Speech Production and Modeling

Simon Leglaive

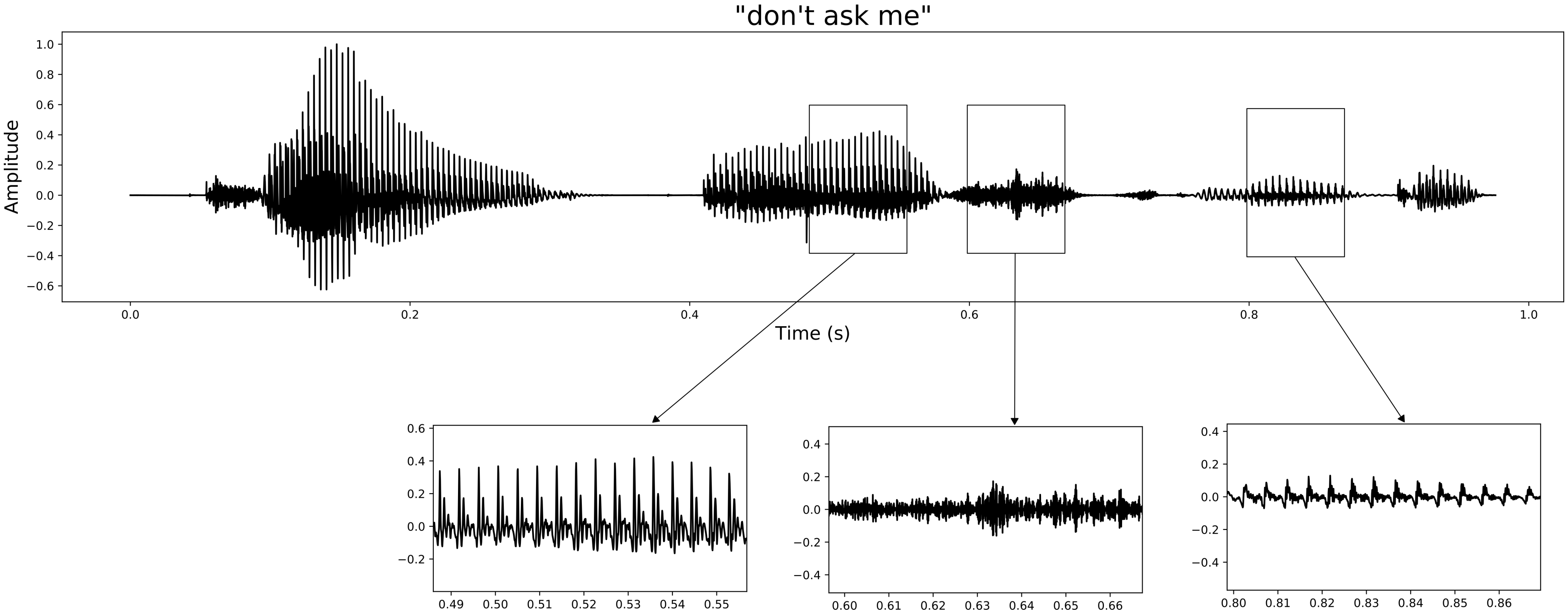
2D-3D Image & Sound, CentraleSupélec

Today

- Speech production
- Characteristics of speech signals
- Analysis, transformation and synthesis of speech signals with the source-filter model

Speech production

Speech signal

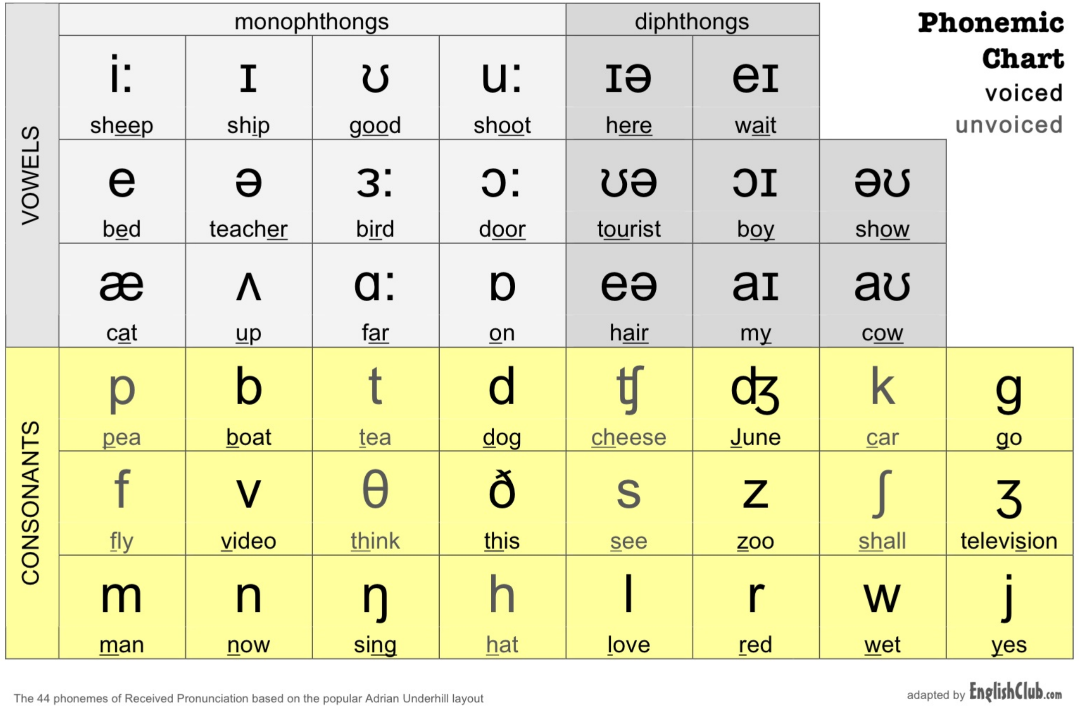


Can you guess to what “speech sound” each bloc corresponds?

Phonemes

Elementary speech sounds are called phonemes.

- 44 phonemes in English.
- 10–15 phonemes per second in normal English speech.
- We are going to see what are the key differences in the production of the different phonemes.



Phonemic Chart
voiced
unvoiced

	monophthongs				diphthongs			
VOWELS	iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		
	e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
	æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
CONSONANTS	p pea	b boat	t tea	d dog	tʃ cheese	dʒ June	k car	g go
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television
	m man	n now	ŋ sing	h hat	l love	r red	w wet	j yes

