Methodological challenges in studying language variation and change:

combining fieldwork and empirical studies to overcome the observer's paradox

Karolina Broś

University of Warsaw

The observer’s paradox

situation in which we want to know
how speakers talk when not supervised
by a researcher but this knowledge can only be gained
by systematic observation (Labov 1972)

getting naturalistic data from speakers is a challenge for the researcher as (s)he needs to get a controlled sample that allows a reliable (statistical) analysis of the collected data

Problems

speakers tend to react to the presence of the linguist

(by suppressing certain features, using hypercorrection or a more formal register)

the social setting (or 'recording environment') itself can also alter the behaviour of the speaker

(presence of a recording device, having to sit in a lab or recording studio, the awareness of the fact that the recording will be heard or analysed by someone, the nature of the task: reading or repeating words and phrases)

What type of data do you use/collect?

How to study language variation and change?

inter- and intradialectal differences

age differences
gender differences
social differences
historical changes

varying data collection methods

How to collect data?

How natural is the speech we collect?

(lab speech, orthography)

Can we collect data remotely?

How does the way in which data are recorded affect data quality?

How to collect data - summary

fieldwork (different methods - elicitation, reading tasks, interaction/conversational speech, semi-structured interviews…)

experimental paradigms (word lists, sentence lists, carrier phrases, frame sentences, repetitions, orthographic bias…)

perception (perception tests: identification, discrimination, sound comparisons, accuracy, reaction times, online methods…)

articulation (EMA, EPG, electroglottography, ultrasound, motion capture)

remote data collection (Zoom, Google Meet, Microsoft Teams, different recording devices: iPad, iPhone, smartphone, laptop, professional recorders, guided/supervised vs self-recordings)

social media (WhatsApp, Messenger recordings, informed consent, data quality and filtering)

How to classify data?

Is segmentation appropriate for a continuous signal?

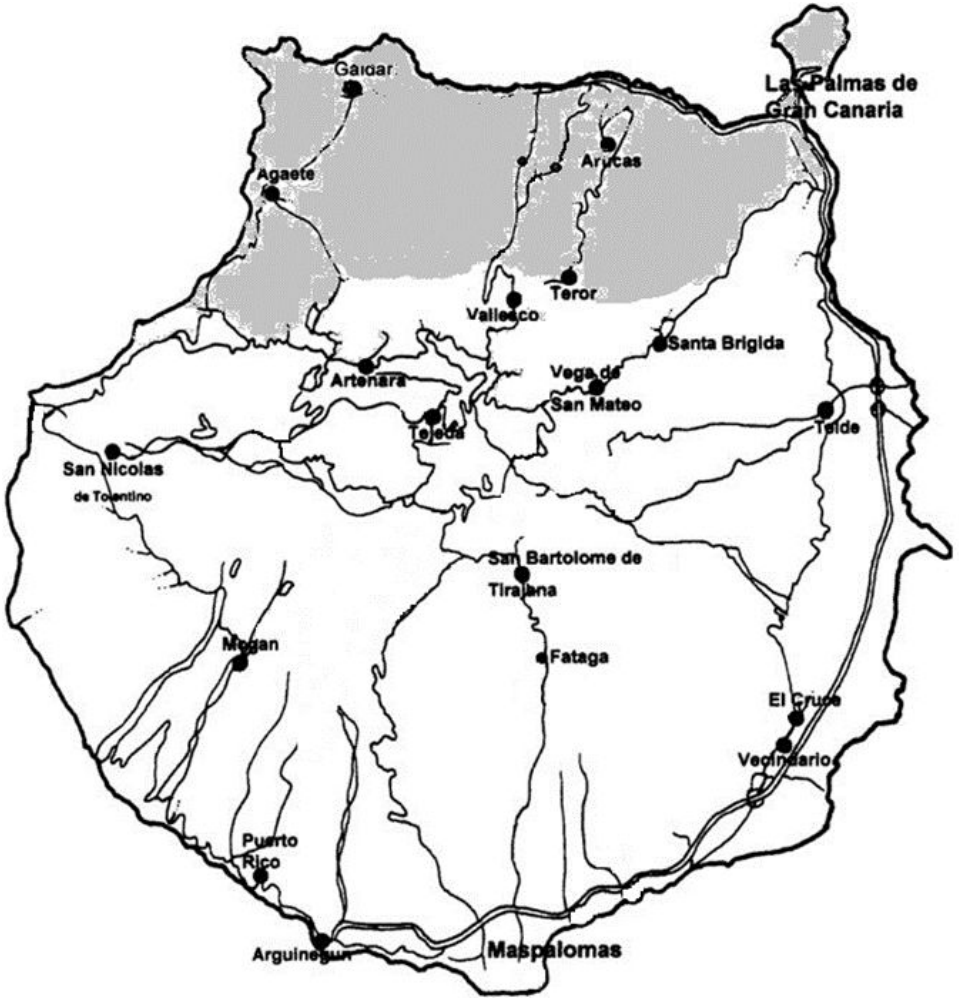
The data

[focus on one language variety]

Spanish spoken on Gran Canaria

Sources:

1) **Fieldwork/corpus: 44 native speakers, 111,317 phones, 16,454 post-vocalic /p t k b d g/**



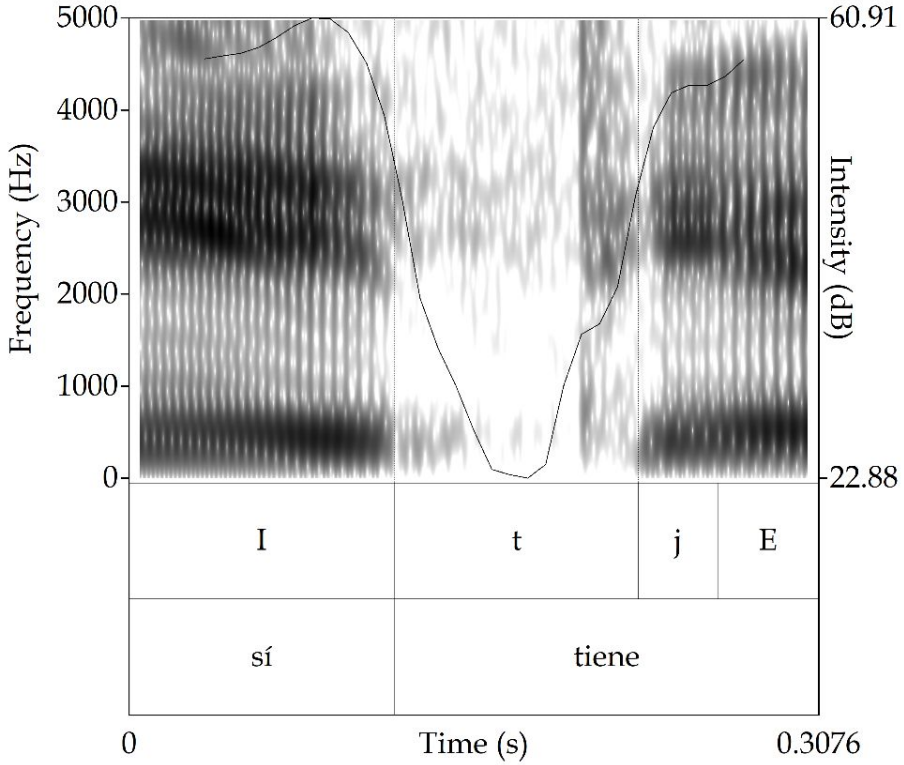
Sociolinguistic study of the dialect based on fieldwork data

