UZH-UNIGE Workshop on Computers in L2 Learning and Assessment.
University of Zurich (Switzerland), 3 May 2019

# Towards the automatic evaluation of L2 pronunciation using Pillai scores and LDA classification accuracy

Paolo Mairano[1] & Fabian Santiago[2]

[1]University of Lille, UMR 8163 Savoirs, Textes, Langage (France)

[2]University of Paris 8, Structures Formelles du Langage (France)

Université de Lille

STL

UNIVERSITÉ PARIS 8 VINCENNES-SAINT-DENIS

Structures Formelles du Langage

# Introduction

- L2 pronunciation assessment → underrepresented in models of communicative competence/language ability
- One of these reasons is:
  - conciliating concepts of the L2 teaching field (intelligibility, comprehensibility…) with those of research in L2 phonetics → difficult task

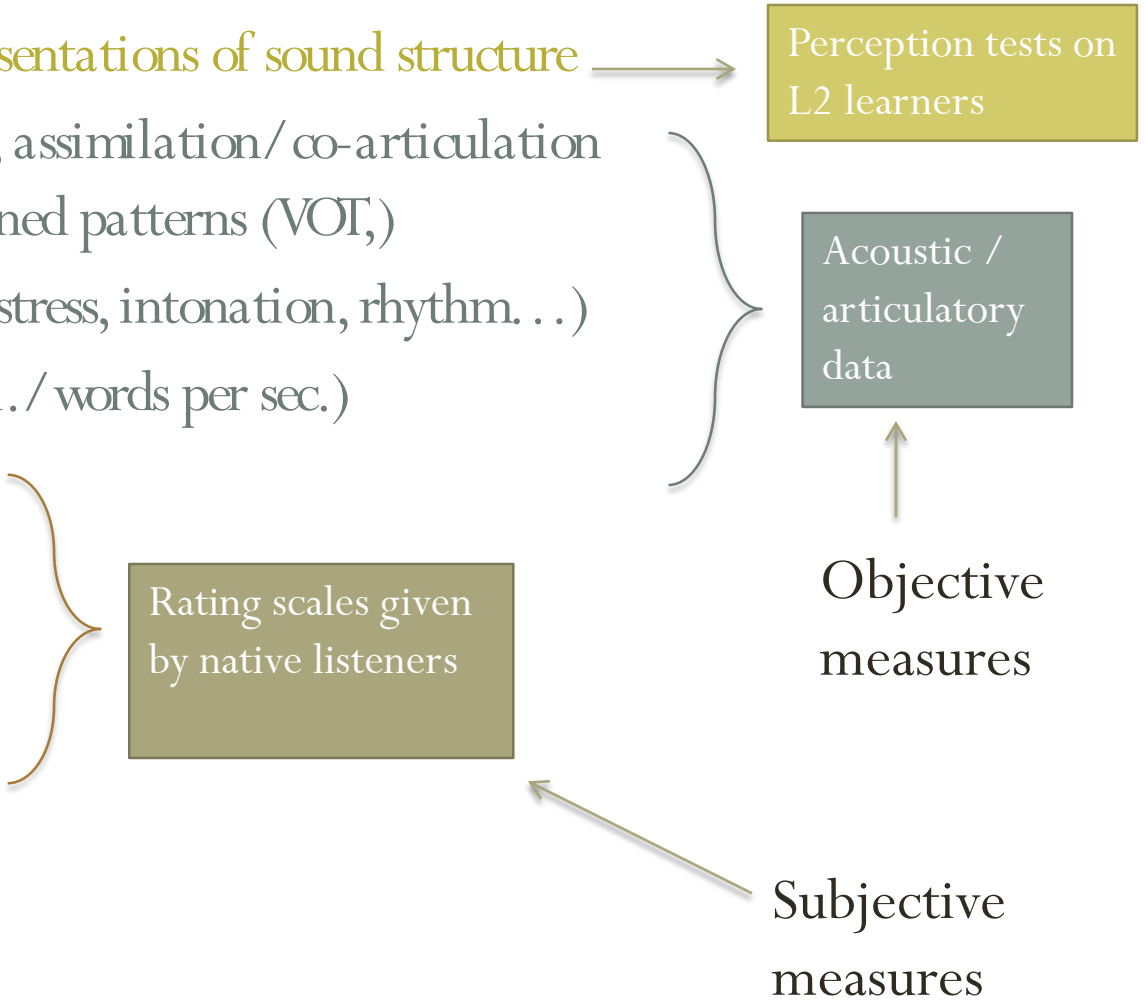
(Isaacs 2014, Derwin & Munro 2009)

# What can be assessed in L2 pronunciation?

- Phonological representations of sound structure
- Vowels, consonants, assimilation/co-articulation processes, fine-grained patterns (VOT)
- Prosodic patterns (stress, intonation, rhythm…)
- Speech rate (# syll./words per sec.)
- Intelligibility
- Comprehensibility
- Accentedness
- Fluency

# How can it be assessed ?

- Phonological representations of sound structure
- Vowels, consonants, assimilation/co-articulation processes, fine-grained patterns (VOT,)
- Prosodic patterns (stress, intonation, rhythm…)
- Speech rate (# syll./words per sec.)
- Intelligibility
- Comprehensibility
- Accentedness
- Fluency



# Assessing for which purpose?

- Phonological representations of sound structure
  Vowels, consonants, assimilation/co-articulation processes, fine-grained patterns (VOT)
- Prosodic patterns (stress, intonation, rhythm…)
- Speech rate (# syll./words per second)
- Intelligibility
- Comprehensibility
- Accentedness
- Fluency

Understanding the L2 phonology acquisition process

Measuring native-likeness

Measuring comprehensible speech in context

# Human assessment of L2 pronunciation (Derwing & Munro, 2005)

**TABLE 1**

**Intelligibility, Comprehensibility, and Accentedness**

| Term | Definition | Measure |
|---|---|---|
| Intelligibility | The extent to which a listener actually understands an utterance | Transcription task<br>% words correct |
| Comprehensibility | A listener's perception of how difficult it is to understand an utterance | Scalar judgment task<br>1 = extremely easy to understand<br>9 = extremely difficult to understand |
| Accentedness | A listener's perception of how different a speaker's accent is from that of the L1 community | Scalar judgment task<br>1 = no accent<br>9 = extremely strong accent |

# Example

Intelligibility task

1. Transcribe the sentence : _____
2. Fill the blanks : La semana pasada un turista _____ se _____ y me dijo.

Comprehensibility scores

1 2 3 4 5 6 7 8 9

extremly easy to understand — impossible to understand

Accentendess

1 2 3 4 5 6 7 8 9

no foreign accent — very strong accent

# Assessing L2 pronunciation with machines

- Automatic Speech Recognition (ASR)
  1. Speech recognition: L2 speech signal → sequence of words
  2. Scoring: comparison of speech rate/acoustic properties between the L2 utterance and model
  3. Error detection: signalling a certain sound within a problematic word to the learner
  4. Error diagnosis: identification of the specific error

(Neri et al. 2003)

- Systems using purely acoustic features (cepstral, mfcc) and/or phonetic features (e.g. VOT, formants)

L2 speech is assessed with reference to a native model (i.e. extrinsically)

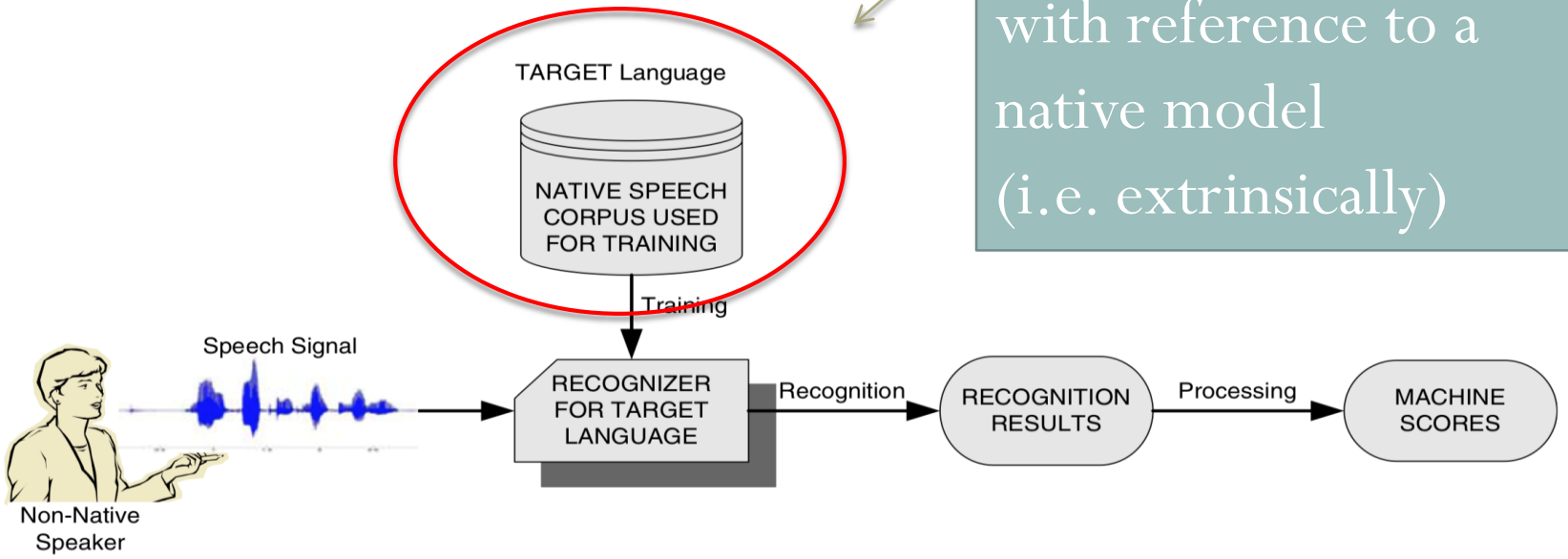



Fig. 1. The VILTS system. A non-native speaker is evaluated on his pronunciation of the TARGET language according to how close his speech characteristics are to those of the native speakers. The text must be known to the system.