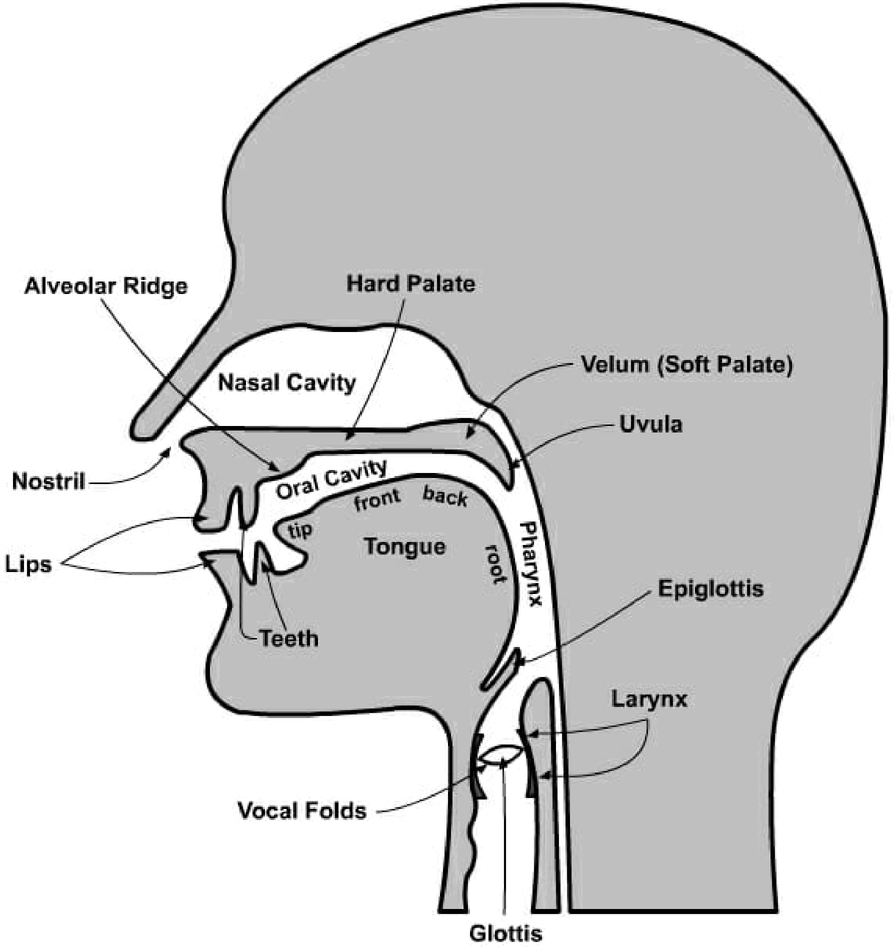
The 44 phonemes of Received Pronunciation based on the popular Adrian Underhill layout

adapted by EnglishClub.com

Image credit: https://www.englishclub.com/pronunciation/phonemic-chart.htm

Speech production – the global view

- The energy comes from air expelled from the lungs.
- At the larynx, this airflow passes between the vocal folds.
- Then it goes through the vocal tract, which is made of three cavities:
    1. the pharynx
    2. the oral cavity
    3. the nasal cavity
- Finally, sound goes out of the mouth and nose openings.



Articulators

We consider as articulator any mobile part of the vocal tract on which we can act voluntarily and which is used in the production of speech sounds.

Tongue

- Very mobile and flexible
- Very important for phonation

Jaw

- Little degrees of freedoms and rigid
- Less important for phonation

Lips

- Very mobile and flexible
- Important movements for phonation:
  - occlusion
  - protrusion
  - raising and lowering
  - stretching, raising and lowering of lip corners

Speech sound sources

We distinguish 3 types of sound sources, which can be combined or occur individually:

- Quasi-periodic source resulting from the vibration of the vocal folds.

  We say that the sound is voiced.

  It can be arbitrarily long (in the limits of an exhalation).

- Fricative noise source produced by a turbulent airflow with a constriction in the vocal tract.

  It can also be arbitrarily long.

- Plosive noise source produced by quick occlusions of the vocal tract and generating an acoustic impulse.

  Here the duration is short.

Voice production

Credits: Joe Wolfe, https://vimeo.com/234805962, more info at https://www.animations.physics.unsw.edu.au/waves-sound/human-sound/index.html

Vocal folds and pitch

- The vibration of the vocal folds defines the pitch of the speech signal (i.e. its fundamental frequency).
- Variations of pitch along time define the melody of the voice.

	Average pitch (Hz)	Pitch range (Hz)
Male	100 - 130	90 - 270
Female	150 - 300	120 - 360
Child	350 - 400	200 - 600

Pitch and mechanisms

Credits: Joe Wolfe, https://vimeo.com/128430263, more info at https://www.animations.physics.unsw.edu.au/waves-sound/human-sound/index.html

Vocal tract and formants

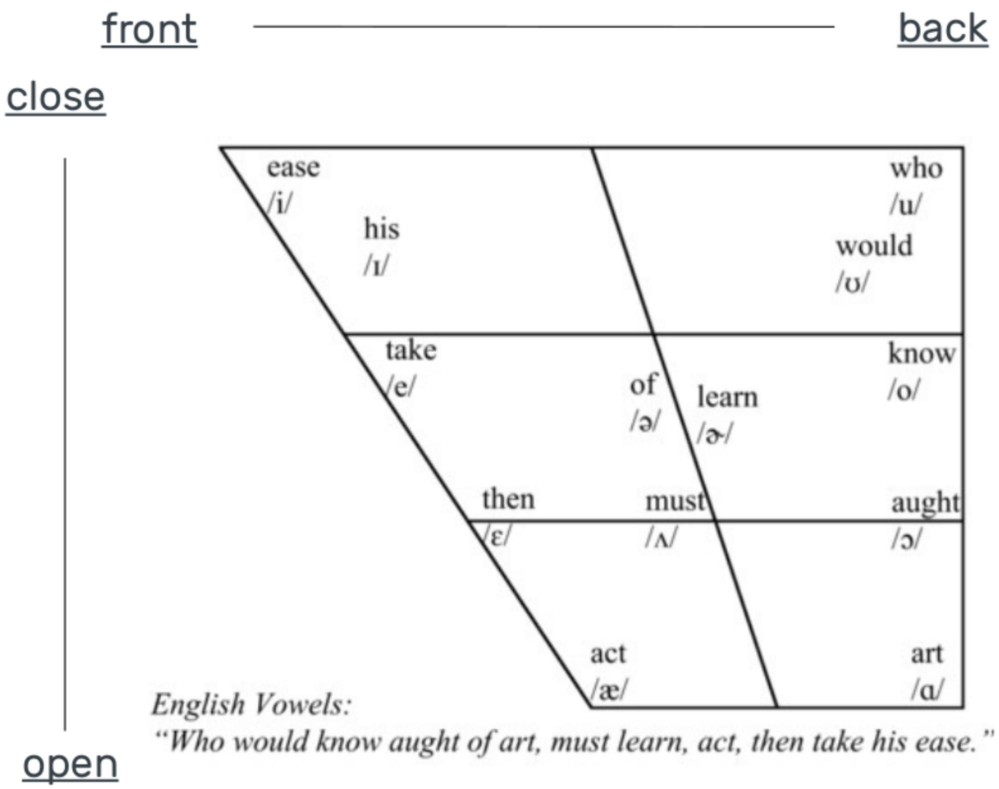
- The three elementary sound sources are modified by the vocal tract, before propagating out of the phonatory system, through the mouth and nose openings.
- The vocal tract actually corresponds to an acoustic filtering of the source signal.
- The cavities in the vocal tract give rise to resonances, that are called the formants.
- By modifying the shape of the vocal tract, we change the acoustic filter and the associated resonances.
- We can change the formants independently of the pitch, or in signal processing terms, we can change the filter independently of the source

Resonances and formants

Credits: Joe Wolfe, https://vimeo.com/128430264, more info at https://www.animations.physics.unsw.edu.au/waves-sound/human-sound/index.html

Distinctive articulatory features of vowels

- Opening of the mouth
  - Opened vowel [a] in “hat"
  - Closed vowel [i] in “meet”
- “Frontness” of the tongue
  - Front vowel [i] in “meet”
  - Back vowel [u] in “boot”
- Rounding of the lips
  - Rounded vowel [ɔ] in “not”
  - Not rounded vowel [i] in “meet”
- Nasalization: sound comes out of the mouth only, or out of the mouth and nose.
  - Nasal vowel [ɑ̃] in “pente” in French
  - Oral vowel [a] in “hat”



English Vowels:
“Who would know aught of art, must learn, act, then take his ease.”