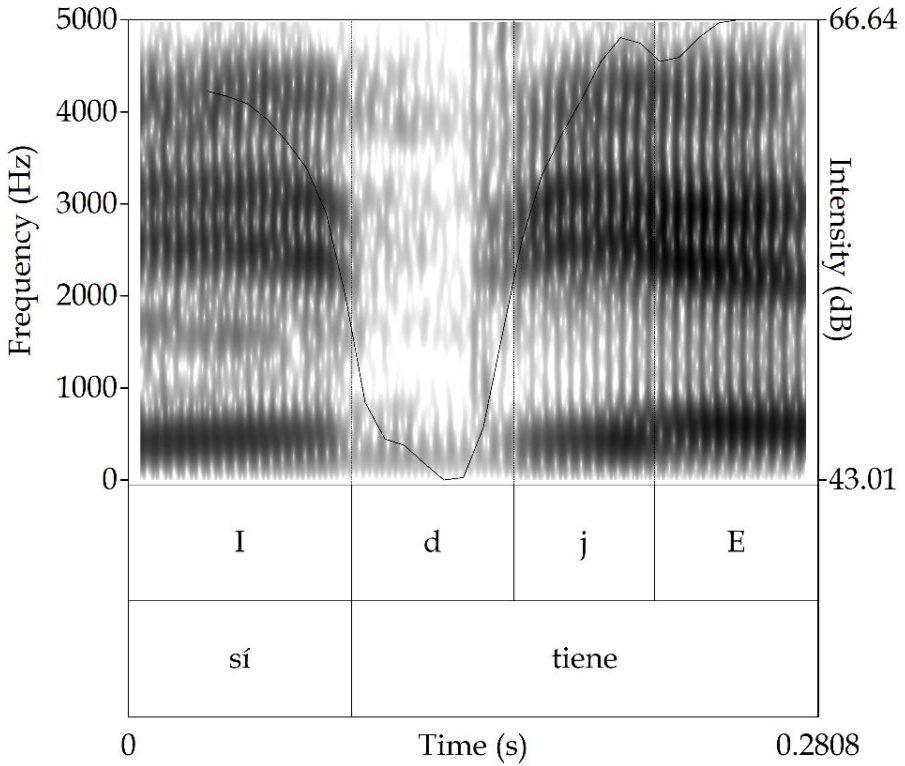
Factors: UR, phonology, social variables

