

# Speech Recognition

Frapiccini Cecilia

Gilardi Luca

Micheli Giacomo

Saviello Raffaele

# Index

- Introduction and data presentation
- First Analysis and PCA
- Conclusion (Future steps)



## Introduction

**Our dataset is contained in the “emuR” package and it is called “dip”**

**It contains several samples along with their associated information**

- actual data (dip\$data)
- speaker label (dip.spkr)
- diphthong label (dip.l)



## Our dataset

Two speakers, a male and a female

Three diphthongs (*aI*, *aU*, *OY*) for a total of 186 samples, 93 for each speaker, divided as:

- 59 times for *aI*
- 22 times for *aU*
- 12 times for *OY*

## Samples and Observations



Each sample is generated by a variable number of observations taken every 5 ms during the time window needed to record the entire diphthong.



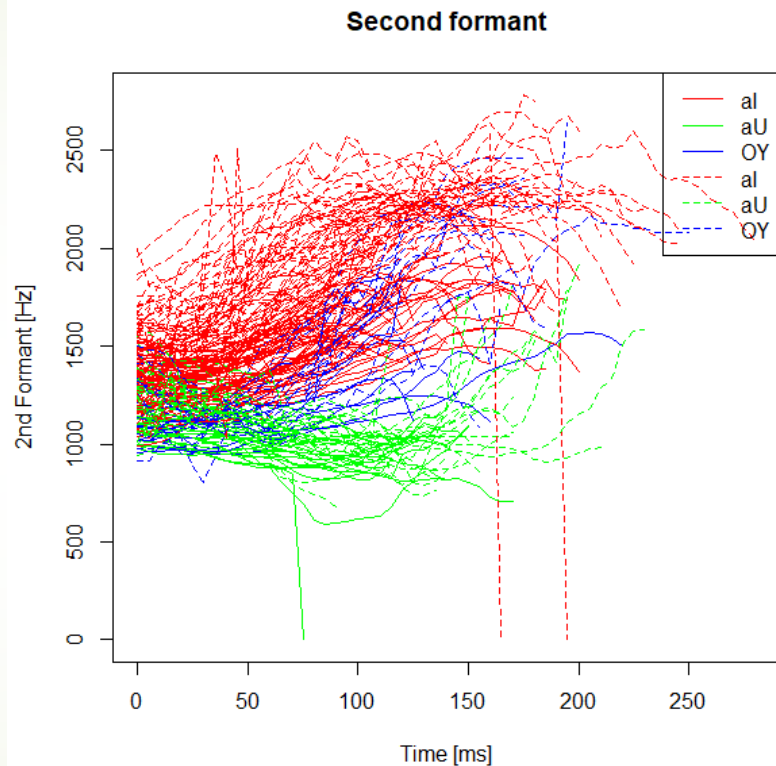
Time windows last on average 145 ms



Each observation is composed of 4 numerical values referring to the four formants

# What is a Formant?

► In speech science and phonetics, a formant is the spectral shaping that results from an acoustic resonance of the human vocal tract



# Goals

- Analyzing separately male and female in order to point out differences and common issues
- Join the results above and highlight pattern to recognize a diphthong when pronounced (regardless of speaker's gender)

# Are two speakers enough for our goals?

## ► YES

- The differences between different genders are more relevant than differences within the same gender
- If we discover some similarities between our two speakers we may generalize them to an entire population
- However, analyzing only two speakers allows us to discover relevant patterns at least inside the same gender






## Different datasets

Creation of six different datasets with normal time in order to analyze the main patterns on each diphthong for both genders:

- Female\_al
- Female\_aU
- Female\_OY
- Male\_al
- Male\_aU
- Male\_OY



## Handling with zero values

### Issue:

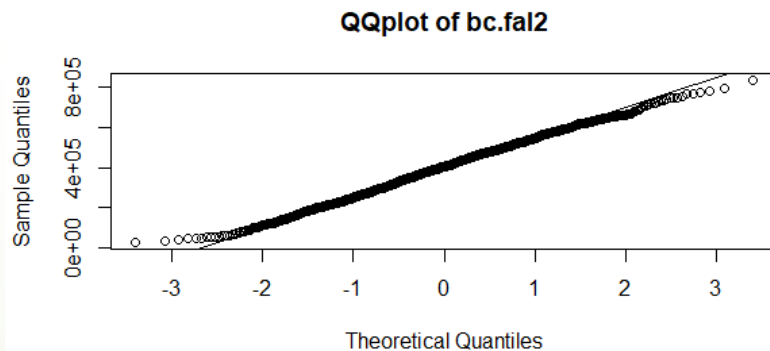
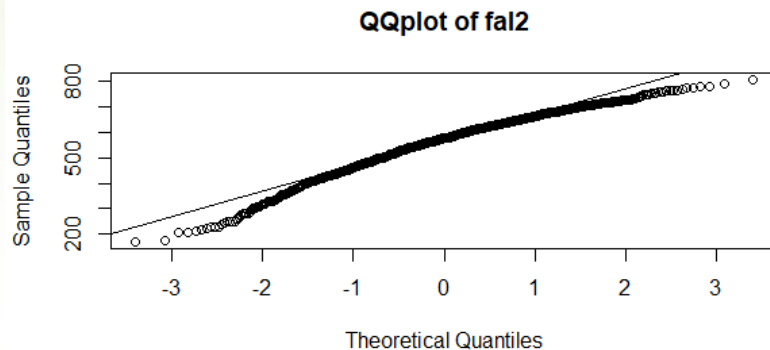
- Dealing with null values produced by technical errors in each sub-dataset