Vowel chart with audio: https://en.wikipedia.org/wiki/IPA_vowel_chart_with_audio

Vowels and formants

We can distinguish between vowels using the position of the first formants

- high/low F1 ↔ opened/closed
- high/low F2 ↔ front/back
- high/low F3 ↔ not rounded/rounded lips

By moving articulators, the shape of the vocal tract varies, formants move in frequency, and vowels change.

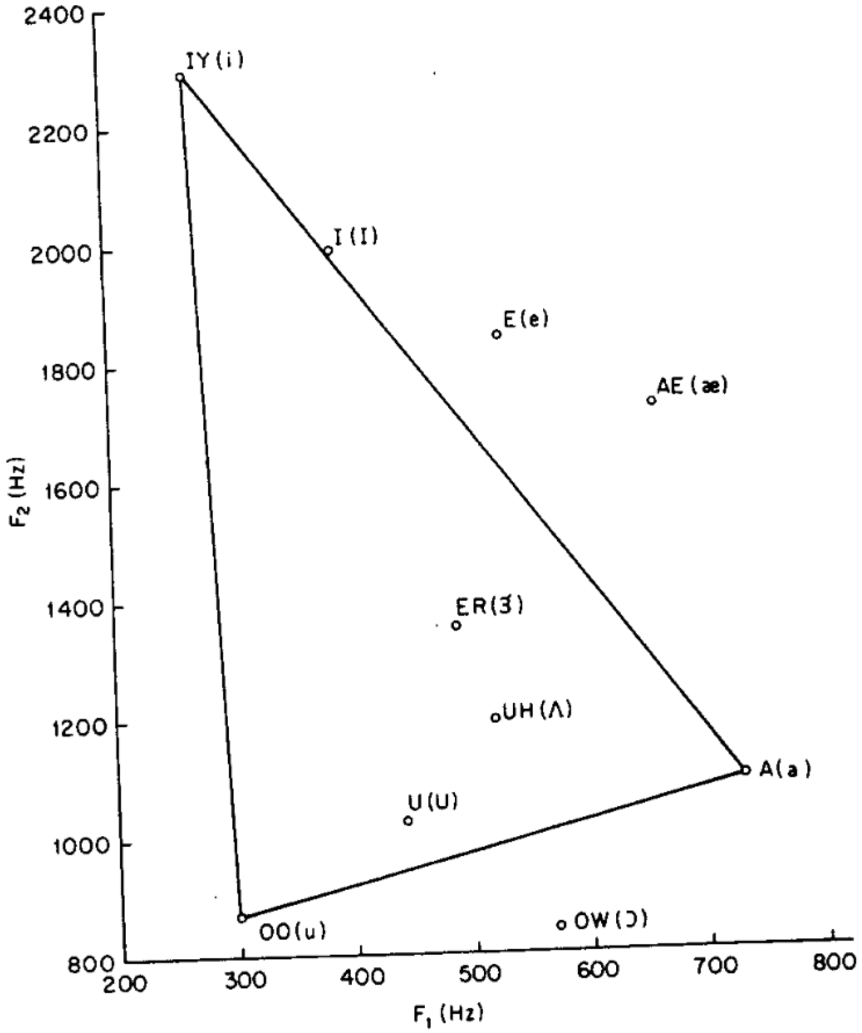


Fig. 3.5 The vowel triangle.

Image credit: L.R. Rabiner and R.W. Shafer, "Digital Processing of Speech Signals", Prentice-Hall, 1978

Vowels clustering in the formants space

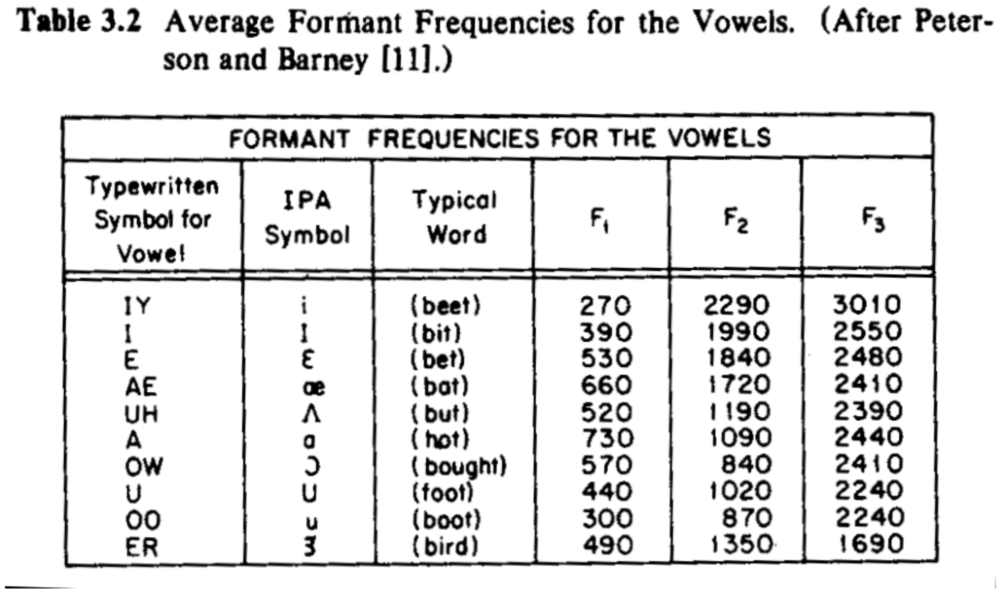


Table 3.2 Average Formant Frequencies for the Vowels. (After Peterson and Barney [11].)

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	ɜ	(bird)	490	1350	1690

male speakers

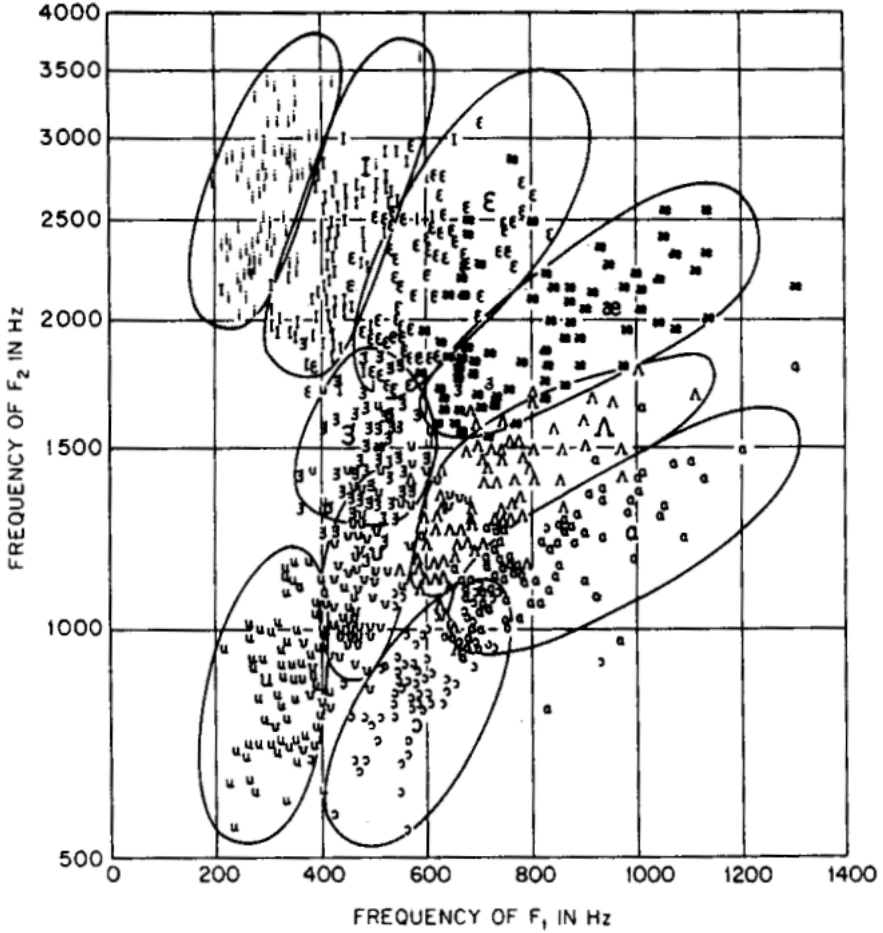


Fig. 3.4 Plot of second formant frequency versus first formant frequency for vowels by a wide range of speakers. (After Peterson and Barney [11].)