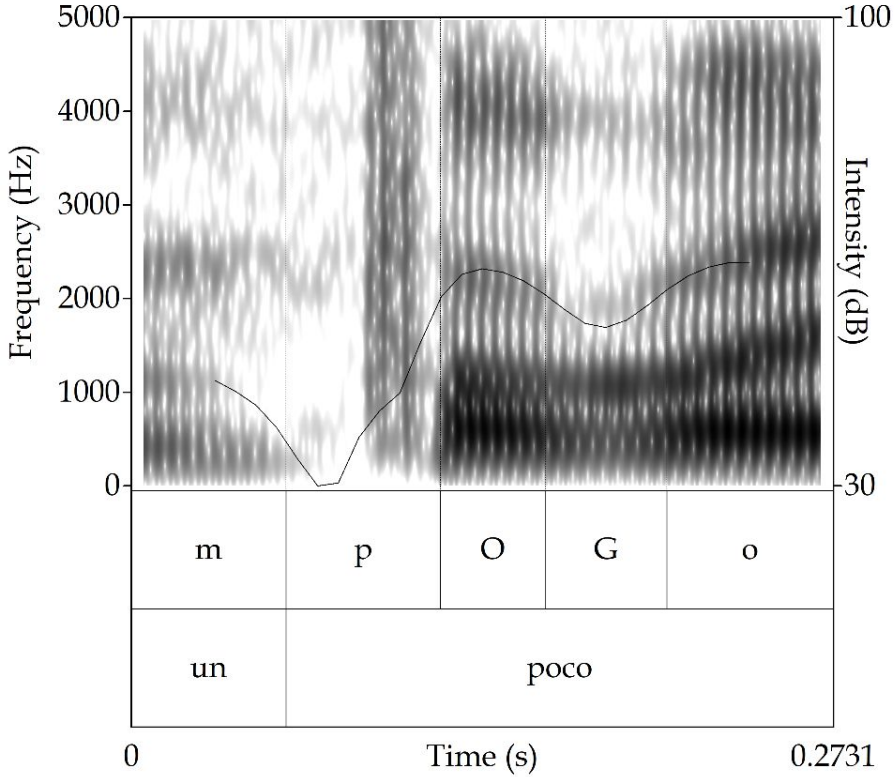
spontaneous speech

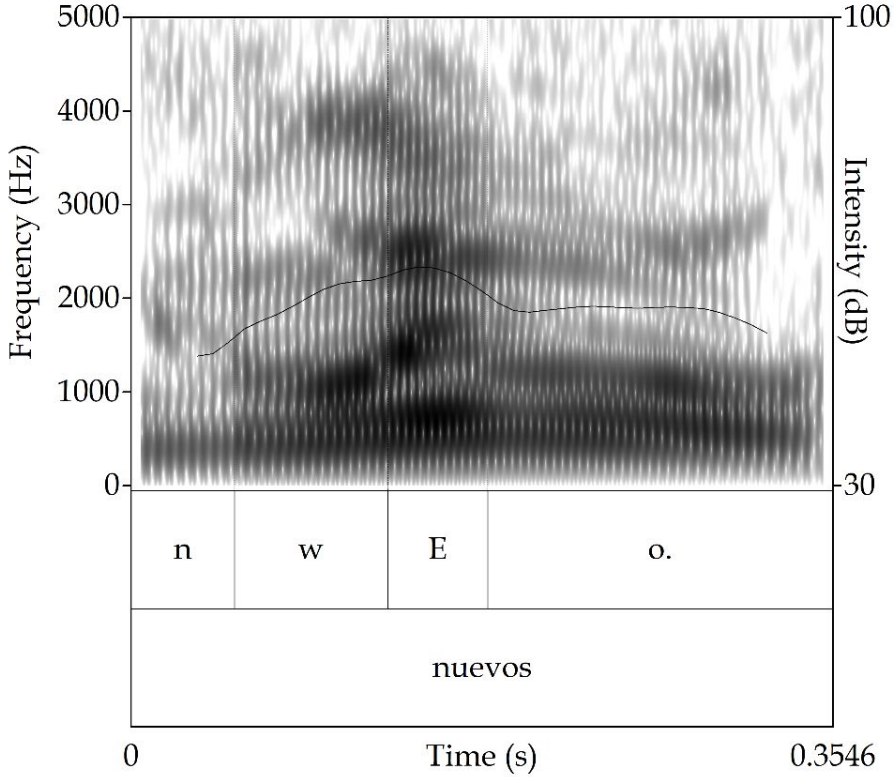
Full-fledged variation on Gran Canaria

UR	Example	voiceless stop	voiced stop	approximant	∅
/p/	*guapo* ‘pretty’	[ˈgwa.po]	[ˈgwa.bo]	[ˈgwa.β̞o]	[ˈgwa.o]
	se parece ‘is similar’	[se.pa.ˈɾe.se]	[se.ba.ˈɾe.se]	[se.β̞a.ˈɾe.se]	[se.a.ˈɾe.se]
	después ‘afterwards’	[de.ˈpwe]	[de.ˈbwe]	[de.ˈβ̞we]	
/b/	*cabeza* ‘head’			[ka.ˈβ̞esa]	[ka.ˈesa]
	la vela ‘the candle’		[la.ˈbe.la]	[la.ˈβ̞ela]	[la.ˈela]
	las velas ‘the candles’	[la.ˈpe.la]	[la.ˈbe.la]	[la.ˈβ̞ela]	


5000
4000
3000
2000
1000
0
Frequency (Hz)
60.91
Intensity (dB)
22.88
I
t
j
E
sí
tiene
0
Time (s)
0.3076


5000
4000
3000
2000
1000
0
Frequency (Hz)
66.64
Intensity (dB)
43.01
I
d
j
E
sí
tiene
0
Time (s)
0.2808


5000
4000
3000
2000
1000
0
Frequency (Hz)
100
30
Intensity (dB)
m
p
O
G
o
un
poco
0
Time (s)
0.2731


5000
4000
3000
2000
1000
0
Frequency (Hz)
100
30
Intensity (dB)
n
w
E
o.
nuevos
0
Time (s)
0.3546

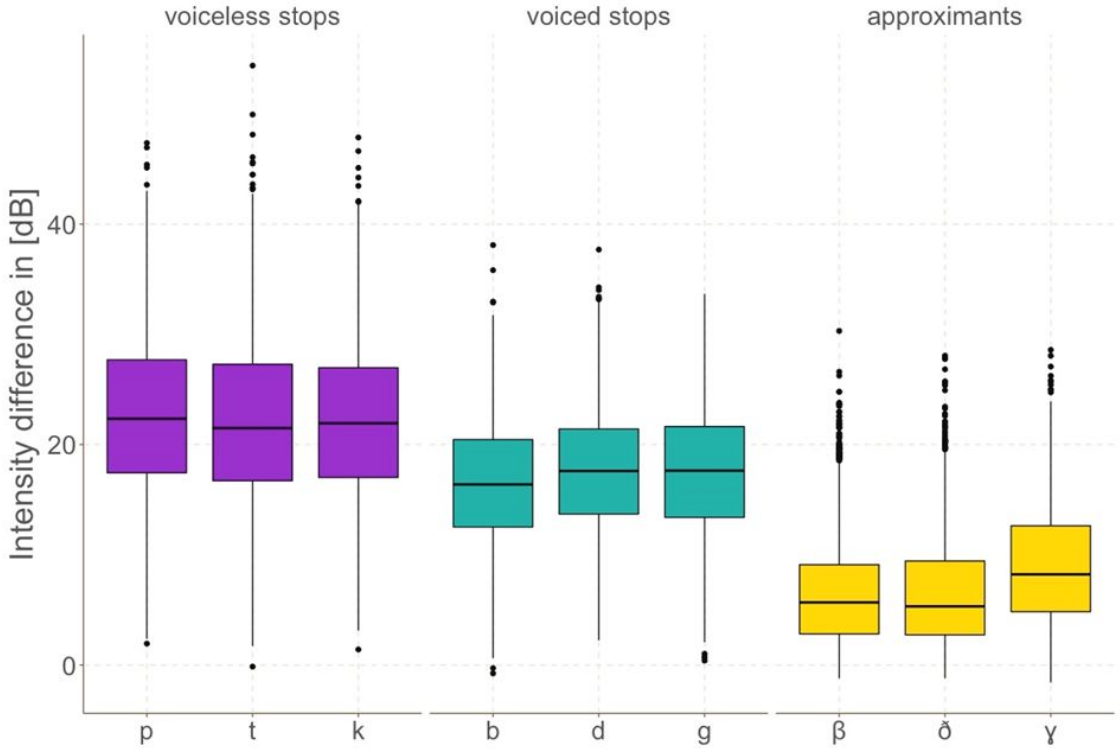
Research questions

- How systematic are the differences between surface sounds?
- Are underlying contrasts preserved?
- Which factors influence surface variation?

