]	[b er d]
[aɪ]	[ay]	iris	['aɪɹɨs]	[ay r ix s]
[aʊ]	[aw]	sunflower	['sʌnflaʊɚ]	[s ah n f l aw axr]
[ɔɪ]	[oy]	poinsettia	[pɔɪn sɛɾiə]	[p oy n s eh dx iy ax]
[ju]	[y uw]	feverfew	[fivɚfju]	[f iy v axr f y u]
[ə]	[ax]	woodruff	['wʊdrəf]	[w uh d r ax f]
[ɨ]	[ix]	tulip	['tulɨp]	[t uw l ix p]
[ɚ]	[axr]	heather	['hɛðɚ]	[h eh dh axr]
[ʉ]	[ux]	dude[2]	[dʉd]	[d ux d]

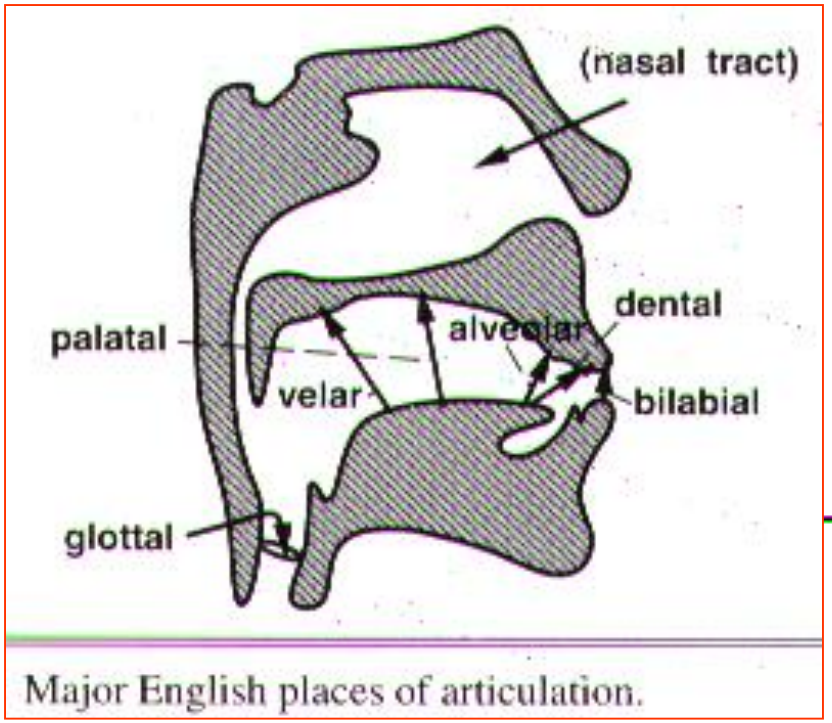
IPA and **ARPAbet** symbols for transcription of English vowels

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[p]	[p]	parsley	['parsli]	[p aa r s l iy]
[t]	[t]	tarragon	['tærəgɑn]	[t ae r ax g aa n]
[k]	[k]	catnip	['kætnɪp]	[k ae t n ix p]
[b]	[b]	bay	[beɪ]	[b ey]
[d]	[d]	dill	[dɪl]	[d ih l]
[g]	[g]	garlic	['gɑrlɪk]	[g aa r l ix k]
[m]	[m]	mint	[mɪnt]	[m ih n t]
[n]	[n]	nutmeg	['nʌtmɛg]	[n ah t m eh g
[ŋ]	[ng]	ginseng	['dʒɪnsɪŋ]	[jh ih n s ix ng]
[f]	[f]	fennel	['fɛnl]	[f eh n el]
[v]	[v]	clove	[kloʊv]	[k l ow v]
[θ]	[th]	thistle	['θɪsl̩]	[th ih s el]
[ð]	[dh]	heather	['hɛðɚ]	[h eh dh axr]
[s]	[s]	sage	[seɪdʒ]	[s ey jh]
[z]	[z]	hazelnut	['heɪzl̩nʌt]	[h ey z el n ah t]
[ʃ]	[sh]	squash	[skwɑʃ]	[s k w a sh]
[ʒ]	[zh]	ambrosia	[æm'broʊʒə]	[ae m b r ow zh ax]
[tʃ]	[ch]	chicory	['tʃɪkɚi]	[ch ih k axr iy]
[dʒ]	[jh]	sage	[seɪdʒ]	[s ey jh]
[l]	[l]	licorice	['lɪkɚɪʃ]	[l ih k axr ix sh]
[w]	[w]	kiwi	['kiwi]	[k iy w iy]
[r]	[r]	parsley	['pɑrsli]	[p aa r s l iy]
[j]	[y]	yew	[yu]	[y uw]
[h]	[h]	horseradish	['hɔrsrædɪʃ]	[h ao r s r ae d ih sh]
[ʔ]	[q]	uh-oh	[ʔʌʔoʊ]	[q ah q ow]
[ɾ]	[dx]	butter	['bʌɾɚ]	[b ah dx axr]
[ɾ̃]	[nx]	wintergreen	[wɪɾ̃ɚgrin]	[w ih nx axr g r i n]
[l̩]	[el]	thistle	['θɪsl̩]	[th ih s el]