male and children speakers

Image credit: L.R. Rabiner and R.W. Shafer, “Digital Processing of Speech Signals”, Prentice-Hall, 1978

Consonants

Fricatives

- fricative noise source
- voiced [v, z, j] or unvoiced [?, ?, ?]
- locally stationary

Plosives

- plosive noise source
- voiced [?, ?, ?] or unvoiced [p, t, k]
- highly non-stationary

Nasal

- voiced
- sound comes mostly from the nose
- examples: [m, n]

Liquids

- voiced
- the vocal tract changes rapidly, especially using the tongue
- examples: [l, r]

Consonant chart with audio: https://en.wikipedia.org/wiki/IPA_pulmonic_consonant_chart_with_audio

Consonants

Fricatives

- fricative noise source
- voiced [v, z, j] or unvoiced [f, s, ch]
- locally stationary

Plosives

- plosive noise source
- voiced [b, d, g] or unvoiced [p, t, k]
- highly non-stationary

Nasal

- voiced
- sound comes mostly from the nose
- examples: [m, n]

Liquids

- voiced
- the vocal tract changes rapidly, especially using the tongue
- examples: [l, r]

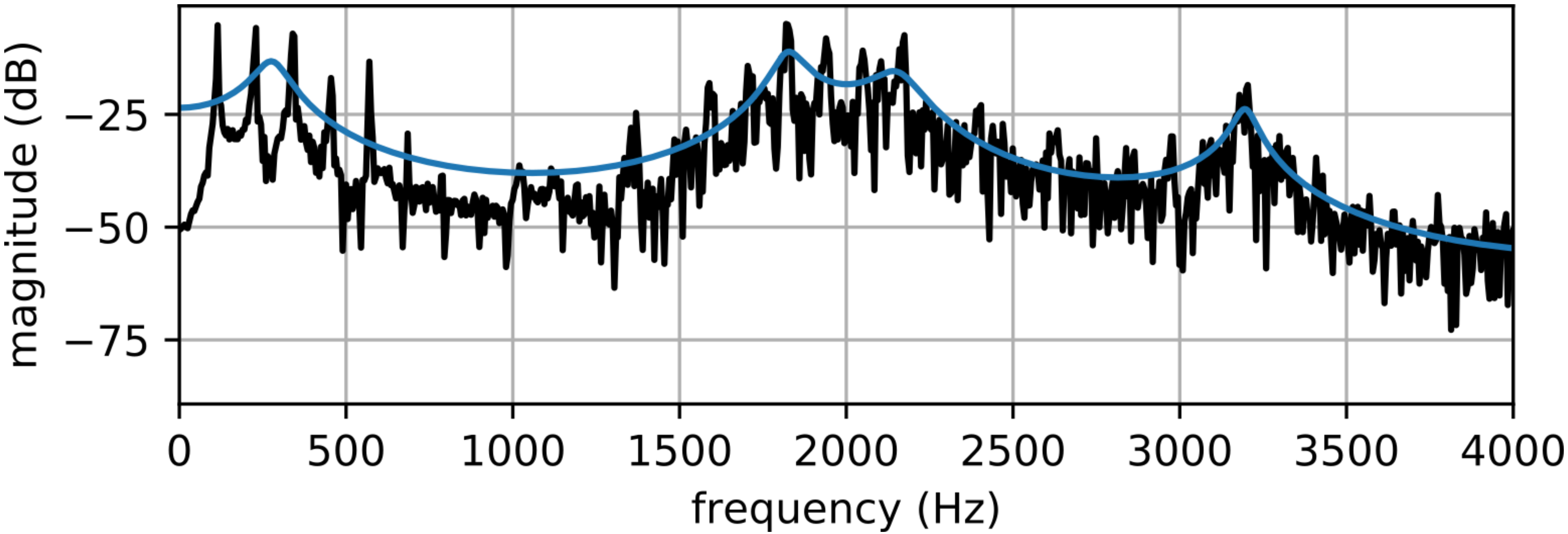
Consonant chart with audio: https://en.wikipedia.org/wiki/IPA_pulmonic_consonant_chart_with_audio

Prosody

- Prosody is on top of the flow of phonemes.
- Prosodic variables:
  - pitch (fundamental frequency)
  - speech rate (number of speech units, e.g. phonemes, per second)
  - loudness (intensity)
  - timbre (spectral characteristics such as amplitude of harmonics)
- Different combinations of these variables are exploited for intonation and accentuation.
- Prosody may reflect various features of the speaker or the utterance:
  - the identity of the speaker
  - the emotional state of the speaker
  - the form of the utterance (statement, question, or command)
  - the presence of irony or sarcasm
  - emphasis