Moustroufas et al. (2007)

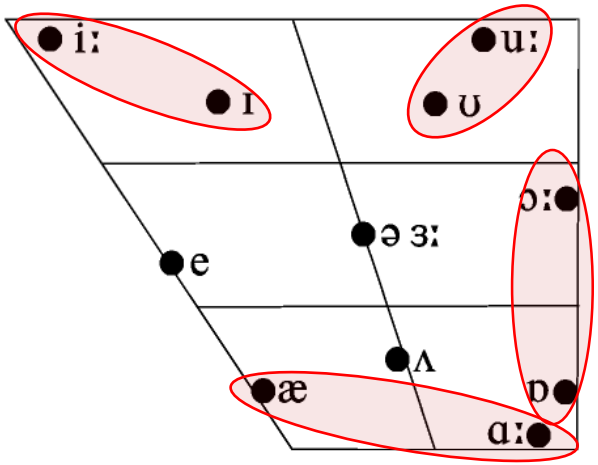
# Outline of this presentation

- Goal: proposing metrics for the intrinsic assessment of L2 vowels, in the spirit of intelligibility/comprehensibility, but with objective measures.
- Structure of the presentation
  - Establish metrics for the intrinsic evaluation of L2 vowels.
  - Test 1: L2 English learners (L1: Italian & French)
  - Test 2: L2 French learners (L1: Italian)
  - Test 3: intrinsic vs extrinsic assessment of L2 French learners (L1: Spanish and English)
  - Conclusions

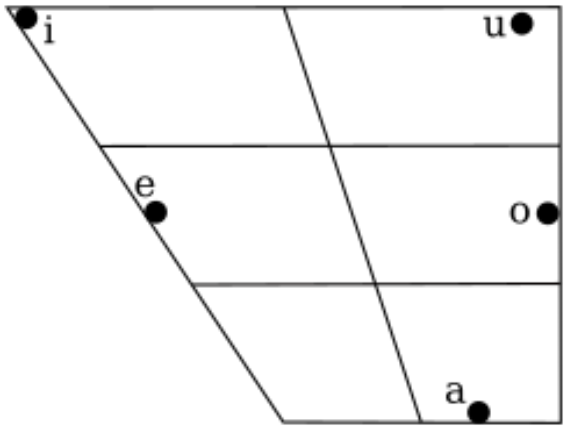
# Assessing L2 English & L2 French vowels *intrinsically*

# Learning the English vowel system

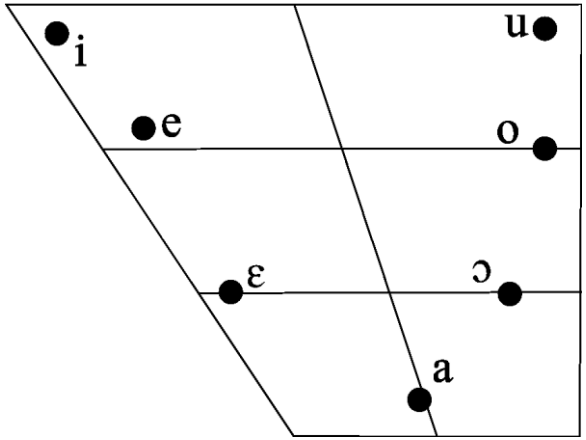
SBE



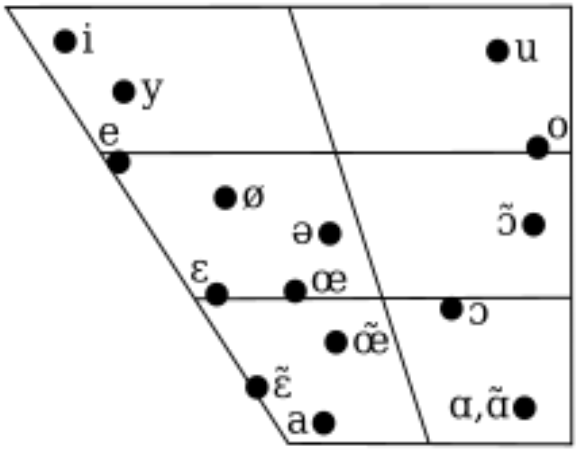

Spanish & Northern Italian




Standard Italian

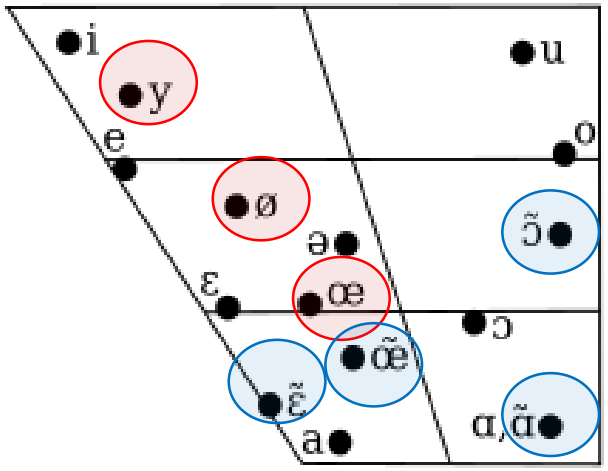



Parisian French

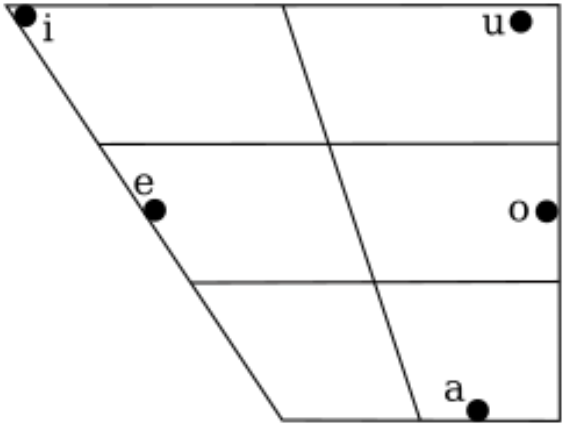


# Learning the French vowel system

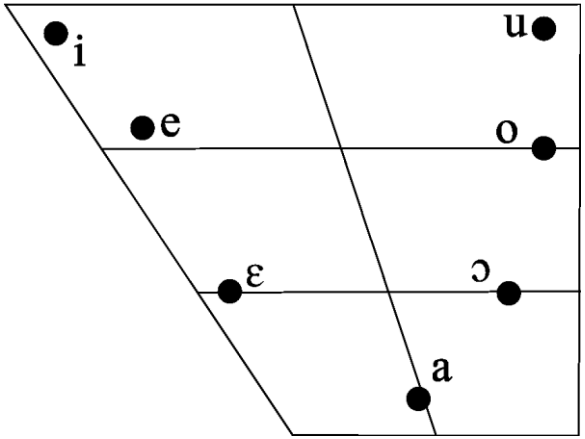
Parisian French



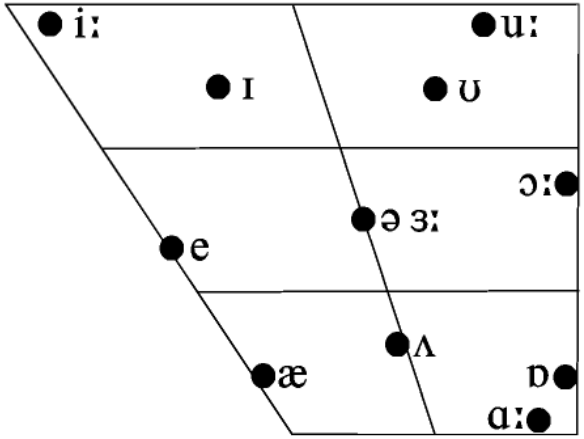

Spanish & Northern Italian




Standard Italian



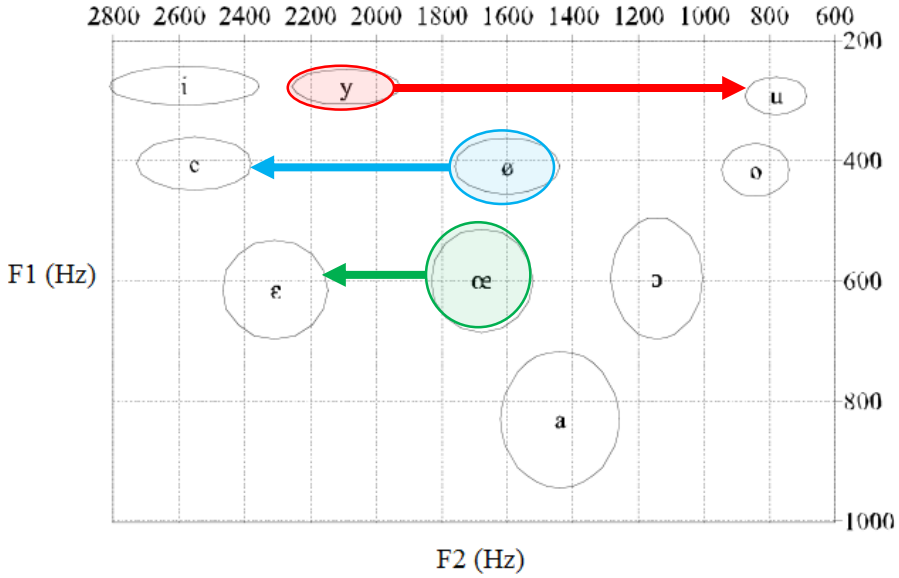

SBE

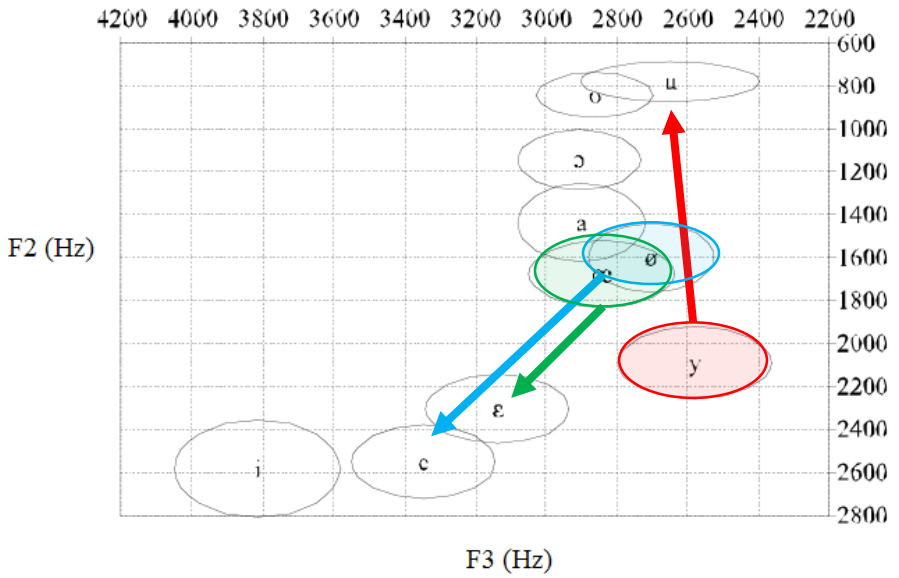


In order to assess the pronunciation
of L2 vowels intrinsically,
we use 3 metrics…