IPA and **ARPAbet** symbols for transcription of English consonants



The major places of consonant articulation with respect to the human mouth.



Major English places of articulation.

Manner of articulation of English consonants.

Consonant Labels	Consonant Examples	Voiced?	Manner
b	**b**ig, a**b**le, ta**b**	+	plosive
p	**p**ut, o**p**en, ta**p**	-	plosive
d	**d**ig, i**d**ea, wa**d**	+	plosive
t	ta**l**k, sa**t**	-	plosive
g	**g**ut, a**ng**le, ta**g**	+	plosive
k	**c**ut, oa**k**en, ta**k**e	-	plosive
v	**v**at, o**v**er, ha**v**e	+	fricative
f	**f**ork, a**f**ter, i**f**	-	fricative
z	**z**ap, la**z**y, ha**z**e	+	fricative
s	**s**it, ca**s**t, to**ss**	-	fricative
dh	**th**en, fa**th**er, scy**th**e	+	fricative
th	**th**in, no**th**ing, tru**th**	-	fricative
zh	ge**n**re, a**z**ure, bei**g**e	+	fricative
sh	**sh**e, cu**sh**ion, wa**sh**	-	fricative
jh	**j**oy, a**g**ile, e**dg**e	+	affricate
ch	**ch**in, ar**ch**er, ma**rch**	-	affricate
l	**l**id, e**l**bow, sa**l**	+	lateral
r	**r**ed, pa**r**t, fa**r**	+	retroflex
y	**y**acht, o**n**ion, **y**ard	+	glide
w	**w**ith, a**w**ay	+	glide
hh	**h**elp, a**h**ead, **h**otel	+	fricative
m	**m**at, a**m**id, ai**m**	+	nasal
n	**n**o, e**n**d, pa**n**	+	nasal
ng	si**ng**, a**ng**er, dri**nk**	+	nasal

Feature Extraction

We can now summarise the process of extraction of spectral features, beginning with the sound wave itself and ending with a feature vector.

An input sound-wave is first digitised.

This process of analogue-to-digital conversion has two steps:

- Sampling and
- Quantisation.

A signal is sampled by measuring its amplitude at a particular time.

The sampling rate is the number of samples taken per second.

Common sampling rates are 8,000Hz and 16,000Hz.

In order to accurately measure a wave, it is necessary to have at least two samples in each cycle. One measuring the positive part of the wave and the other measuring the negative part.

More that two samples per cycle increases the amplitude accuracy, but less that two samples will cause the frequency of the wave to be complete missed.

Therefore, the maximum frequency wave that can be measured is one whose frequency is half that of the sample rate (since every cycle needs two samples).

The maximum frequency for a given sampling rate is called the **Nyquist** frequency.

Most information in human speech is in frequencies below 10,000 Hz, therefore a 20,000 Hz sampling rate would be necessary for complete accuracy.