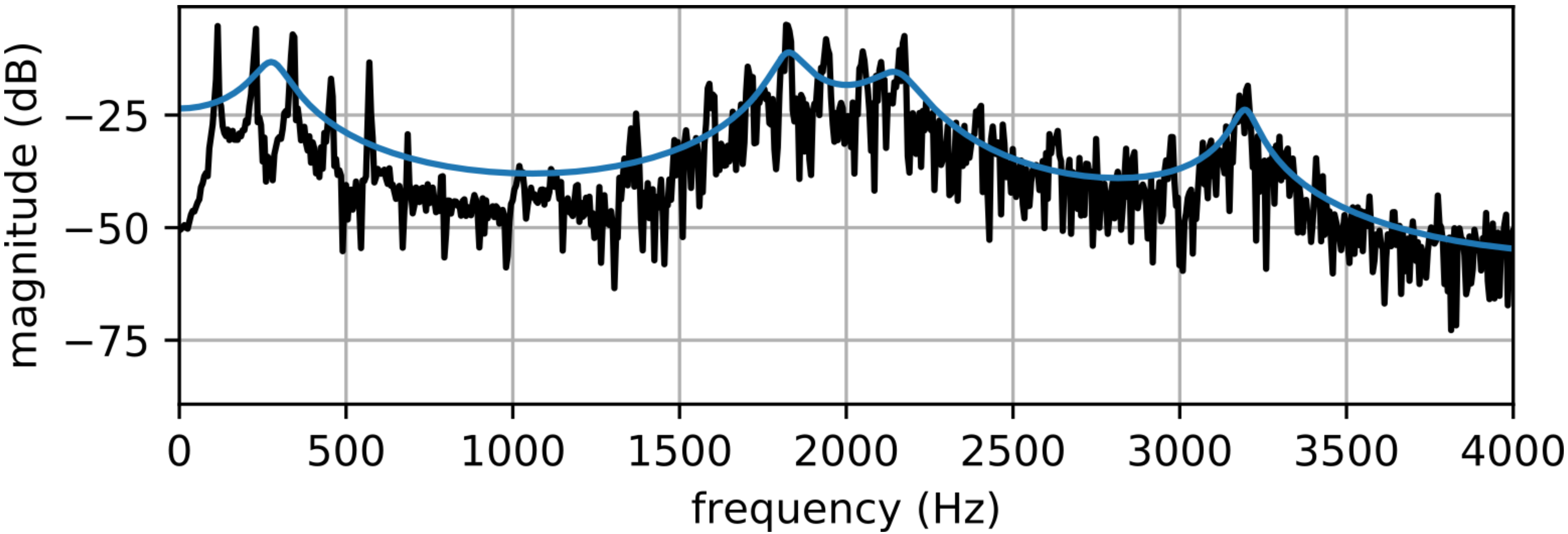
Spectrum/spectrogram reading

The spectral envelope



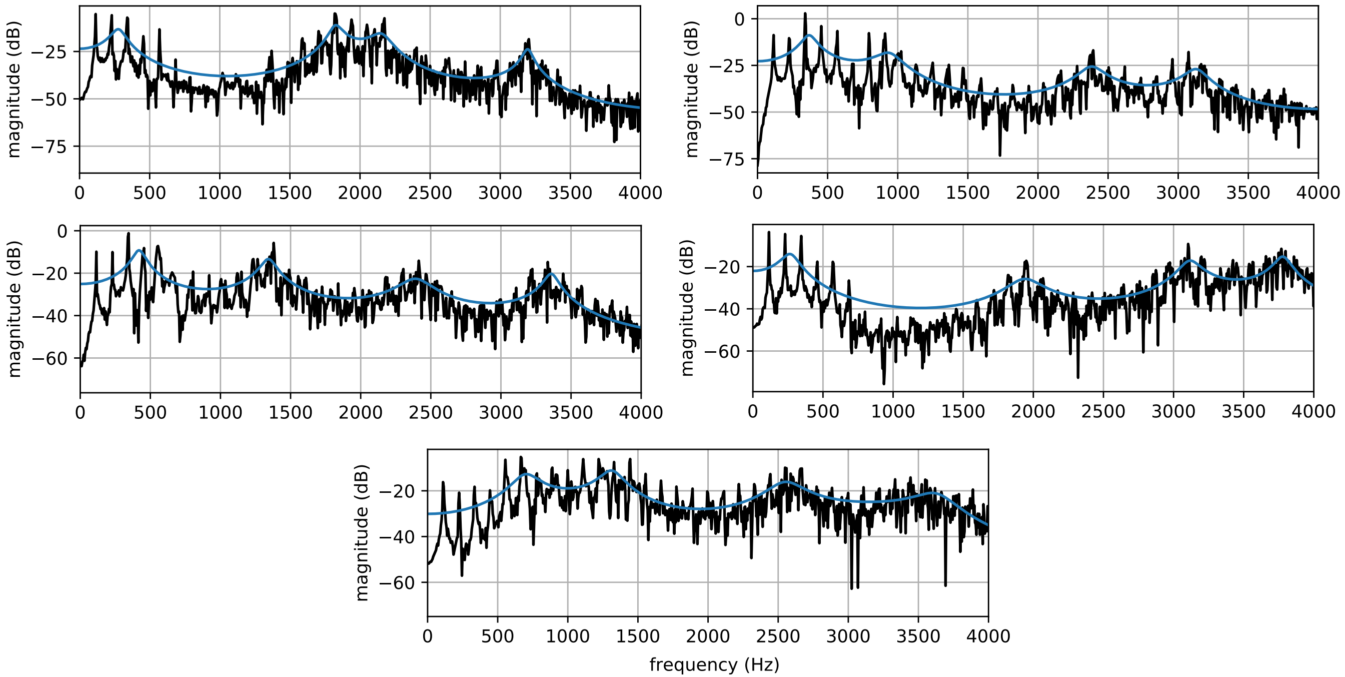
- Black curve: power spectrum (in dB) of the recording of a vowel, computed with the DFT.
- Blue curve: spectral envelope showing the formant resonances, computed with linear predictive coding (will be discussed in the lab session).

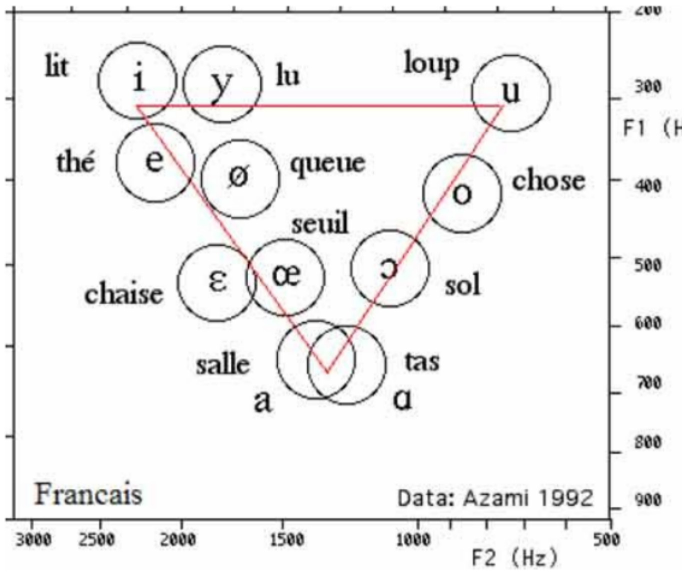
The spectral envelope



- Black curve: power spectrum (in dB) of the recording of a vowel, computed with the DFT.
- Blue curve: spectral envelope showing the formant resonances, computed with linear predictive coding (will be discussed in the lab session).

Male or female speaker?