# 1. Acoustic distance of vowel pairs: *Euclidean distances*

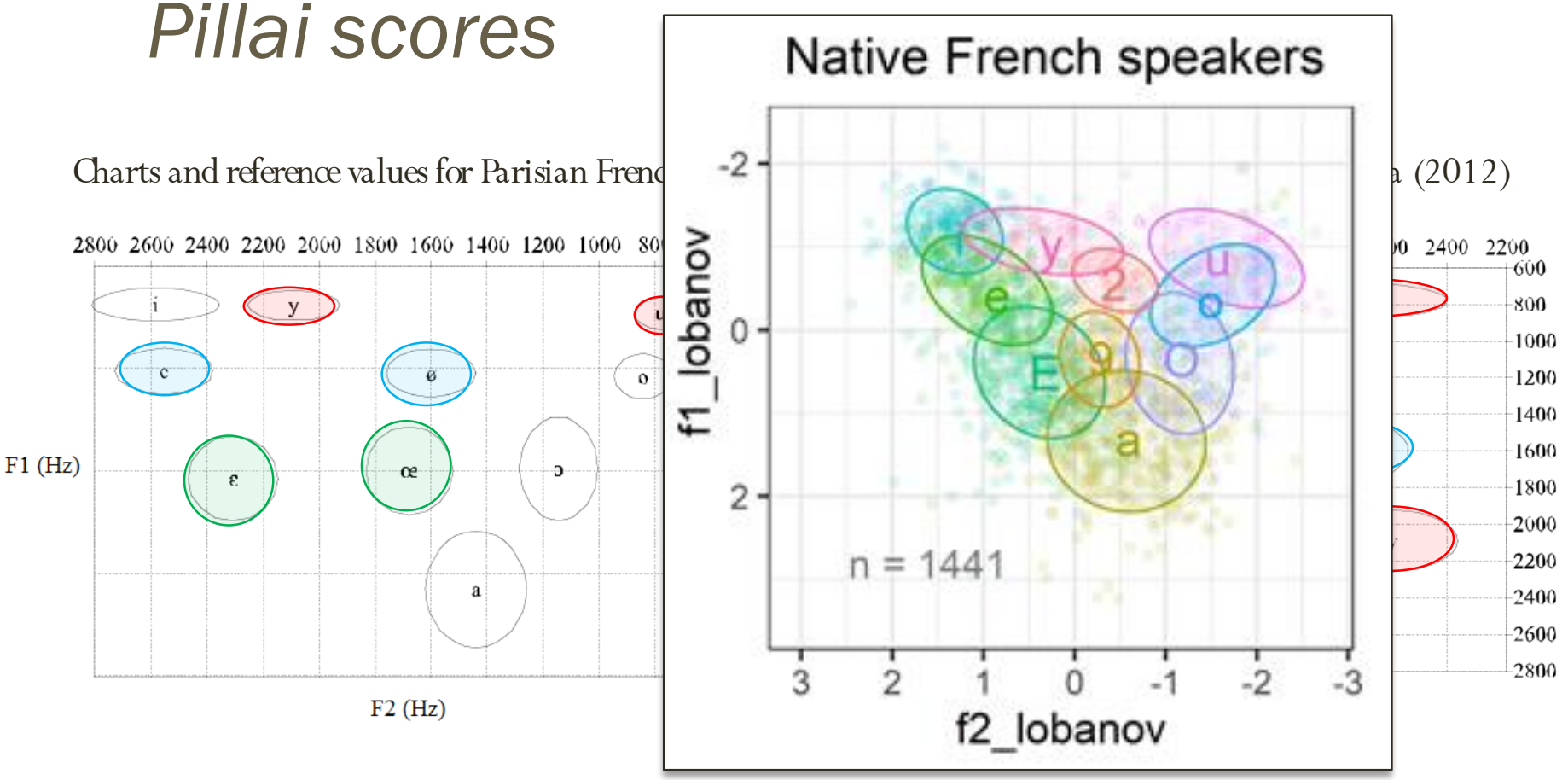
Charts and reference values for Parisian French from: Georgeton, Paillerau, Landron, Gao & Kamiyama (2012)



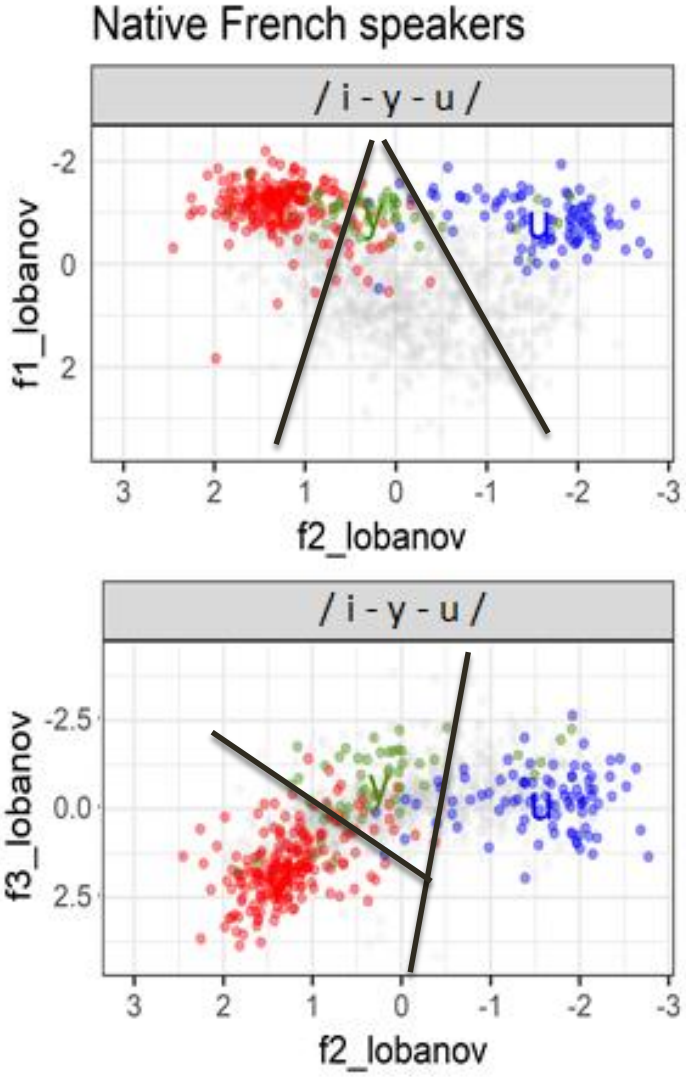

Assumption: the acoustic distance between target vowel pairs will be smaller (or null) for learners who have not yet established phonological categories for L2 French vowels /y, ø, œ/.

# 2. Acoustic overlap of vowel pairs: *Pillai scores*

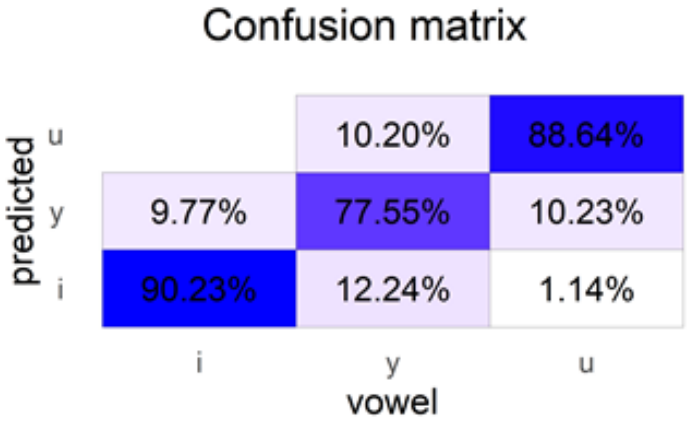



Assumption: the acoustic overlap between target vowel pairs will be greater for learners who have not yet established phonological categories for L2 French vowels /y, ø, œ/.

# 3. Acoustic overlap of vowel pairs: *LDA classification accuracy*




Linear discriminant analysis (LDA): statistical method for separating objects or observations that belong to distinct categories based on measured characteristics, and for classifying new observations in these categories.



# Test 1:
# Italian & French learners of L2 English

Data from the ICE-IPAC corpus of L2 English speech

Mairano, Bouzon, Capliez & De Iacovo (2019) Acoustic distances, Pillai scores and LDA classification scores as metrics of L2 comprehensibility and nativelikeness. Proc. of ICPHS2019.

# Metrics recap

- Vowel metrics
  - Euclidean distances
  - Pillai scores
  - LDA classification accuracy

1. /iː/ – /ɪ/
2. /ɑː/ – /æ/
3. /ɔː/ – /ɒ/
4. /uː/ – /ʊ/

- Other metrics – for comparison
  - AR (Articulation Rate)
  - SR (Speech Rate)
  - PSR (Pause/Speech Ratio)
  - AVL (Average Pause Length)

fluency

  - VOT (Voice Onset Time for /p, t, k/)
  - Native ratings of nativelikeness (5 listeners, ICC = .92)
  - Native ratings of comprehensibility (5 listeners, ICC = .94)

# Participants

- 25 learners of L2 English from the ICE-IPAC corpus.

| Group | University | N | Age | Level | Gender |
|---|---|---|---|---|---|
| IT | Turin | 15 | 22.3 (±2.46) | B1-C1 | 11 F + 4 M |
| FR | Lille | 10 | 22.5 (±3.44) | B1-C1 | 8 F + 2 M |

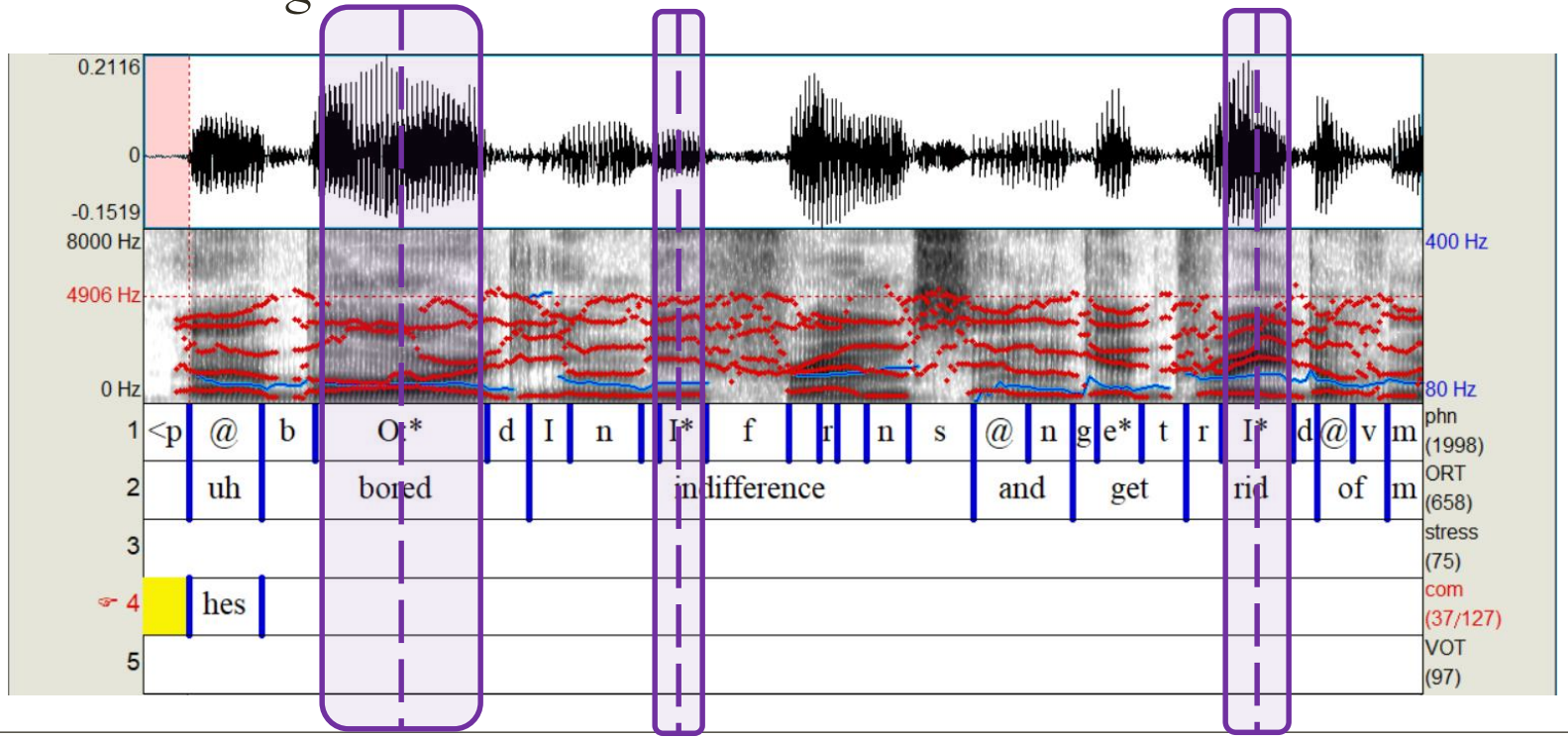
# Tasks

Protocol of the ICE-IPAC L2 English corpus
(Andreassen, Herry-Bénit, Kamiyama & Lacoste, 2015)