### Possible solutions:

- ~~Erase them~~
- Substitute them with the mean of formant's value for that diphthong and the associated speaker

# Gaussianity

- ▶ QQplot and Shapiro test on each dataset
- ▶ Unsatisfactory results
- ▶ With a Box Cox transformation results didn't change too much (got even worse)
- ▶ We postpone the problem until we will have knowledge of more efficient methods to solve it



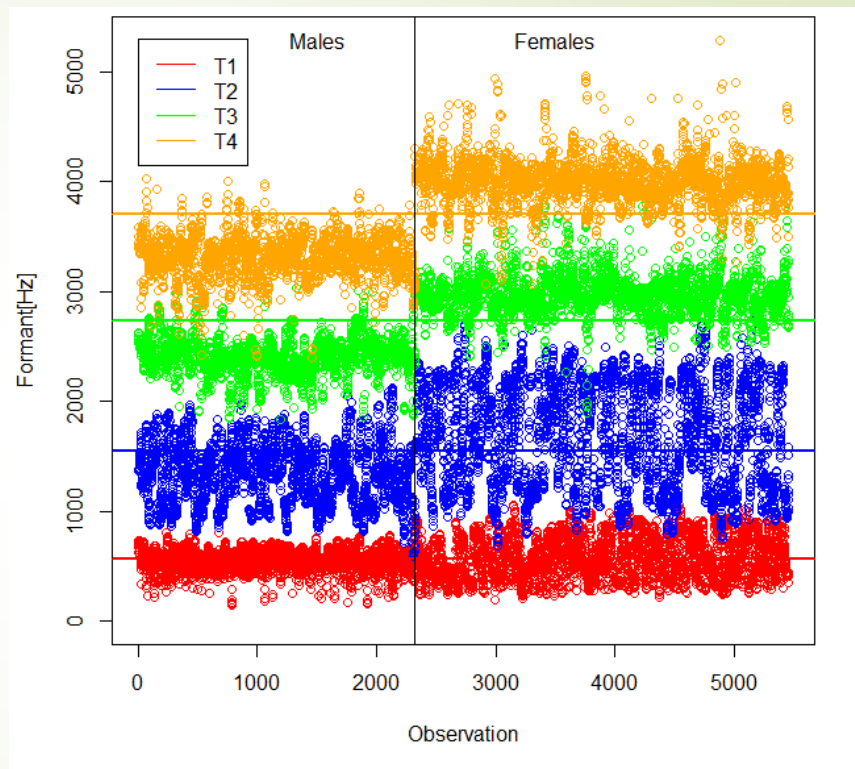


## What we have done so far

- Analysis of the behaviour of the formants for:
  - Samples for each speaker
  - Speaker for each diphthong
  - For each pair of speaker and diphthong

# Analysis of the samples for each speaker

- Third and fourth formants:
  - Clear distinction between the frequencies of female and male
- Second formant:
  - Highest variability



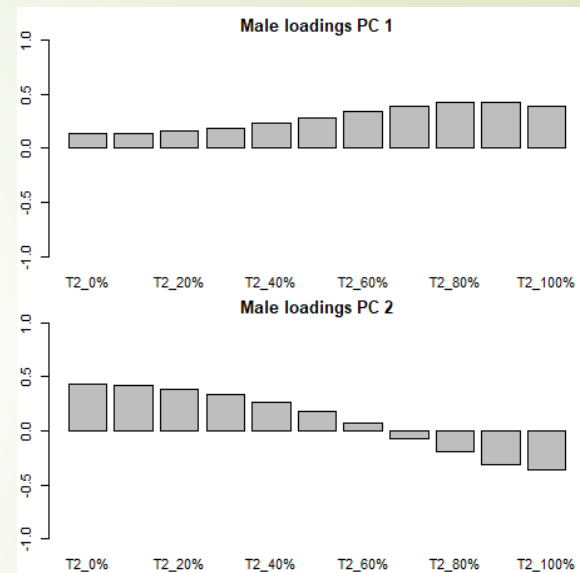
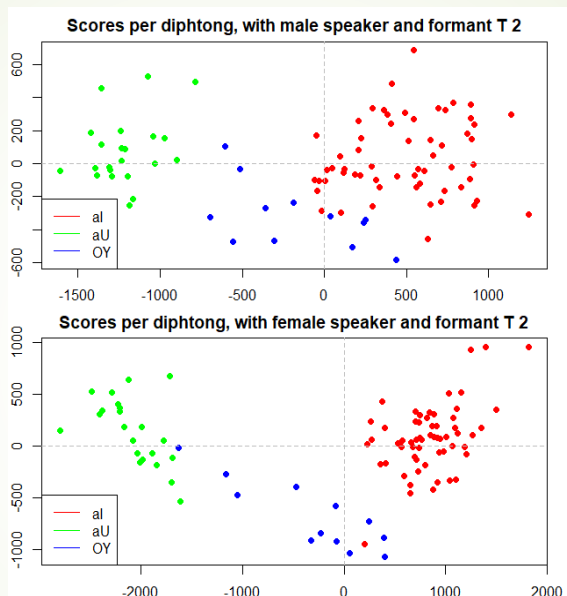
# Data handling

- Normalization of time values
- Computation of the approximated values of the formants
- Usage of the approximated values for the computation of relevant features

# PCA on speakers

➤ Most interesting results obtained with the second formant

➤ Clear distinction among the diphthongs





## Future Steps

- We will analyze the classification problem of the diphthongs (with and without knowing the speaker)
- Given the data time dependence, we will use the functional analysis to reach final results
- If we will not find out any solution for the normality problem, we will take into account to use non parametric tests for inference





**Thanks for your attention!**  
**Any questions?**