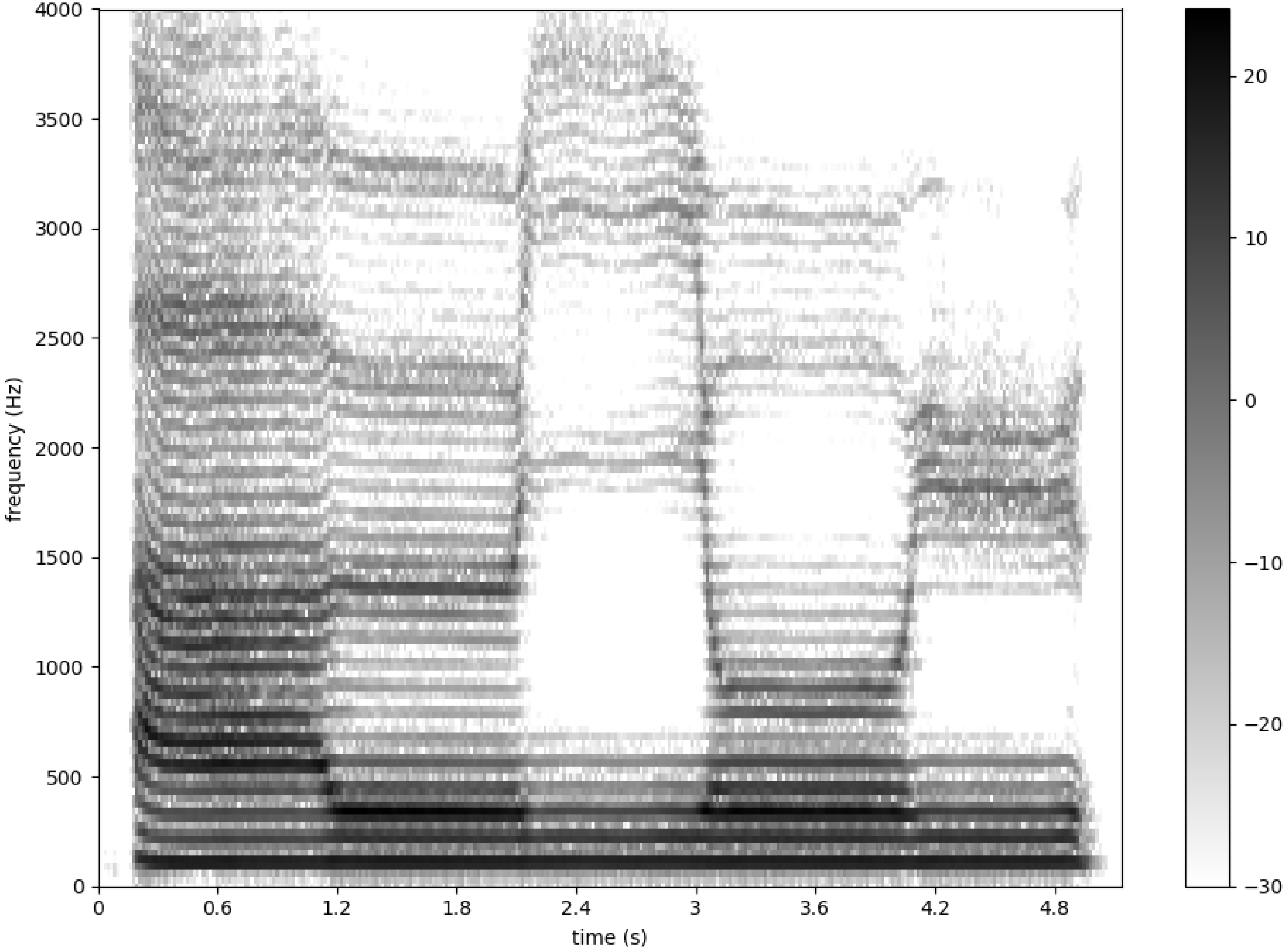


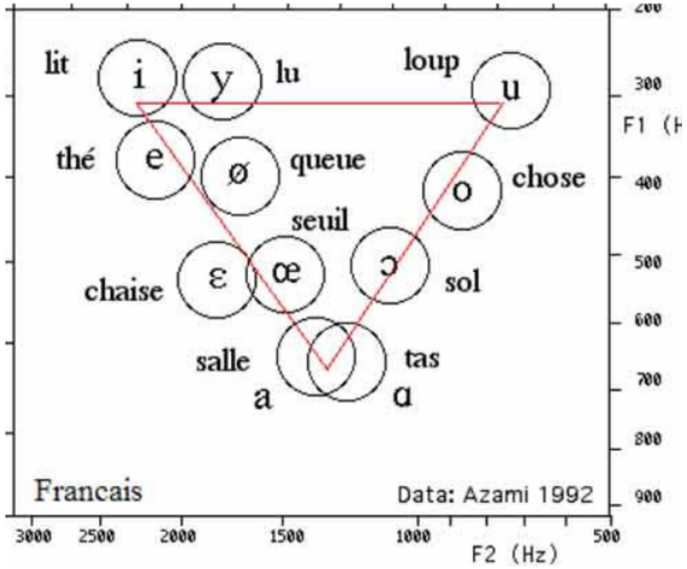


Go to www.wooclap.com/CXIOJL and find the vowel that corresponds to each spectrum, using the above French vocal triangle.

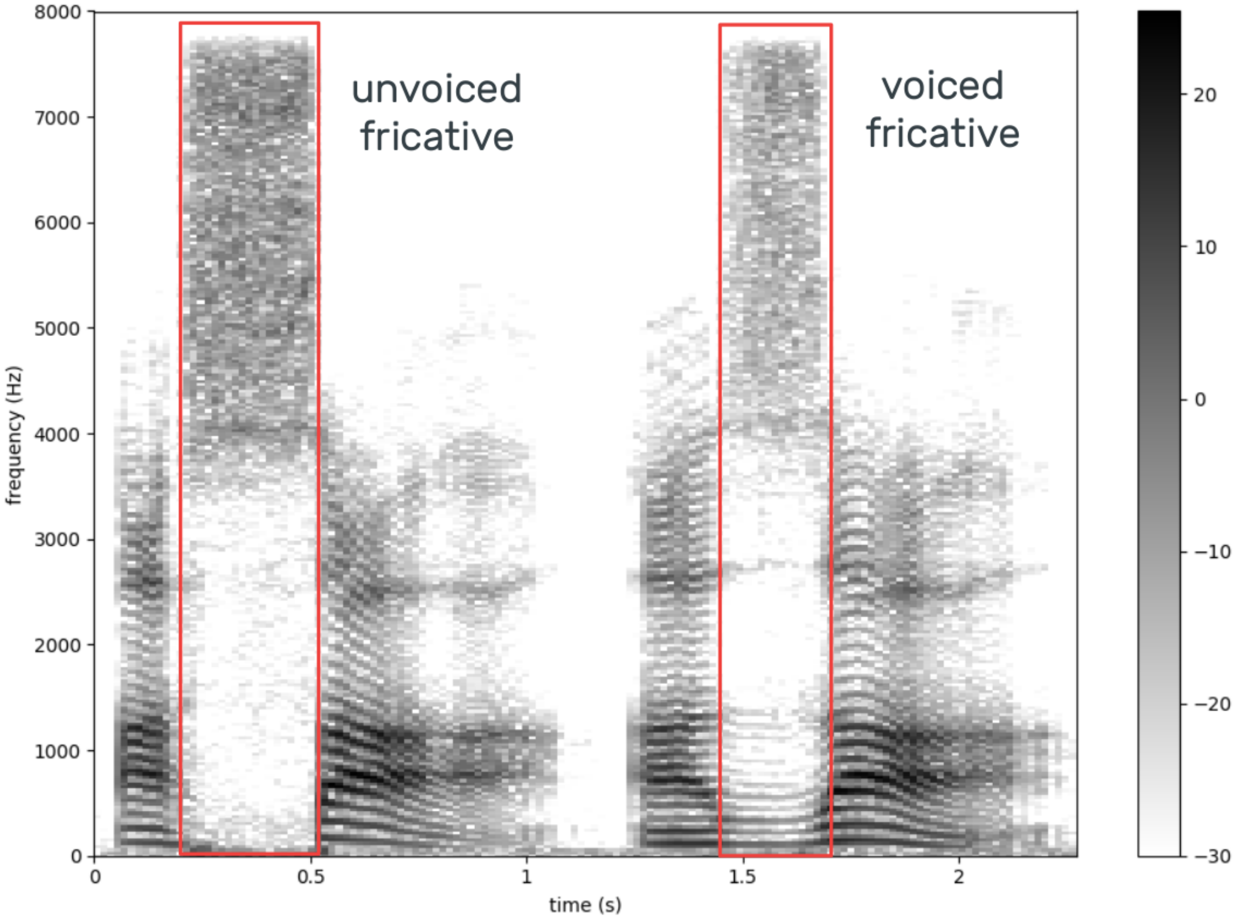
Spectrogram reading - "aeiou"

We could have done the same from a spectrogram representation.