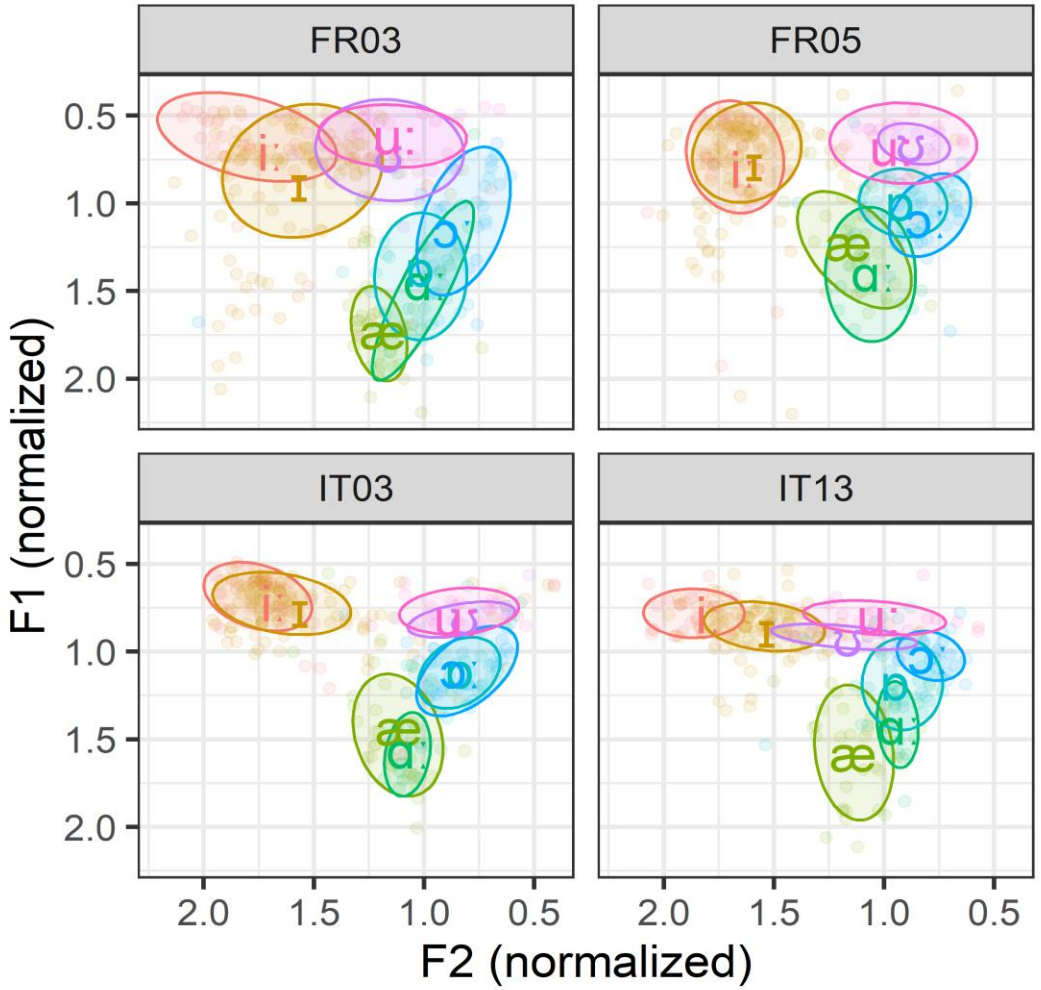
- Linguistic background questionnaire
- Word list read aloud task
- Word list repetition task
- Read aloud task (1 newspaper article, 506 words)
- Dialogue with peer
- Dialogue with native
- Read aloud task in L1 (1 newspaper article, 498 words)

# Data annotation

1. Orthographic transcription of IPUs
2. Automatic phonetization and forced alignment with WebMAUS
3. Manual check in Praat
4. Formant extraction with a Praat script
5. Filtering and normalization in R



# Vowel metrics for four sample learners of L2 English

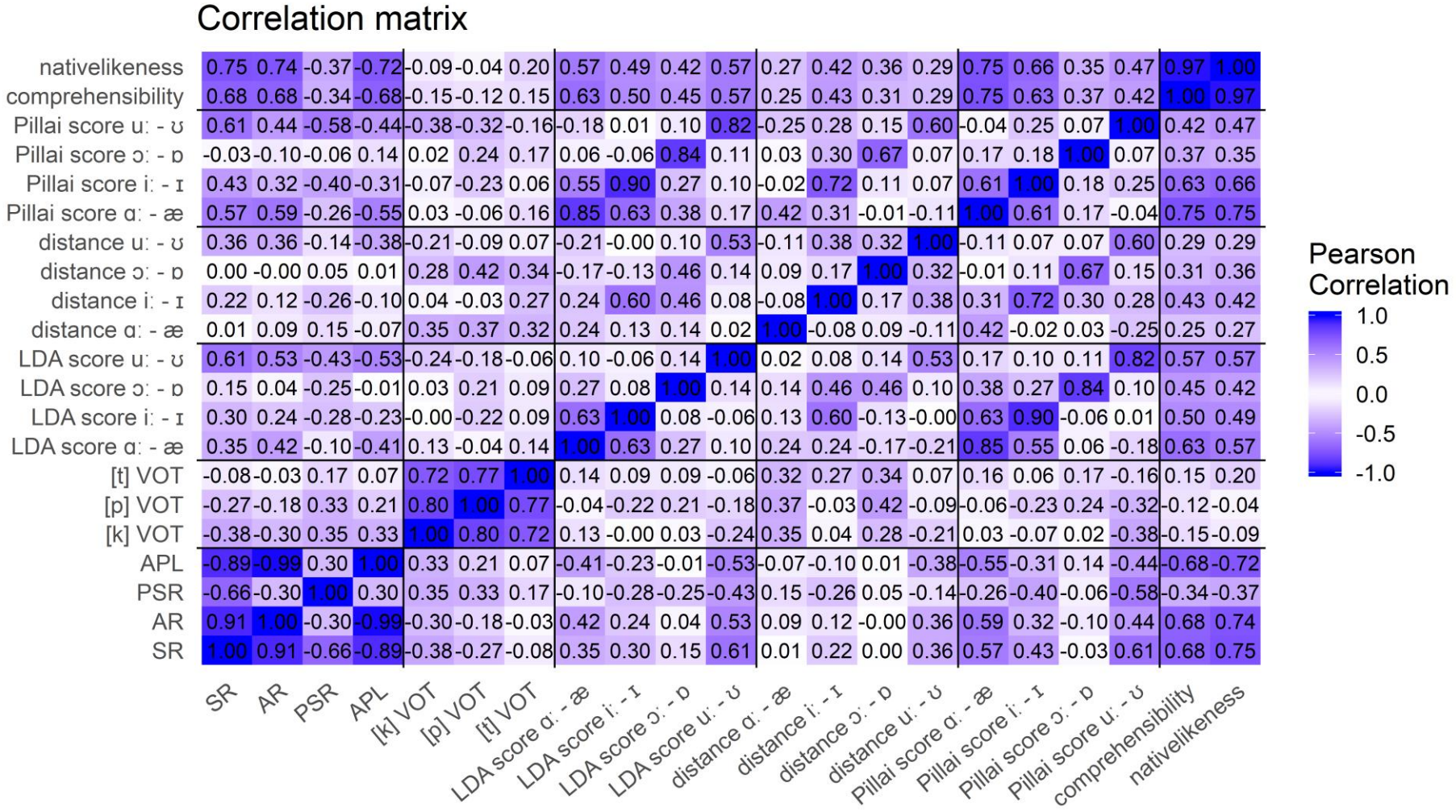



Euclidean distances

| | FR03 | FR05 | IT03 | IT13 |
|---|---|---|---|---|
| /iː - ɪ/ | 0.049 | 0.004 | 0.015 | 0.075 |
| /uː - ʊ/ | 0.011 | 0.008 | 0.002 | 0.035 |
| /ɑː - æ/ | 0.115 | 0.043 | 0.018 | 0.067 |
| /ɔː - ɒ/ | 0.074 | 0.024 | 0.001 | 0.048 |

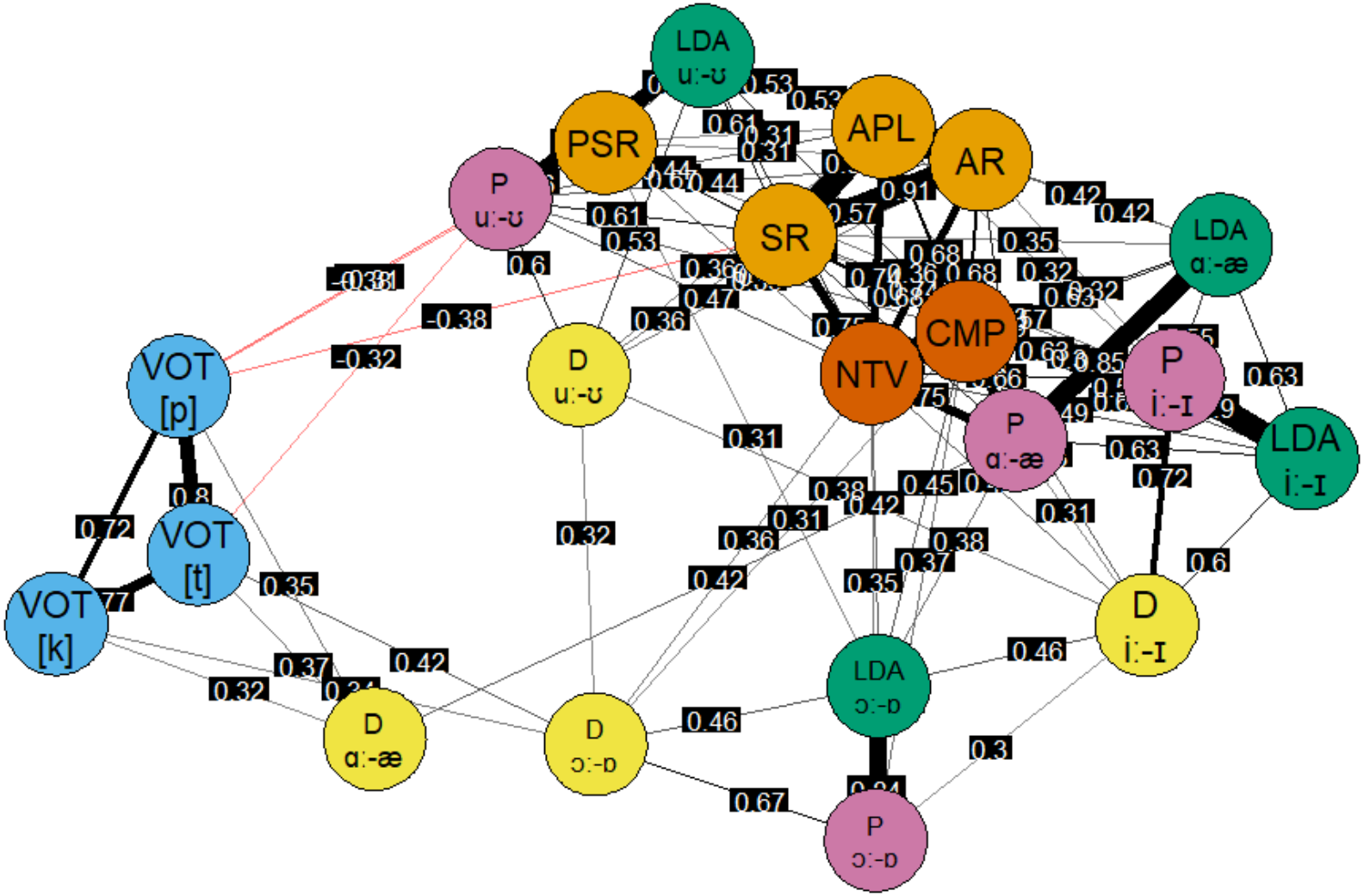
Pillai scores

| | FR03 | FR05 | IT03 | IT13 |
|---|---|---|---|---|
| /iː - ɪ/ | 0.384 | 0.125 | 0.308 | 0.487 |
| /uː - ʊ/ | 0.347 | 0.166 | 0.281 | 0.482 |
| /ɑː - æ/ | 0.449 | 0.270 | 0.362 | 0.445 |
| /ɔː - ɒ/ | 0.110 | 0.100 | 0.034 | 0.492 |