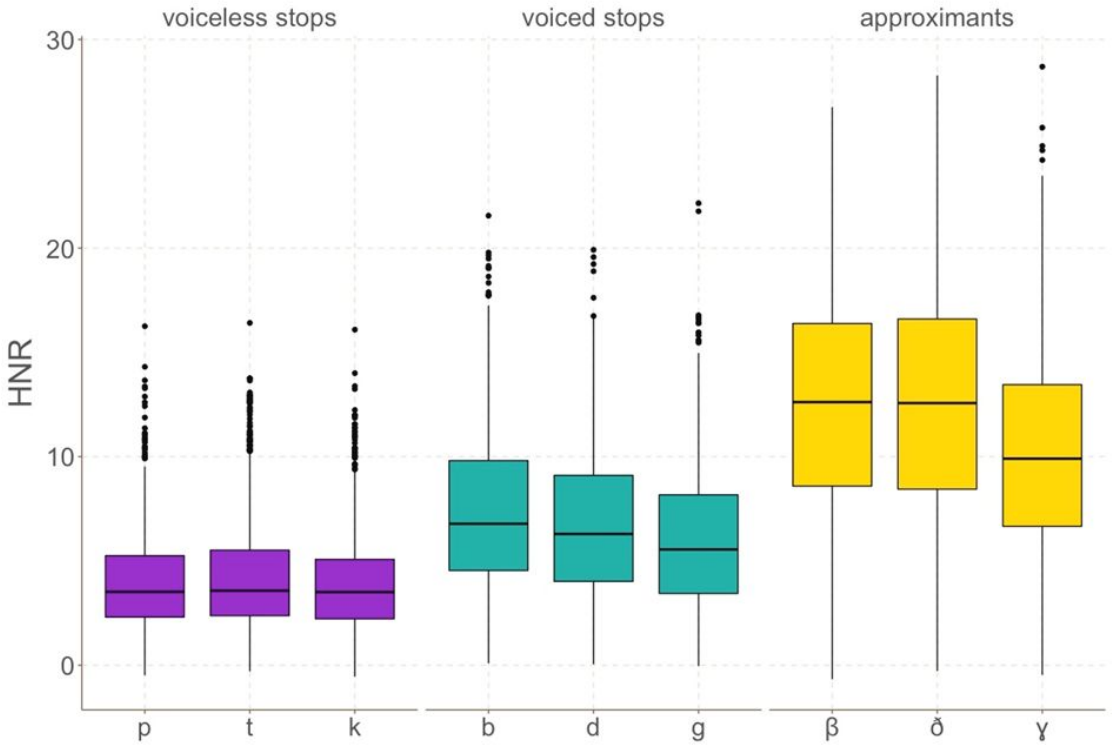
Measurements

- **intensity difference** (max intensity of the preceding vowel - min intensity of the target segment)
- **relative sound duration** (C/VC duration)
- **harmonics-to-noise ratio (HNR)**