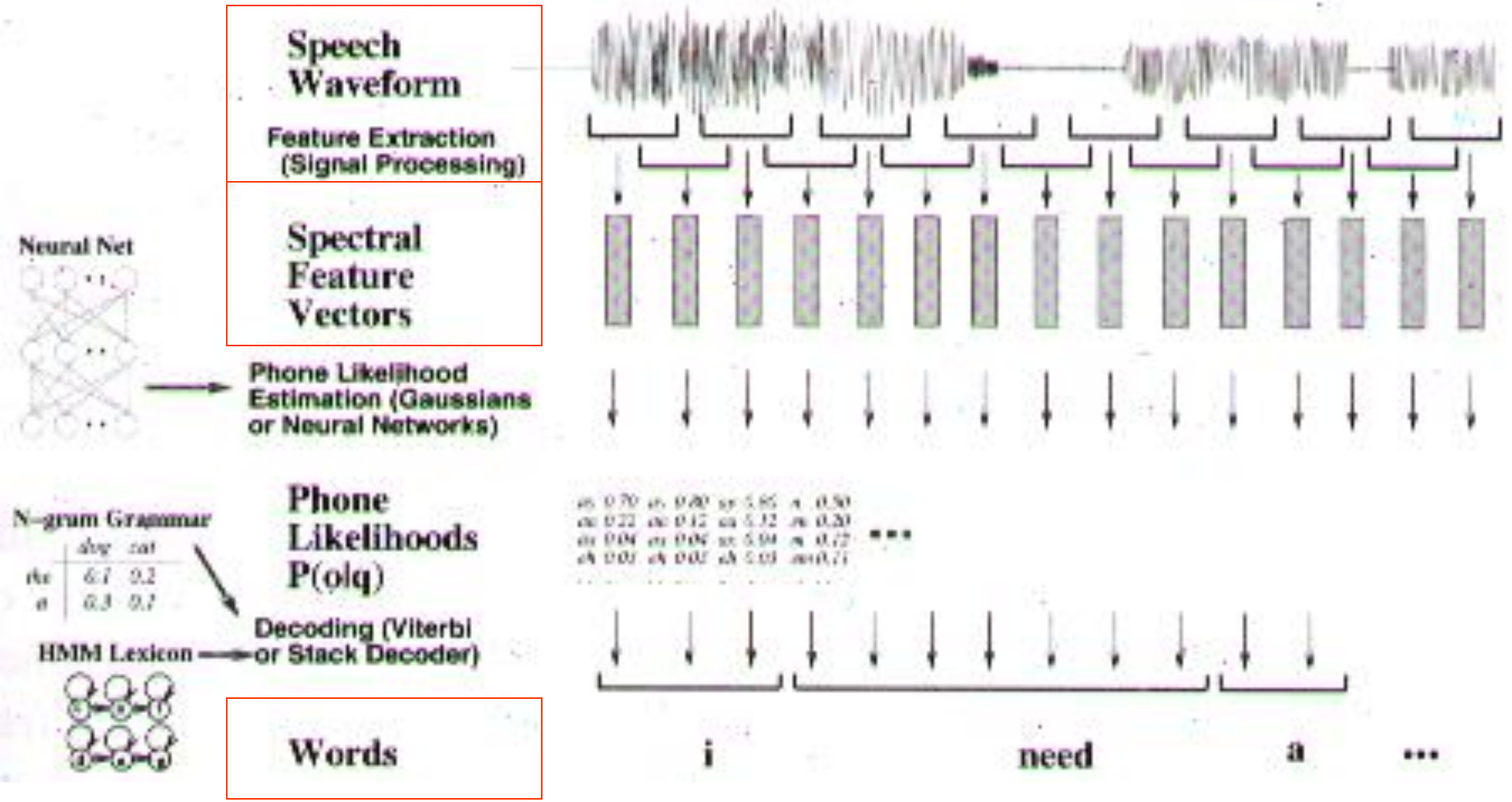
A sampling rate of, for example, 8,000 Hz, will require 8000 amplitude measurements for each second of speech.

It is very important to store amplitude measurements efficiently.

They are usually stored as integers, either 8-bit with values from –128 to 127 or 16-bit with values from –32768 to 32767.

- This process of representing a real-valued number as an integer is **quantisation** because there is a minimum granularity (the quantum size) and all values which are are closer together than this quantum size are represented identically.

- Once a waveform has been digitised, it is converted to some set of spectral features.


Speech Waveform
Feature Extraction (Signal Processing)
Spectral Feature Vectors
Neural Net
Phone Likelihood Estimation (Gaussians or Neural Networks)
Phone Likelihoods P(o|q)
N-gram Grammar
HMM Lexicon
Decoding (Viterbi or Stack Decoder)
Words
i
need
a
...

TO DO THIS WEEK

1. Read chapter 7 of the Jurafsky and Martin textbook