Correlation matrix
nativelikeness 0.75 0.74 -0.37 -0.72 -0.09 -0.04 0.20 0.57 0.49 0.42 0.57 0.27 0.42 0.36 0.29 0.75 0.66 0.35 0.47 0.97 1.00
comprehensibility 0.68 0.68 -0.34 -0.68 -0.15 -0.12 0.15 0.63 0.50 0.45 0.57 0.25 0.43 0.31 0.29 0.75 0.63 0.37 0.42 1.00 0.97
Pillai score uː - ʊ 0.61 0.44 -0.58 -0.44 -0.38 -0.32 -0.16 -0.18 0.01 0.10 0.82 -0.25 0.28 0.15 0.60 -0.04 0.25 0.07 1.00 0.42 0.47
Pillai score ɔː - ɒ -0.03 -0.10 -0.06 0.14 0.02 0.24 0.17 0.06 -0.06 0.84 0.11 0.03 0.30 0.67 0.07 0.17 0.18 1.00 0.07 0.37 0.35
Pillai score iː - ɪ 0.43 0.32 -0.40 -0.31 -0.07 -0.23 0.06 0.55 0.90 0.27 0.10 -0.02 0.72 0.11 0.07 0.61 1.00 0.18 0.25 0.63 0.66
Pillai score ɑː - æ 0.57 0.59 -0.26 -0.55 0.03 -0.06 0.16 0.85 0.63 0.38 0.17 0.42 0.31 -0.01 -0.11 1.00 0.61 0.17 -0.04 0.75 0.75
distance uː - ʊ 0.36 0.36 -0.14 -0.38 -0.21 -0.09 0.07 -0.21 -0.00 0.10 0.53 -0.11 0.38 0.32 1.00 -0.11 0.07 0.07 0.60 0.29 0.29
distance ɔː - ɒ 0.00 -0.00 0.05 0.01 0.28 0.42 0.34 -0.17 -0.13 0.46 0.14 0.09 0.17 1.00 0.32 -0.01 0.11 0.67 0.15 0.31 0.36
distance iː - ɪ 0.22 0.12 -0.26 -0.10 0.04 -0.03 0.27 0.24 0.60 0.46 0.08 -0.08 1.00 0.17 0.38 0.31 0.72 0.30 0.28 0.43 0.42
distance ɑː - æ 0.01 0.09 0.15 -0.07 0.35 0.37 0.32 0.24 0.13 0.14 0.02 1.00 -0.08 0.09 -0.11 0.42 -0.02 0.03 -0.25 0.25 0.27
LDA score uː - ʊ 0.61 0.53 -0.43 -0.53 -0.24 -0.18 -0.06 0.10 -0.06 0.14 1.00 0.02 0.08 0.14 0.53 0.17 0.10 0.11 0.82 0.57 0.57
LDA score ɔː - ɒ 0.15 0.04 -0.25 -0.01 0.03 0.21 0.09 0.27 0.08 1.00 0.14 0.14 0.46 0.46 0.10 0.38 0.27 0.84 0.10 0.45 0.42
LDA score iː - ɪ 0.30 0.24 -0.28 -0.23 -0.00 -0.22 0.09 0.63 1.00 0.08 -0.06 0.13 0.60 -0.13 -0.00 0.63 0.90 -0.06 0.01 0.50 0.49
LDA score ɑː - æ 0.35 0.42 -0.10 -0.41 0.13 -0.04 0.14 1.00 0.63 0.27 0.10 0.24 0.24 -0.17 -0.21 0.85 0.55 0.06 -0.18 0.63 0.57
[t] VOT -0.08 -0.03 0.17 0.07 0.72 0.77 1.00 0.14 0.09 0.09 -0.06 0.32 0.27 0.34 0.07 0.16 0.06 0.17 -0.16 0.15 0.20
[p] VOT -0.27 -0.18 0.33 0.21 0.80 1.00 0.77 -0.04 -0.22 0.21 -0.18 0.37 -0.03 0.42 -0.09 -0.06 -0.23 0.24 -0.32 -0.12 -0.04
[k] VOT -0.38 -0.30 0.35 0.33 1.00 0.80 0.72 0.13 -0.00 0.03 -0.24 0.35 0.04 0.28 -0.21 0.03 -0.07 0.02 -0.38 -0.15 -0.09
APL -0.89 -0.99 0.30 1.00 0.33 0.21 0.07 -0.41 -0.23 -0.01 -0.53 -0.07 -0.10 0.01 -0.38 -0.55 -0.31 0.14 -0.44 -0.68 -0.72
PSR -0.66 -0.30 1.00 0.30 0.35 0.33 0.17 -0.10 -0.28 -0.25 -0.43 0.15 -0.26 0.05 -0.14 -0.26 -0.40 -0.06 -0.58 -0.34 -0.37
AR 0.91 1.00 -0.30 -0.99 -0.30 -0.18 -0.03 0.42 0.24 0.04 0.53 0.09 0.12 -0.00 0.36 0.59 0.32 -0.10 0.44 0.68 0.74
SR 1.00 0.91 -0.66 -0.89 -0.38 -0.27 -0.08 0.35 0.30 0.15 0.61 0.01 0.22 0.00 0.36 0.57 0.43 -0.03 0.61 0.68 0.75
SR
AR
PSR
APL
[k] VOT
[p] VOT
[t] VOT
LDA score ɑː - æ
LDA score iː - ɪ
LDA score ɔː - ɒ
LDA score uː - ʊ
distance ɑː - æ
distance iː - ɪ
distance ɔː - ɒ
distance uː - ʊ
Pillai score ɑː - æ
Pillai score iː - ɪ
Pillai score ɔː - ɒ
Pillai score uː - ʊ
comprehensibility
nativelikeness
Pearson Correlation
1.0
0.5
0.0
-0.5
-1.0

# Relations among metrics

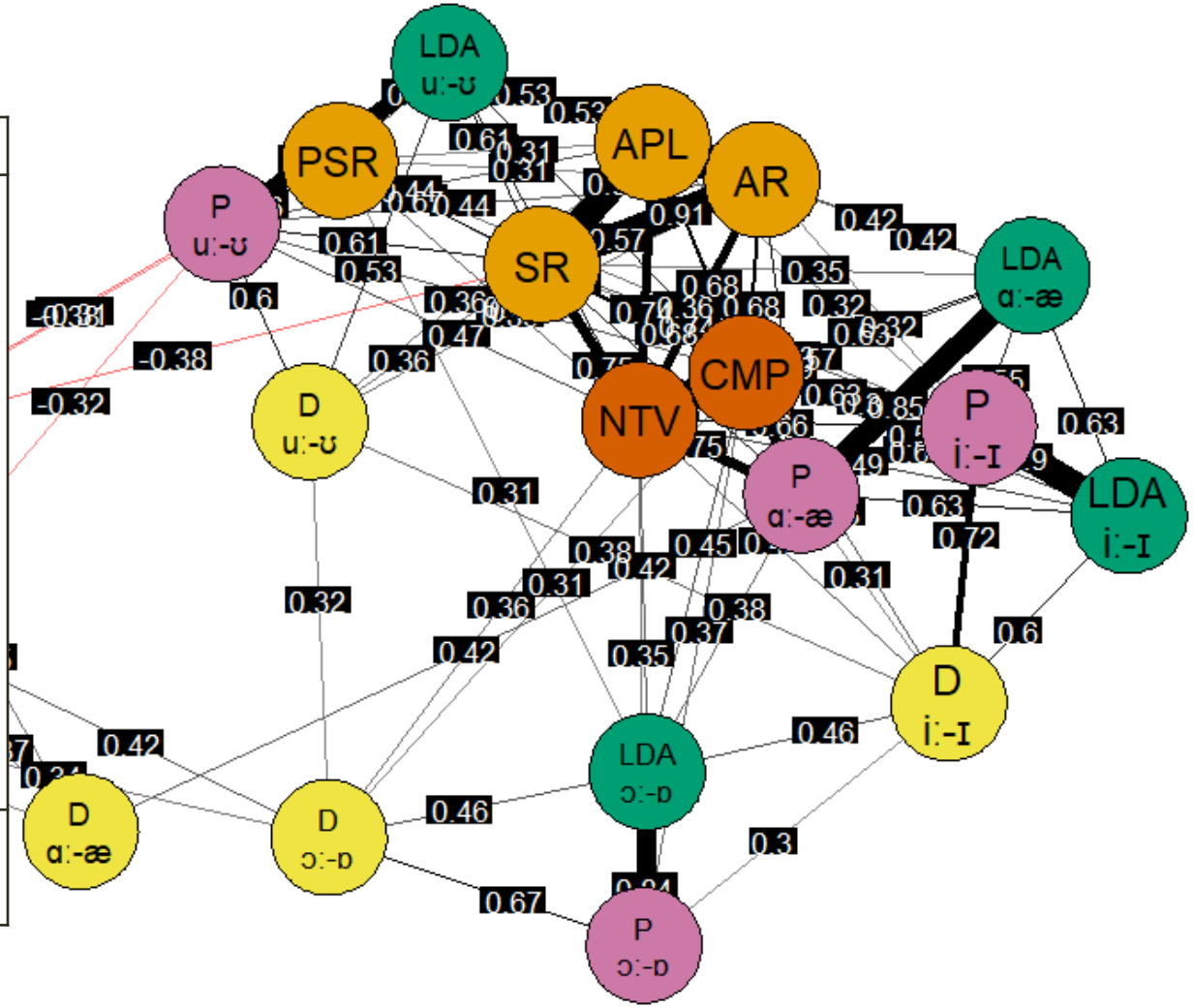
(*Fruchterman-Reingold* graph)



# Relations among metrics

## (*Fruchterman-Reingold* graph)

| Lm for CMP | |
|---|---|
| SR | .049* |
| PSR | .013* |
| VOT /t/ | .002** |
| LDA ɑː-æ | <.001*** |
| LDA iː-ɪ | .087. |
| D ɑː-æ | .032* |
| D iː-ɪ | .066. |
| D ɔː-ɒ | <.001*** |
| D uː-ʊ | .042* |
| P iː-ɪ | .096. |
| P uː-ʊ | <.001*** |
| Mult. $R^2$ | .96 |
| Adj. $R^2$ | .92 |



# Test 2:
# Italian learners of L2 French

Data from the ProSeg corpus of L2 French speech

Mairano & Santiago (under review) What vocabulary size tells us about pronunciation skills: Issues in assessing learners of L2 French.