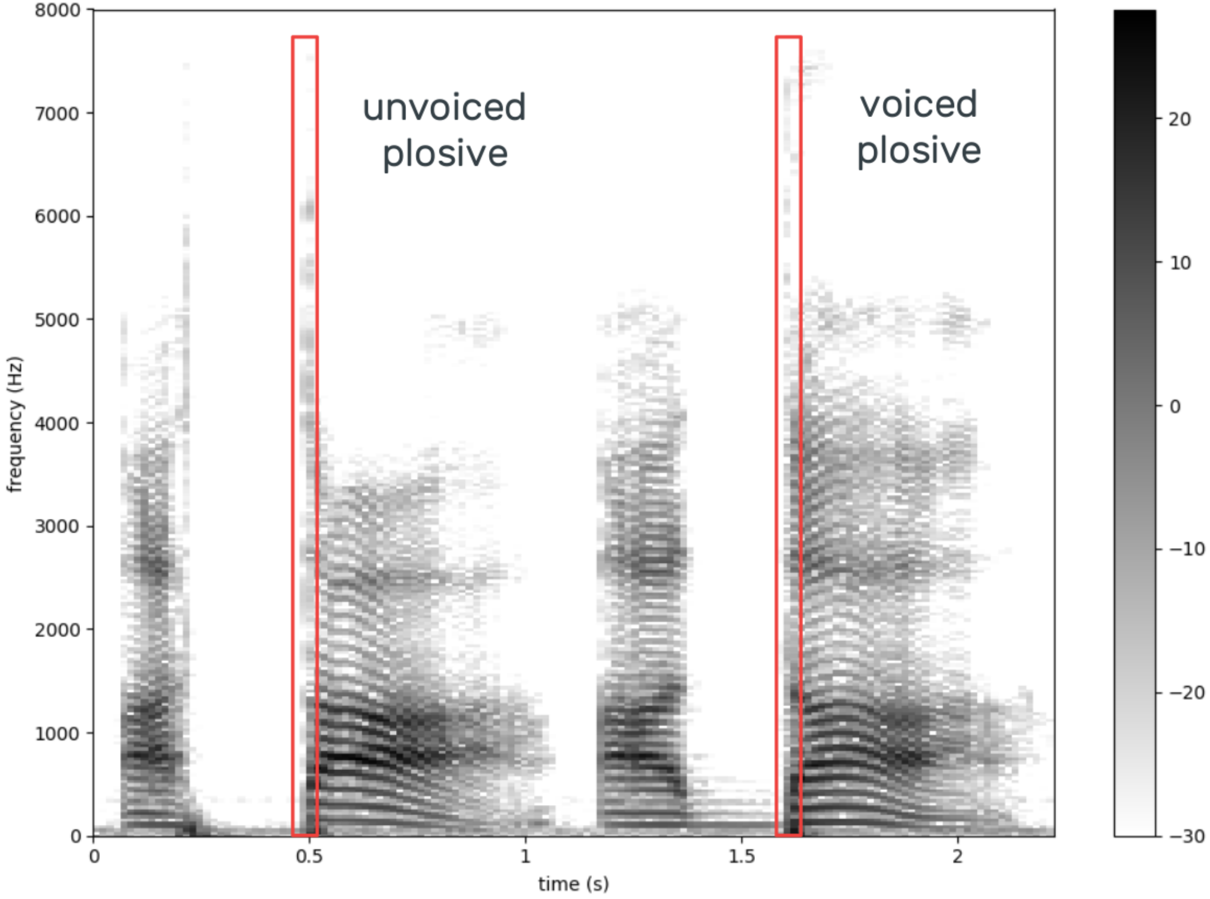


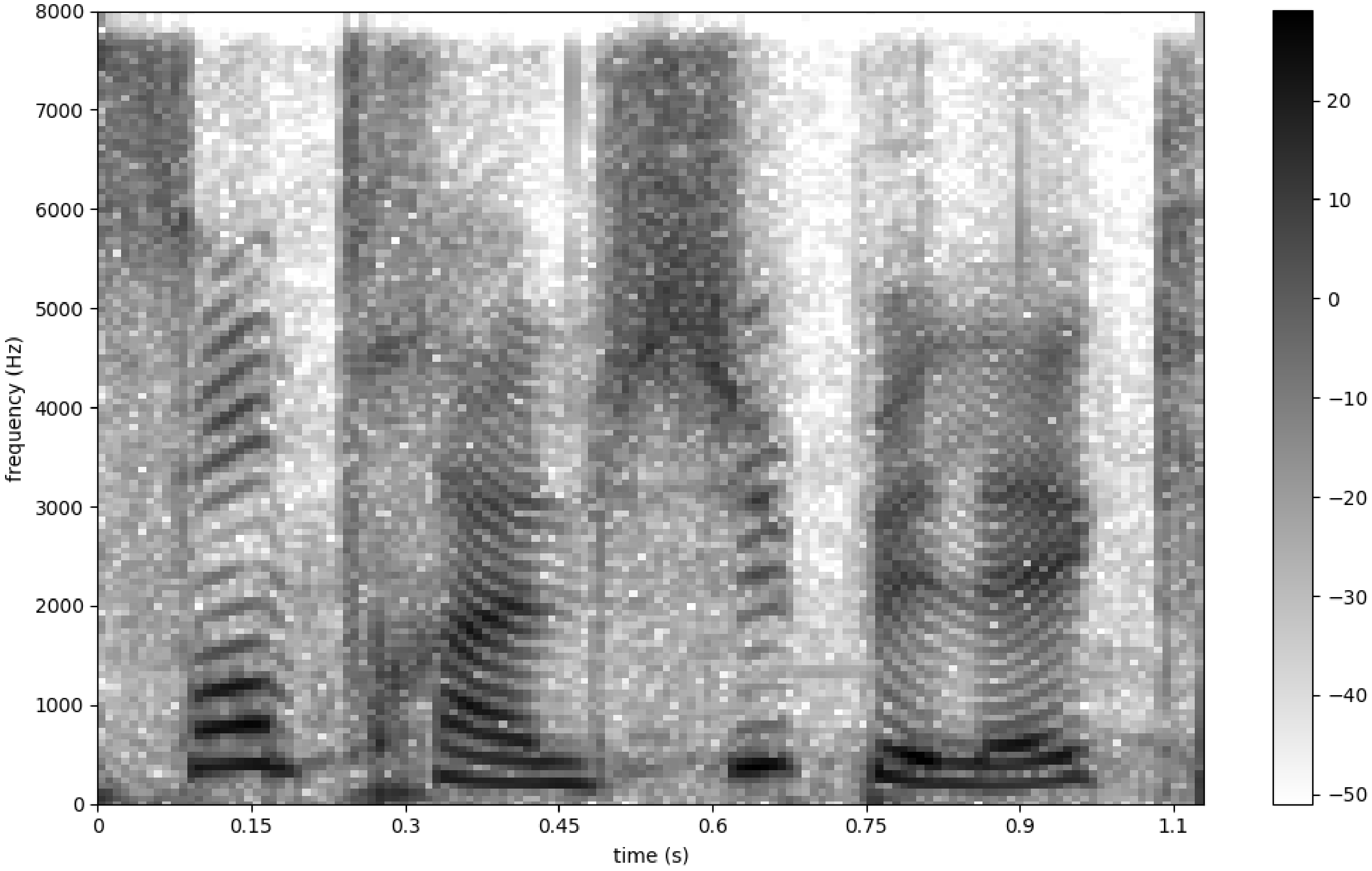


Spectrogram reading - "assa - azza"



Spectrogram reading - "atta - adda"





With a bit of practice you could be able to decode this mystery spectrogram.