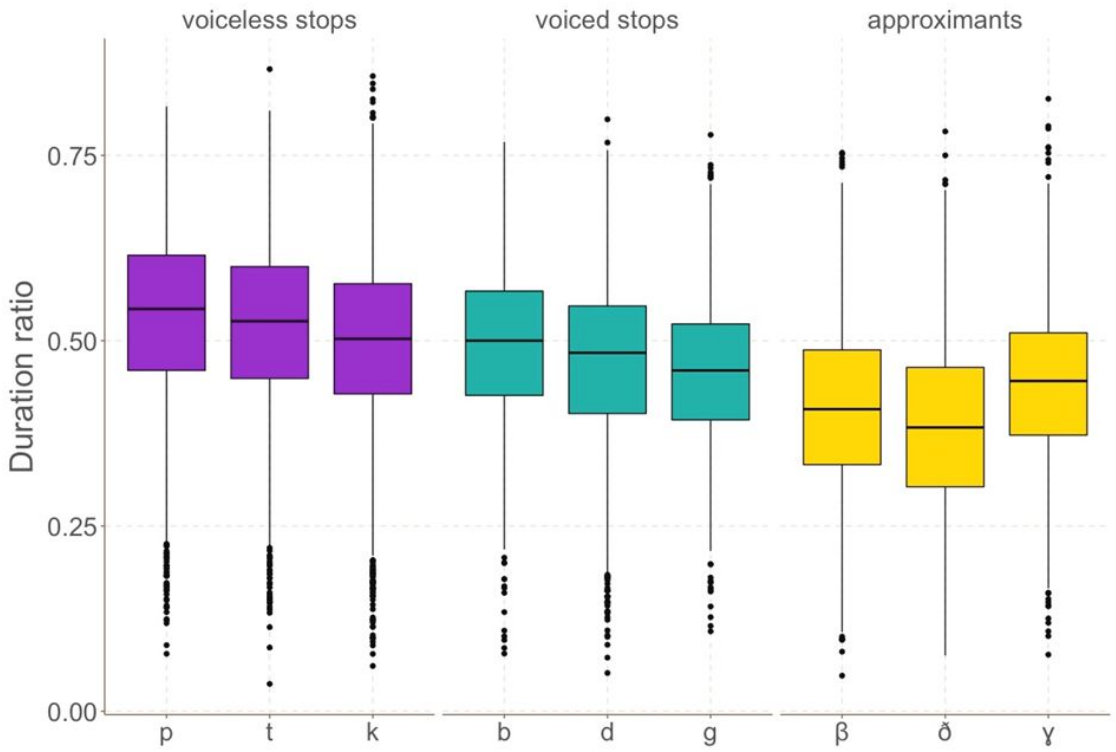
Surface differences



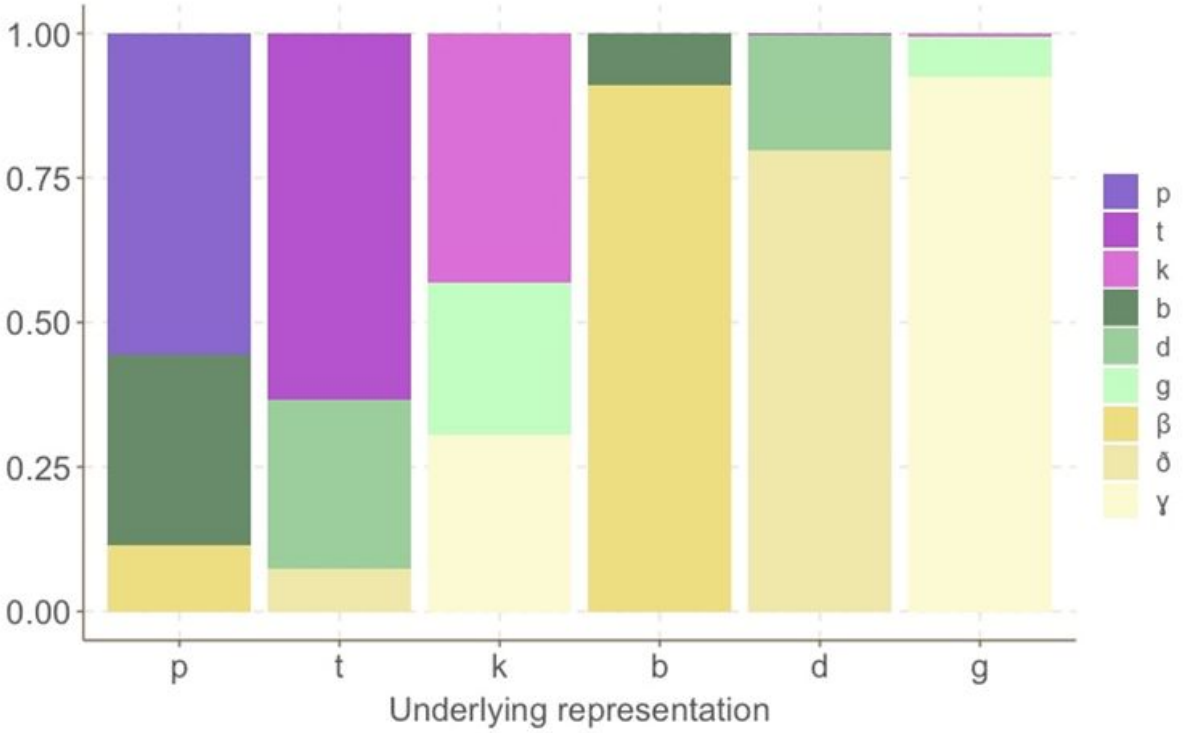
Surface differences



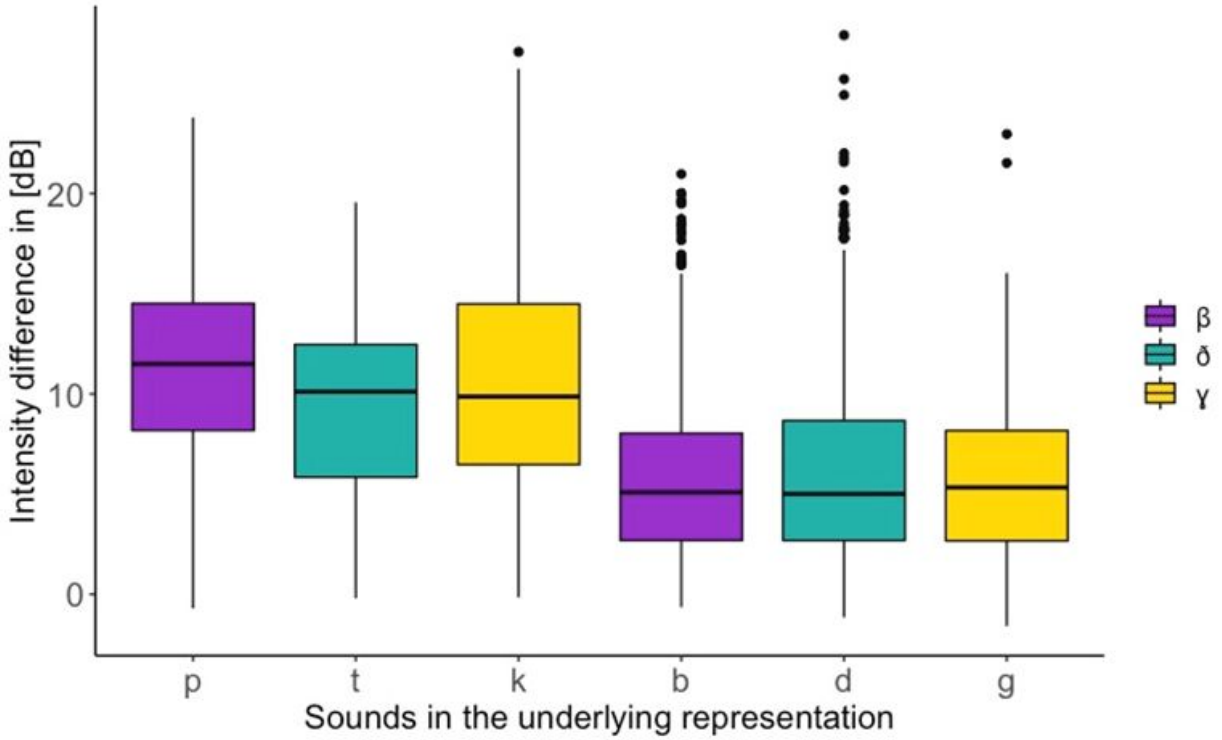
Surface differences



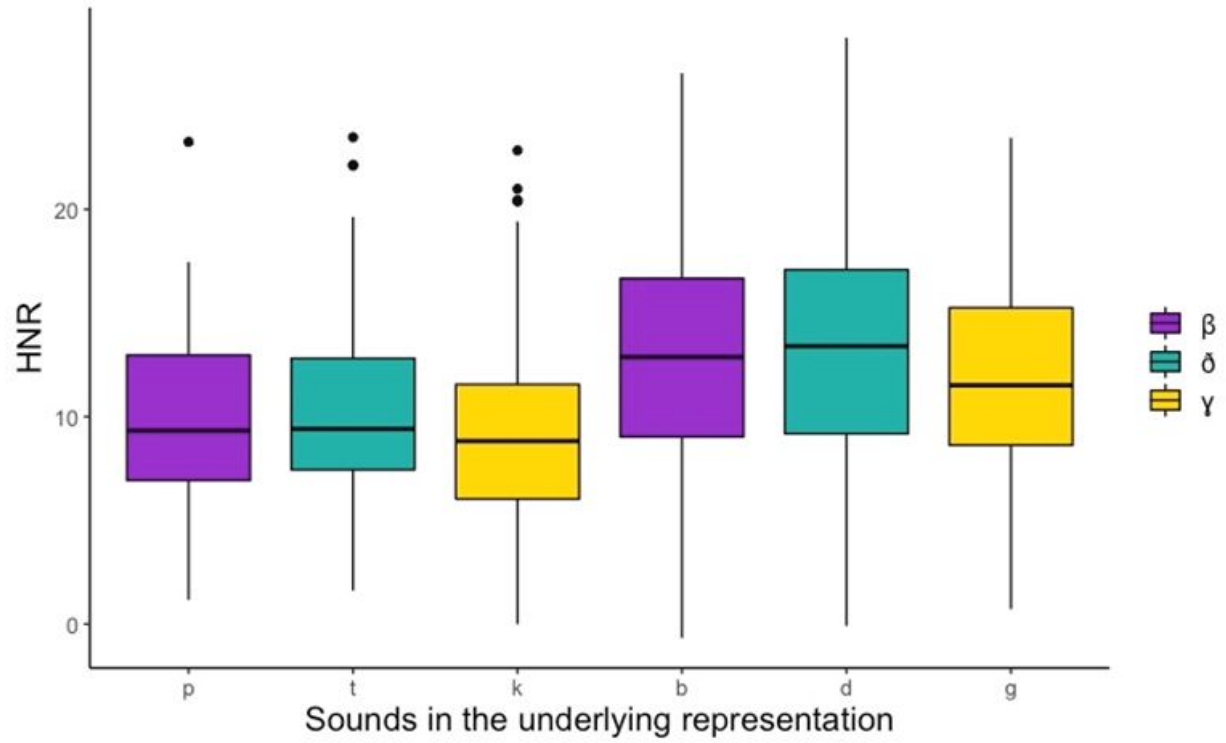
Surface differences



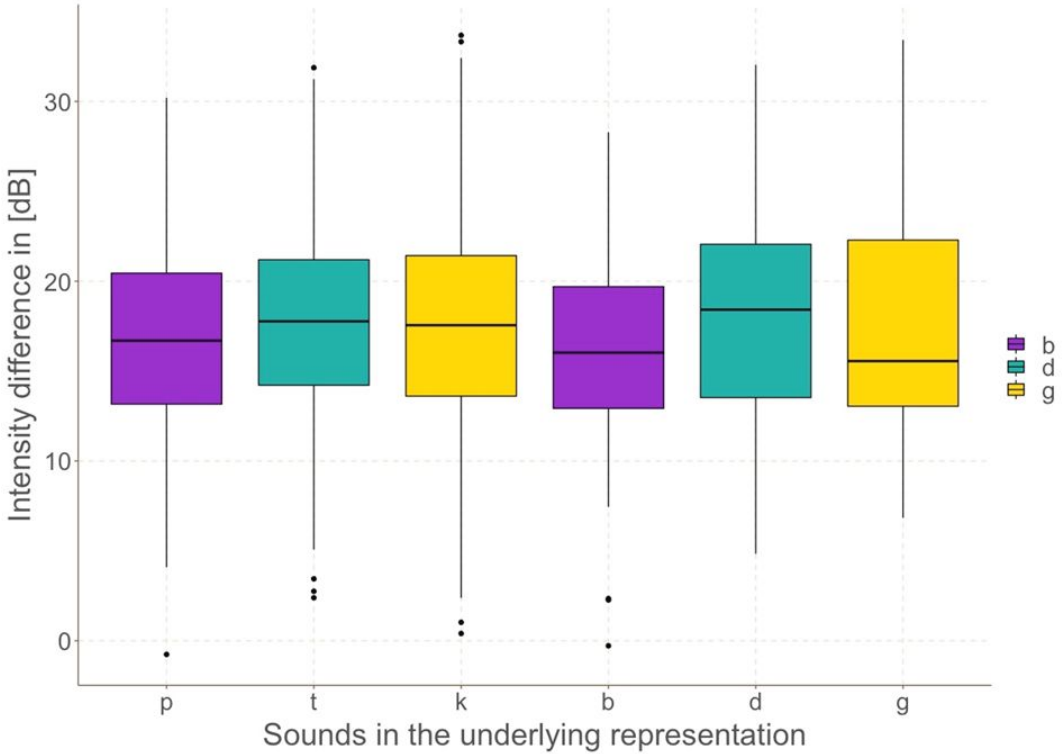
Phonemic status



Phonemic status



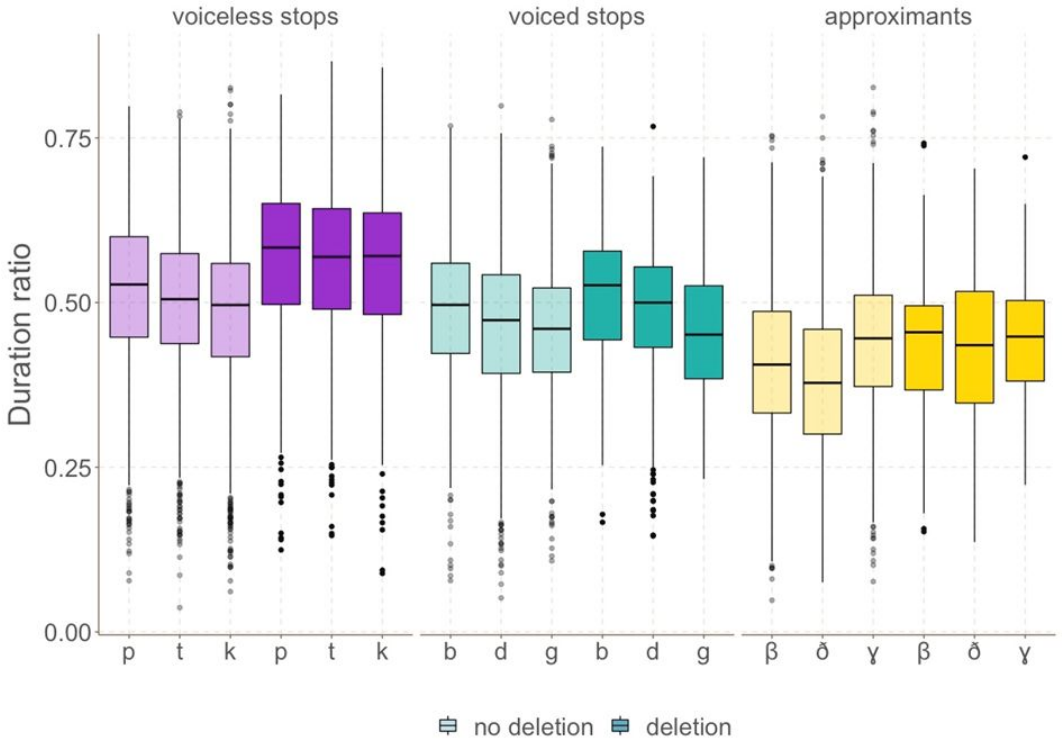
Phonemic status



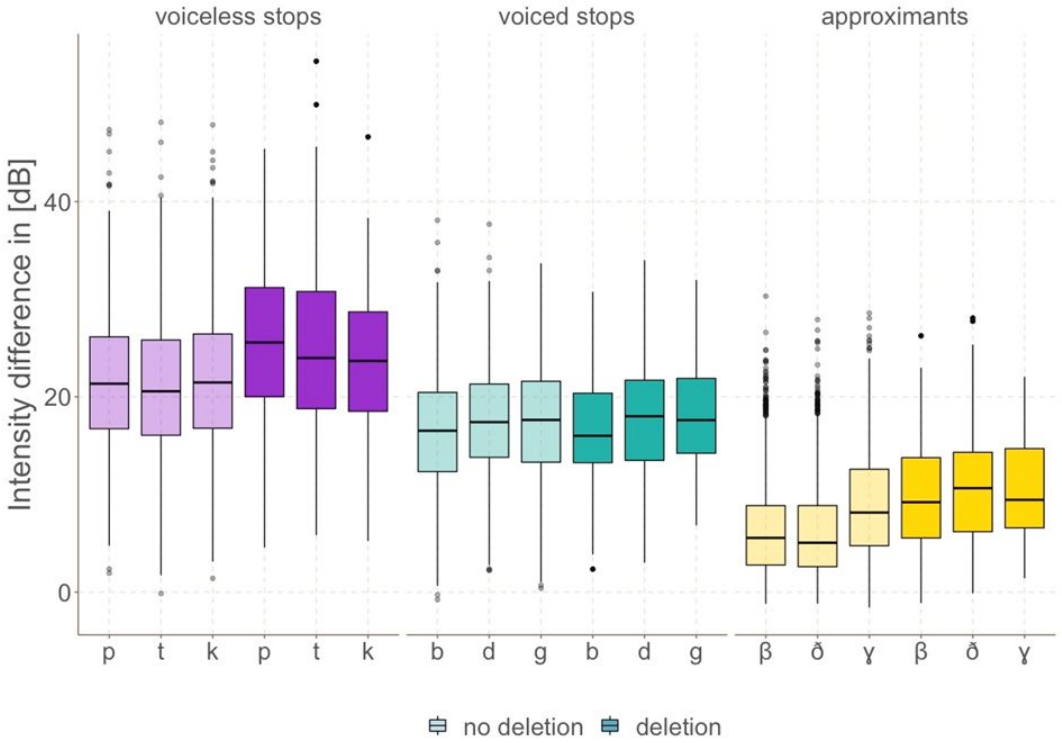
Phonological conditioning

	/p/	/t/	/k/	/b/	/d/	/g/