# Metrics recap

- Vowel metrics
  - Euclidean distances
  - Pillai scores
  - LDA classification accuracy

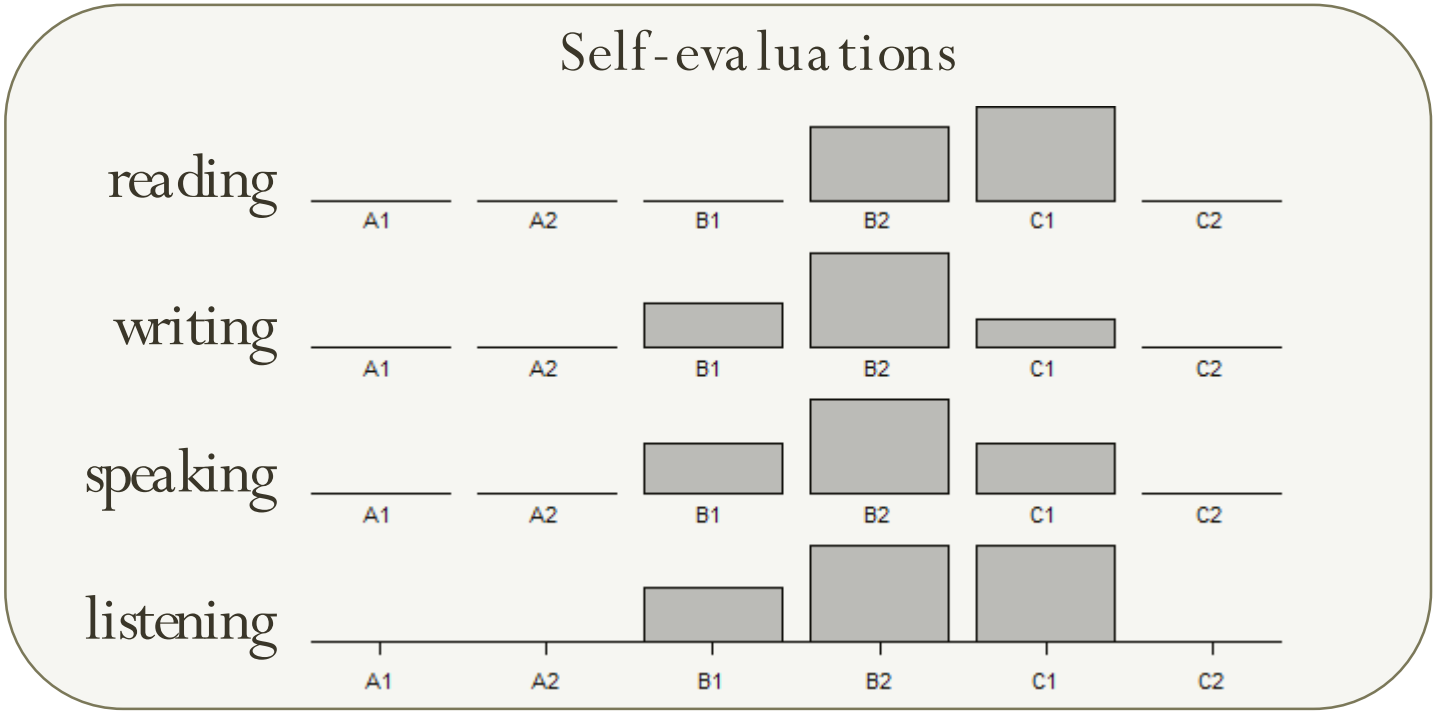
1. /y/ - /u/
2. /ø/ - /e/
3. /œ/ - /ɛ/

- Other metrics – for comparison
  - AR (Articulation Rate)
  - SR (Speech Rate)
  - NP (Number of Pauses in the first 5 mins of speech)
  - FA (ratings of Foreign Accentedness by 3 speakers, ICC = .89)

# Participants

- 25 Italian learners of L2 French recruited at the University of Turin (Italy)

Age: 24 (20 - 34)
Gender: 21 F, 4M
Months in FR speaking country: 3 (0 - 12)



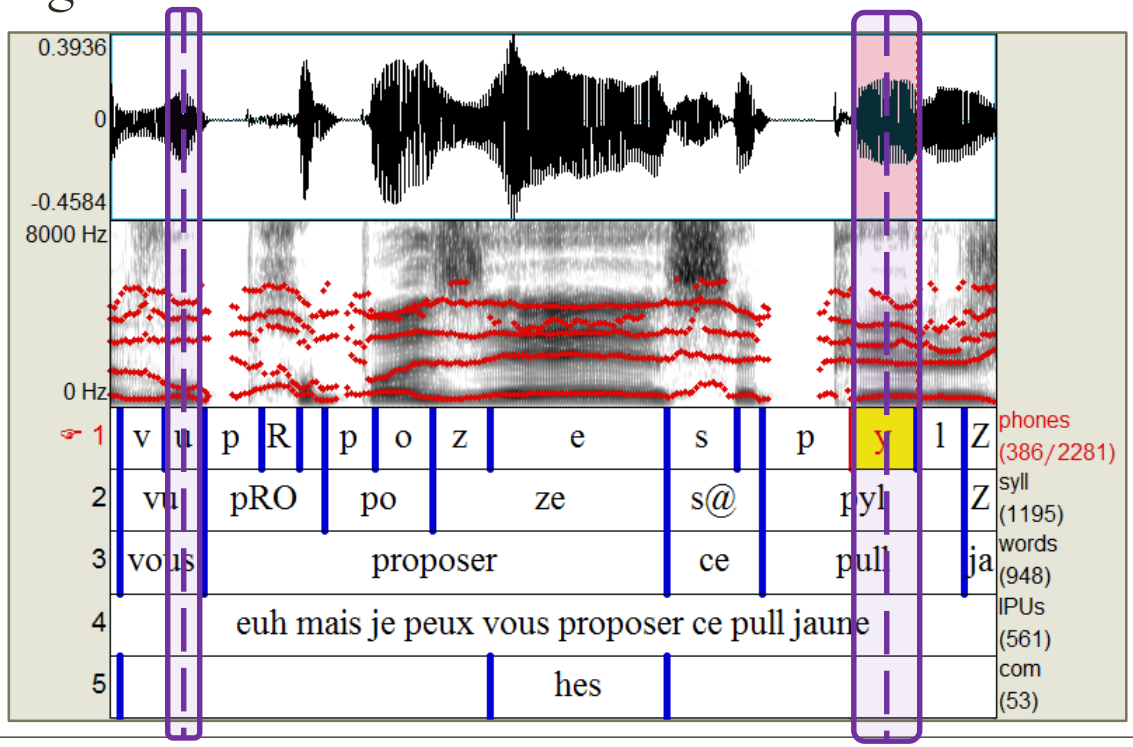
# Tasks

Protocol of the ProSeg L2 French corpus
(Delais-Roussarie, Kupisch, Mairano, Santiago & Splendido, EUROSLA2018)

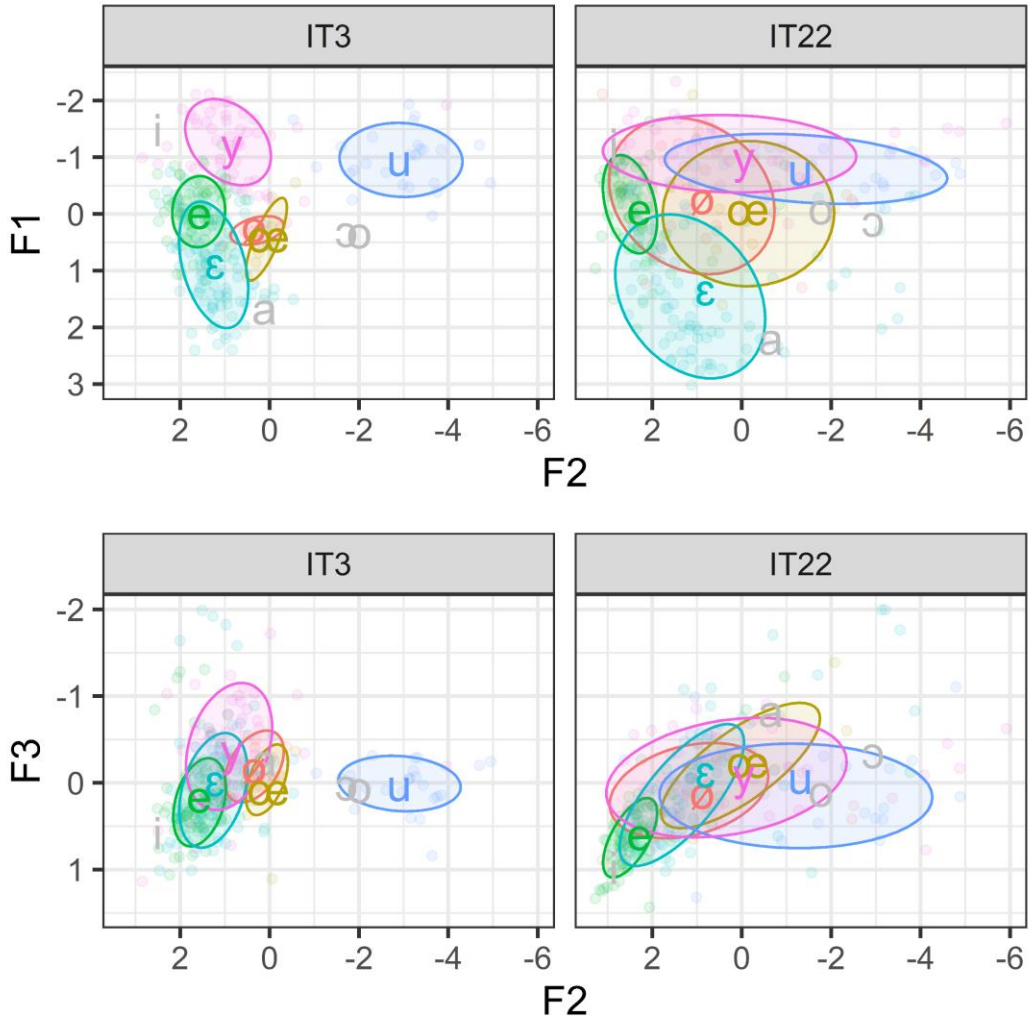
- Linguistic background questionnaire
- Dialang vocabulary test
- Read aloud task
  (8 short texts, 530 words)
- Picture description (~10 mins)
- Monologue (~10 mins)
- Read aloud task in L1
  (8 short texts, 427 words)

# Data annotation

1. Orthographic transcription of IPUs
2. Automatic phonetization and forced alignment with EasyAlign
3. Manual check in Praat
4. Formants' extraction with a Praat script
5. Filtering and normalization in R.