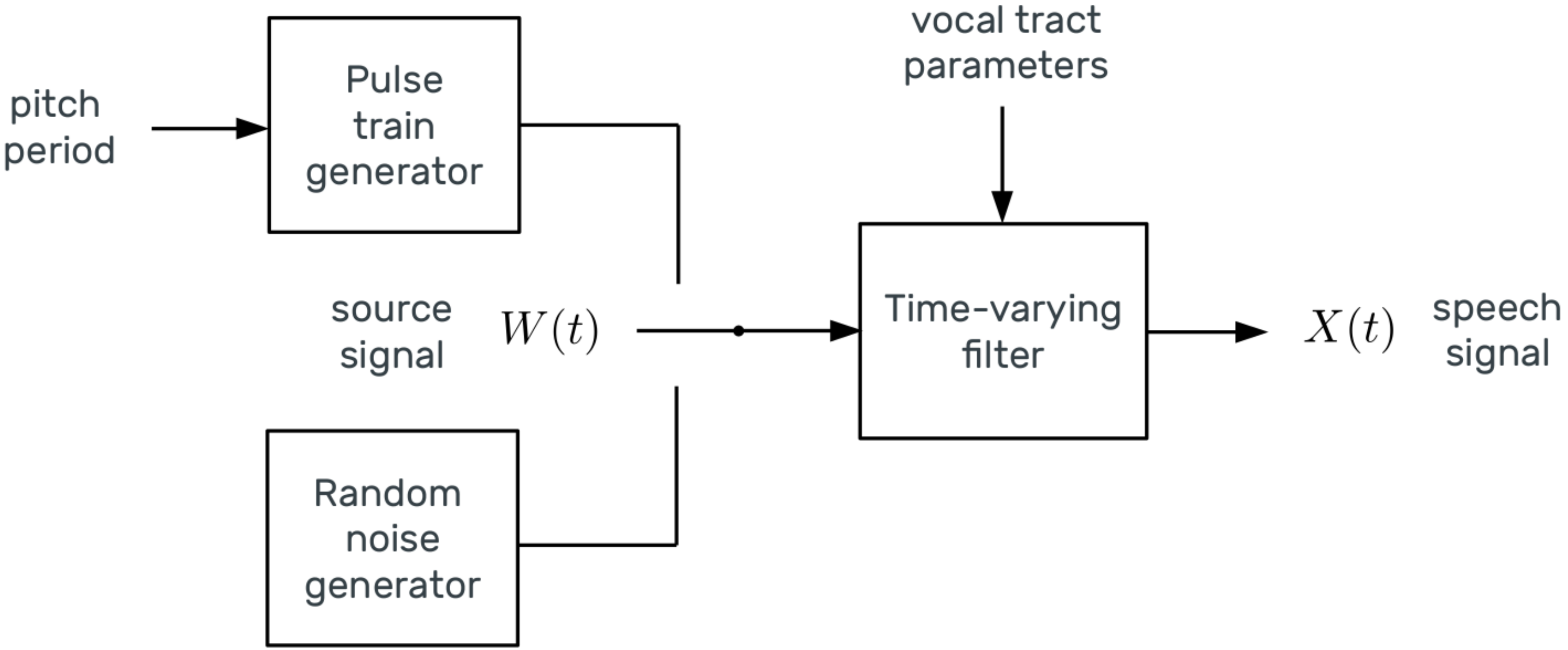
1 virtual bonus point if you do it 😉.

Further reading

Introduction to voice acoustics by Joe Wolfe, Emeritus Professor at the University of New South Wales (Syndney, Australia):

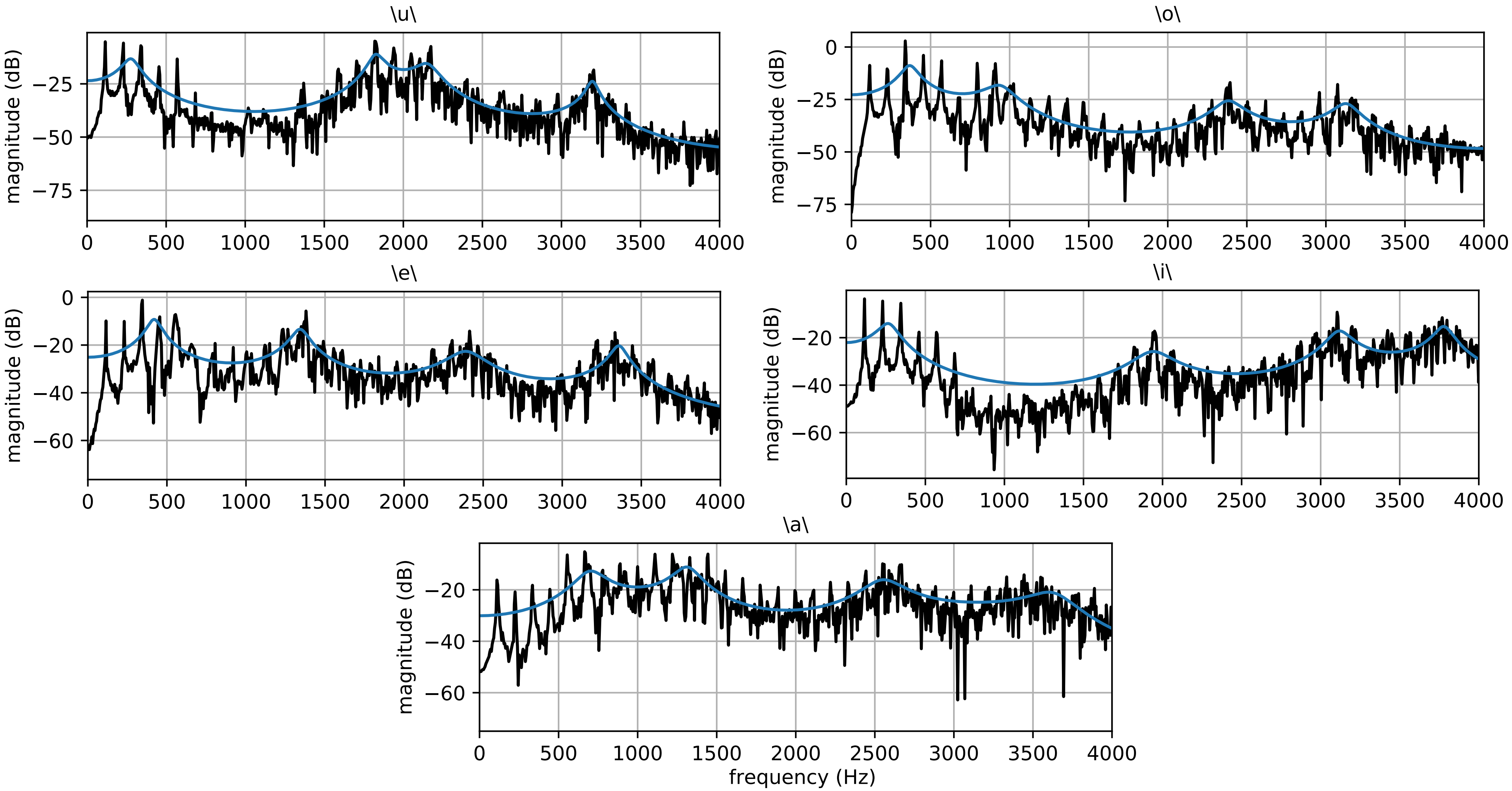
https://newt.phys.unsw.edu.au/jw/voice.html

Practical activity



Analysis, transformation and synthesis of speech signals with the **source-filter model**

Solution to the wooclap


\u\
\o\
\e\
\i\
\a\
magnitude (dB)
−25
−50
−75
0
−20
−40
−60
0
500
1000
1500
2000
2500
3000
3500
4000
frequency (Hz)