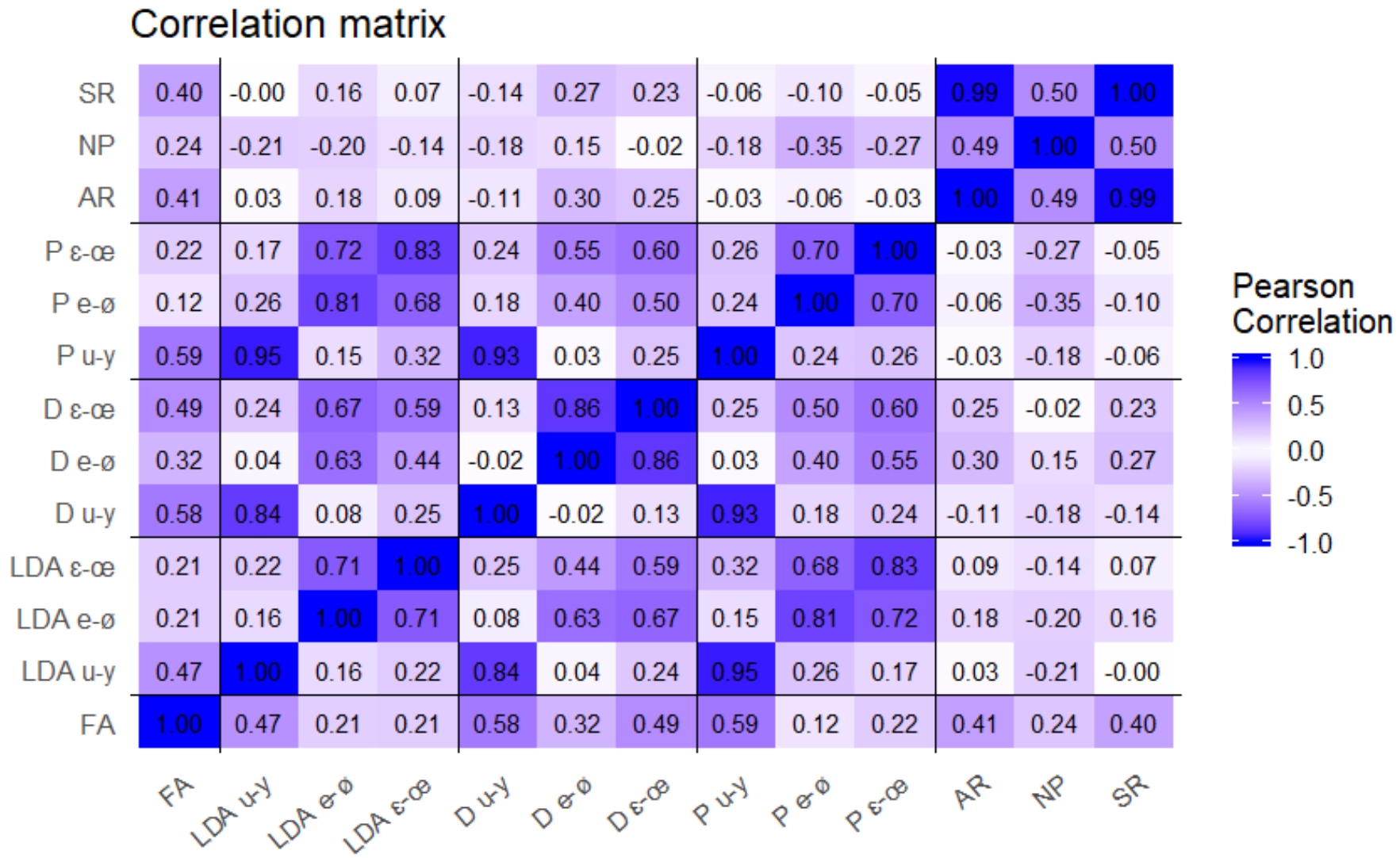


# Vowel metrics for two sample learners of L2 French



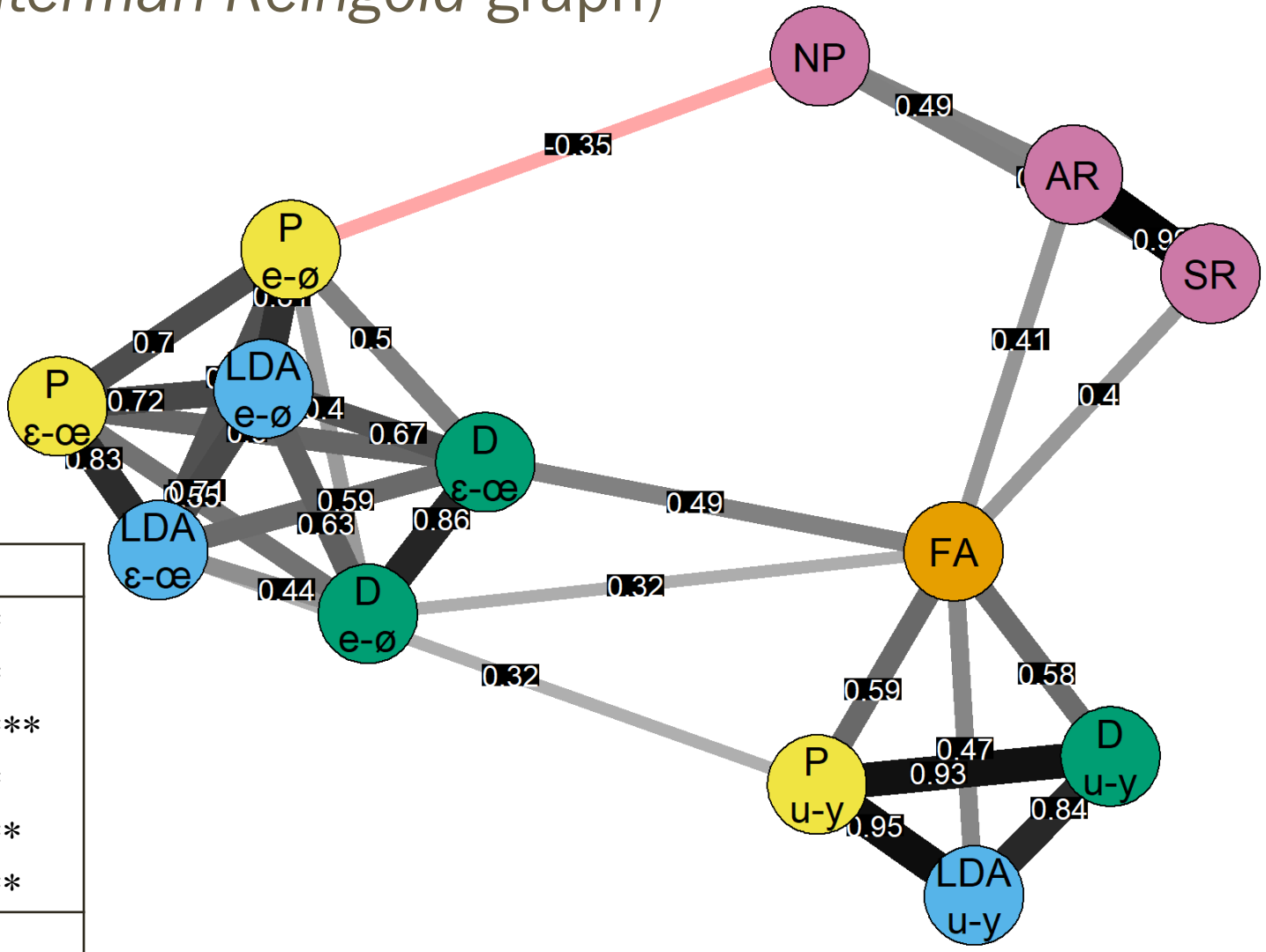

| | Euclidean distances (D) | | |
|---|---|---|---|
| | /y - u / | /ø - e/ | /œ - ɛ/ |
| IT3 | 4.23 | 1.62 | 1.43 |
| IT22 | 1.60 | 2.19 | 0.92 |

| | Pillai scores (P) | | |
|---|---|---|---|
| | /y - u / | /ø - e/ | /œ - ɛ/ |
| IT3 | 0.78 | 0.40 | 0.34 |
| IT22 | 0.10 | 0.22 | 0.18 |


Correlation matrix
SR 0.40 -0.00 0.16 0.07 -0.14 0.27 0.23 -0.06 -0.10 -0.05 0.99 0.50 1.00
NP 0.24 -0.21 -0.20 -0.14 -0.18 0.15 -0.02 -0.18 -0.35 -0.27 0.49 1.00 0.50
AR 0.41 0.03 0.18 0.09 -0.11 0.30 0.25 -0.03 -0.06 -0.03 1.00 0.49 0.99
P ɛ-œ 0.22 0.17 0.72 0.83 0.24 0.55 0.60 0.26 0.70 1.00 -0.03 -0.27 -0.05
P e-ø 0.12 0.26 0.81 0.68 0.18 0.40 0.50 0.24 1.00 0.70 -0.06 -0.35 -0.10
P u-y 0.59 0.95 0.15 0.32 0.93 0.03 0.25 1.00 0.24 0.26 -0.03 -0.18 -0.06
D ɛ-œ 0.49 0.24 0.67 0.59 0.13 0.86 1.00 0.25 0.50 0.60 0.25 -0.02 0.23
D e-ø 0.32 0.04 0.63 0.44 -0.02 1.00 0.86 0.03 0.40 0.55 0.30 0.15 0.27
D u-y 0.58 0.84 0.08 0.25 1.00 -0.02 0.13 0.93 0.18 0.24 -0.11 -0.18 -0.14
LDA ɛ-œ 0.21 0.22 0.71 1.00 0.25 0.44 0.59 0.32 0.68 0.83 0.09 -0.14 0.07
LDA e-ø 0.21 0.16 1.00 0.71 0.08 0.63 0.67 0.15 0.81 0.72 0.18 -0.20 0.16
LDA u-y 0.47 1.00 0.16 0.22 0.84 0.04 0.24 0.95 0.26 0.17 0.03 -0.21 -0.00
FA 1.00 0.47 0.21 0.21 0.58 0.32 0.49 0.59 0.12 0.22 0.41 0.24 0.40
FA
LDA u-y
LDA e-ø
LDA ɛ-œ
D u-y
D e-ø
D ɛ-œ
P u-y
P e-ø
P ɛ-œ
AR
NP
SR
Pearson Correlation
1.0
0.5
0.0
-0.5
-1.0

# Relations among metrics

## (*Fruchterman-Reingold* graph)




| Lm for FA | |
|---|---|
| LDA u-y | .032* |
| LDA ɛ-œ | .041* |
| D u-y | <.001*** |
| D e-ø | .048* |
| D ɛ-œ | .001** |
| SR | .007** |
| Mult. $R^2$ | .82 |
| Adj. $R^2$ | .75 |

# Test 3:
## L1 Spanish & L1 English learners of L2 French

Data from the COREIL and AixOx learner corpora

Goal: test intrinsic vs extrinsic assessment

# Metrics recap

- Vowel metrics

1. /i/ - /y/ - /u/
2. /e/ - /ø/ - /o/
3. /ɛ/ - /œ/ - /ɔ/

  - Extrinsic assessment:
    LDA classification accuracy after
    training on L1 native data
    - measure distinctness of vowel categories in L2 productions
      (=phonological development / intelligibility??)
  - Intrinsic assessment:
    LDA classification accuracy after
    training on each speaker's productions
    - measure similarity between L1 and L2 productions
      (=nativelikeness)

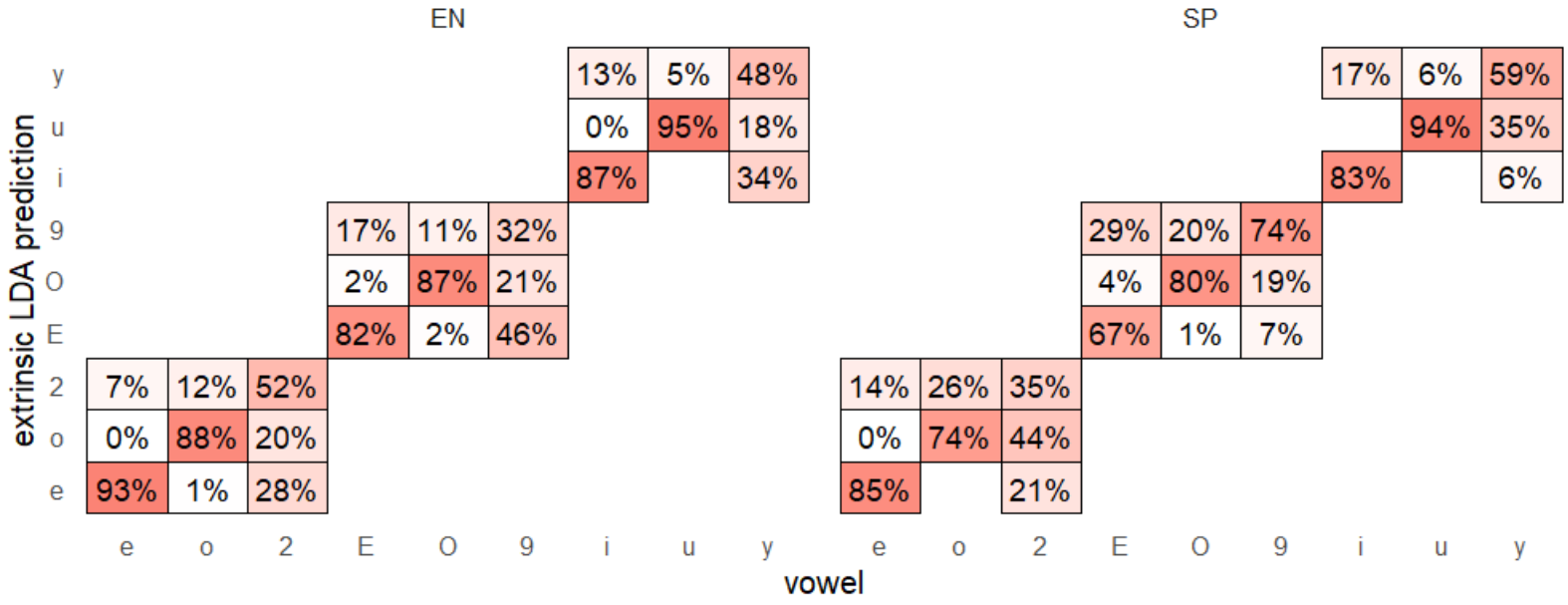
# Participants & task

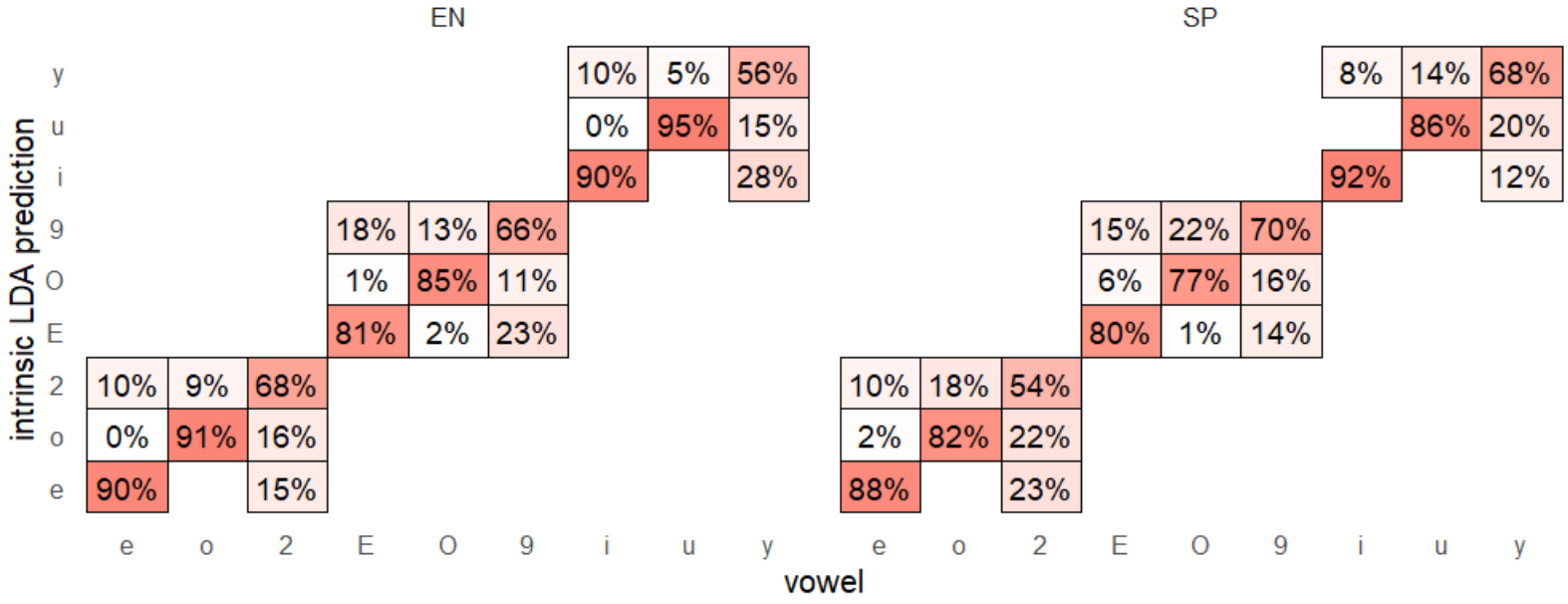
- 20 learners of L2 French (+ 10 control native speakers)

| Group | Origin | N | Age | Level | Corpus |
|---|---|---|---|---|---|
| FR (control) | Paris/Aix-en-Prov. | 5+5 | 35 (±14) | Native | COREIL + AixOx |
| SP | Mexico City | 10 | 25 (±6) | B1-B2 | COREIL |
| EN | Oxford (UK) | 10 | 22 (±2) | B1-B2 | AixOx |

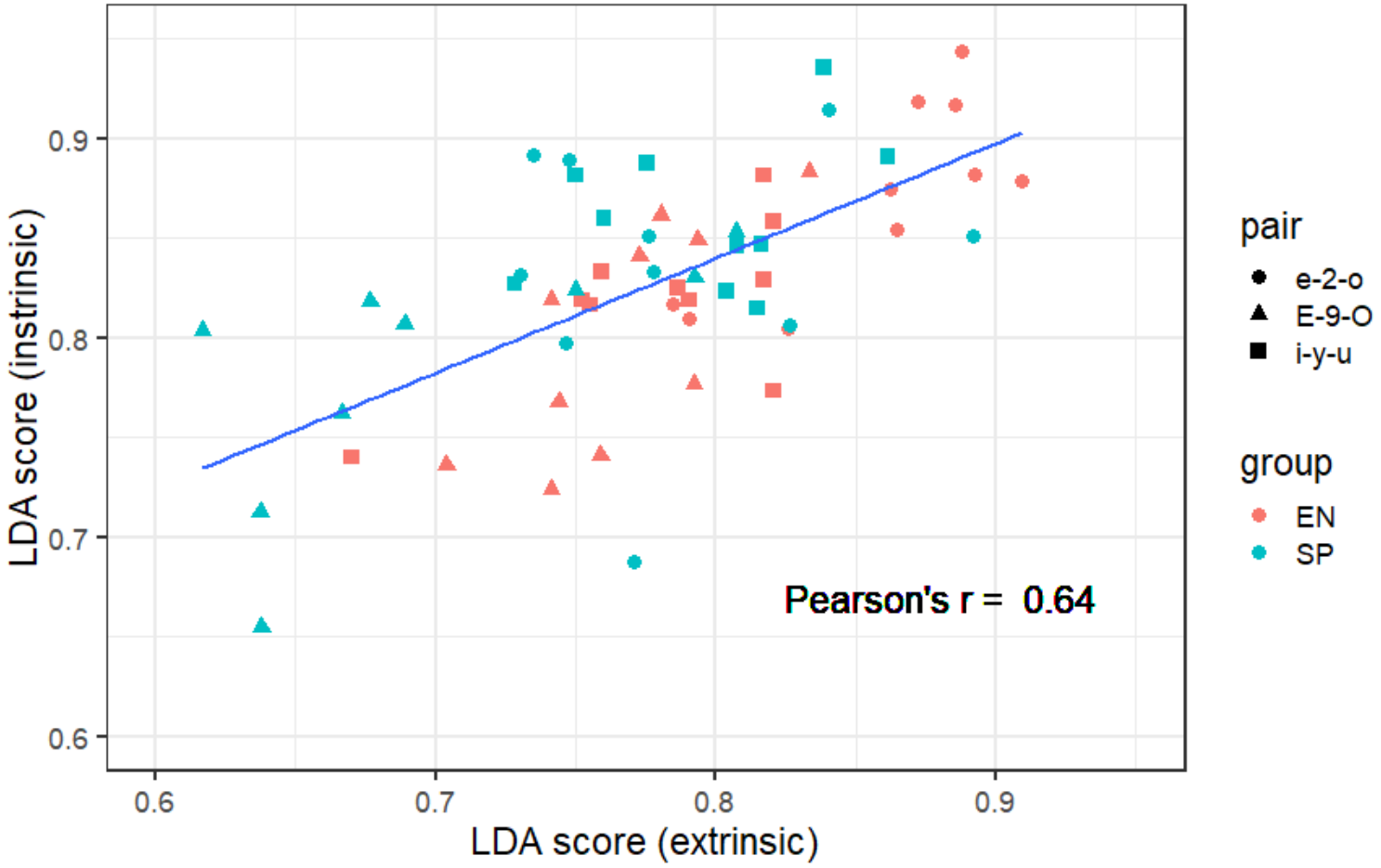
- Read aloud task
- Usual annotation and procedure for extracting formants in Praat

# LDA classification accuracy



# Correlation of extrinsic and intrinsic assessments



# Conclusion

# Summary of tests

- All our metrics for the intrinsic assessment of L2 vowels seem to correlate with native judgments, both in L2 English (test 1) and L2 French (test 2).
- Using our whole set of metrics we are able to explain 92% and 75% of variance for native ratings of comprehensibility / foreign-accentedness in our data of L2 English and L2 French.
- Intrinsic and extrinsic assessments with LDA show a reassuring relationship, represented by a correlation strength of $r = .64$.

# Limits of intrinsic assessment

- Needs a certain amount of speech data by each speaker
- Not suitable for students who expect to be scored on the basis of nativelikeness
- Not suitable for giving immediate feedback to students, only useful for test scoring
- Only for vowels, at the moment
- In its pure form, it can easily be 'fooled'

---

- High-quality audio is needed
- Potential issues caused by formant detection errors, etc.

# Advantages of intrinsic assessment

- Intrinsic assessment does not compare learners' pronunciation to a predefined model:
  - Does not evaluate students wrt a specific standard accent
  - May be useful for assessing the development of relevant phonological categories, i.e. in acquisitional research
  - Works in the spirit of intelligibility / comprehensibility rather than nativelikeness
  - Seems to correlate well with native ratings, as well as with extrinsic assessment

# Towards the automatic evaluation of L2 pronunciation using Pillai scores and LDA classification accuracy

Paolo Mairano & Fabian Santiago

## Thank you!