post-deletion	391	642	410	186	472	46
voiceless stop	88.2%	93.8%	72.0%	0.5%	0.4%	4.3%
voiced stop	7.9%	5.3%	11.7%	62.9%	68.4%	37.0%
approximant	3.8%	0.9%	16.3%	36.6%	31.1%	58.7%
		84.7% vs 15.3%			**56.1% vs 42.1%**	
postvocalic	1769	2225	3177	1902	1854	594
voiceless stop	48.5%	54.7%	39.3%	0.0%	0.3%	0.2%
voiced stop	37.9%	35.9%	28.5%	3.4%	7.5%	4.7%
approximant	13.6%	9.4%	32.3%	96.6%	92.2%	95.1%
		47.5% vs 52.5%			**5.2% vs 94.6%**	

Phonological conditioning



Phonological conditioning

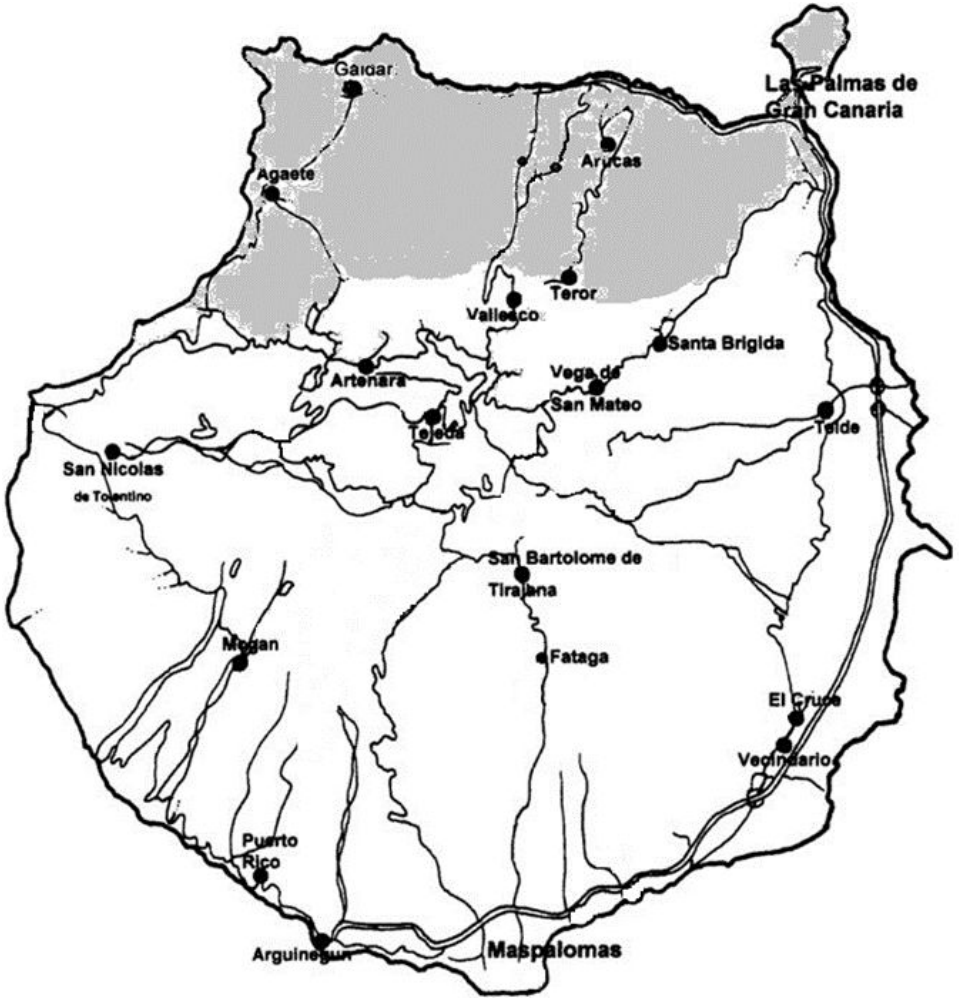


Interim summary 1

- there is a lot of **gradience** and **variability** in the data – probably in any dataset
- some degree of **categoricity** or **allophonic variation** can be identified quantitatively
- **different URs** are produced differently despite partial phonemic overlap
- surface variants depend on **phonological structure**: interaction with deletion

Fieldwork vs lab speech

Spanish spoken on Gran Canaria



Sources:

1) **Fieldwork/corpus: 44 native speakers, 111,317 phones, 16,454 post-vocalic /p t k b d g/**
2) **Experimental data from 20 young speakers, 128 sentences each**

Factor: social setting

lab vs spontaneous speech

Examples of sentences used

He comprado 5 panes de millo	stressed /p/
He comprado 5 pantalones de lana	unstressed /p/
He comprado 5 tarros de garbanzos	stressed /t/
He comprado 5 tenedores de plástico	unstressed /t/
He comprado 5 kilos de tomates	stressed /k/
He comprado 5 calcetines de Adidas	unstressed /k/

Modality 1

aspiration/deletion

/s/ -> [h/H] /_V — prensa[h]idráulicas ‘hydraulic presses’

/s/ -> [h] /_k — chocolate[h]con ‘chocolates with’

/s/ -> [∅] /_d — pane[∅]de ‘breads from’

stop lenition

/b d g/ -> [b d g] /V(C)_ — pane(s)[d]e ‘breads from’

/b d g/ -> [B D G] /V_ — cinco[D]ulces ‘five sweets’

/p t k/ -> [b d g] /V_ — cinco[b]anes ‘five breads’

Modality 2

aspiration/deletion

/s/ -> [/s/ -> [h/H]/_V

prensa[H]idráulicas ‘hydraulic press’

/s/ -> [∅] /_C

chocolate[∅]con ‘chocolates with’

stop lenition

/b d g/ -> [B D G] /V(C)_

pane(s)[D]e ‘breads from’

/b d g/ -> [B D G] /V_

cinco[D]ulces ‘five sweets’

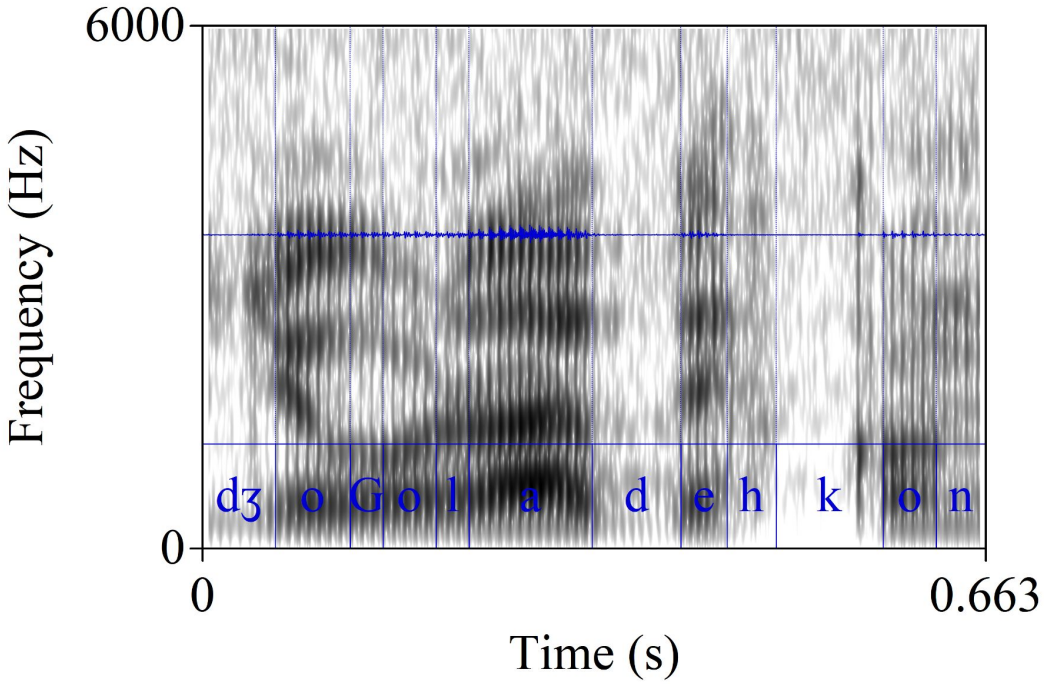
/p t k/ -> [b d g] /V_

cinco[b]anes ‘five breads’

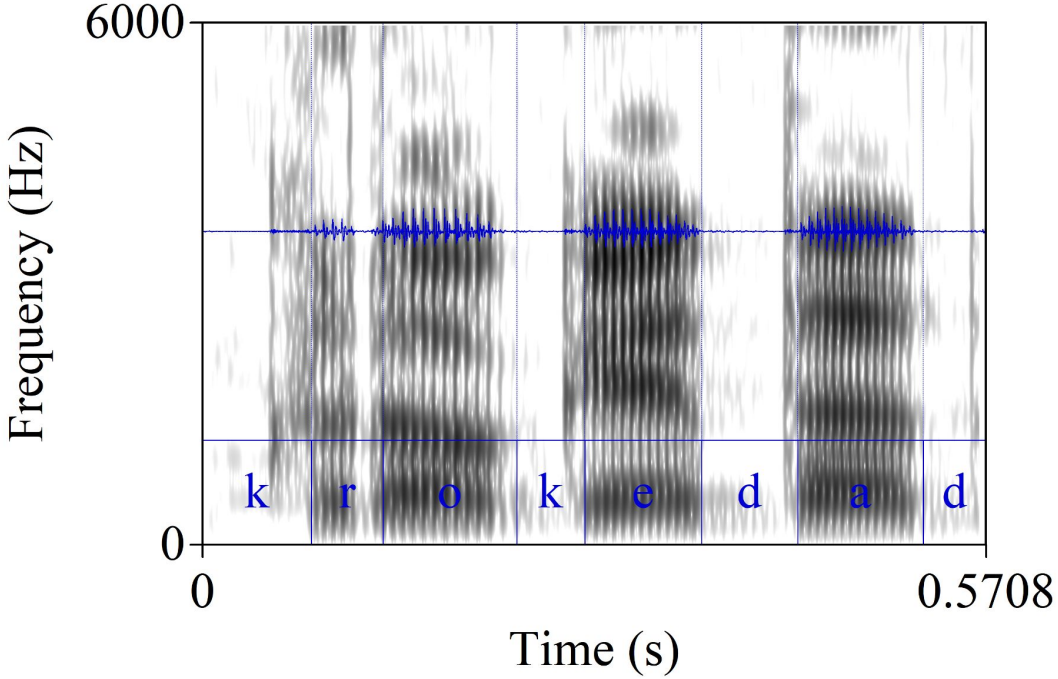
/p t k/ -> [p t k] /V(C)_

chocolate(s)[k]on ‘chocolates with’

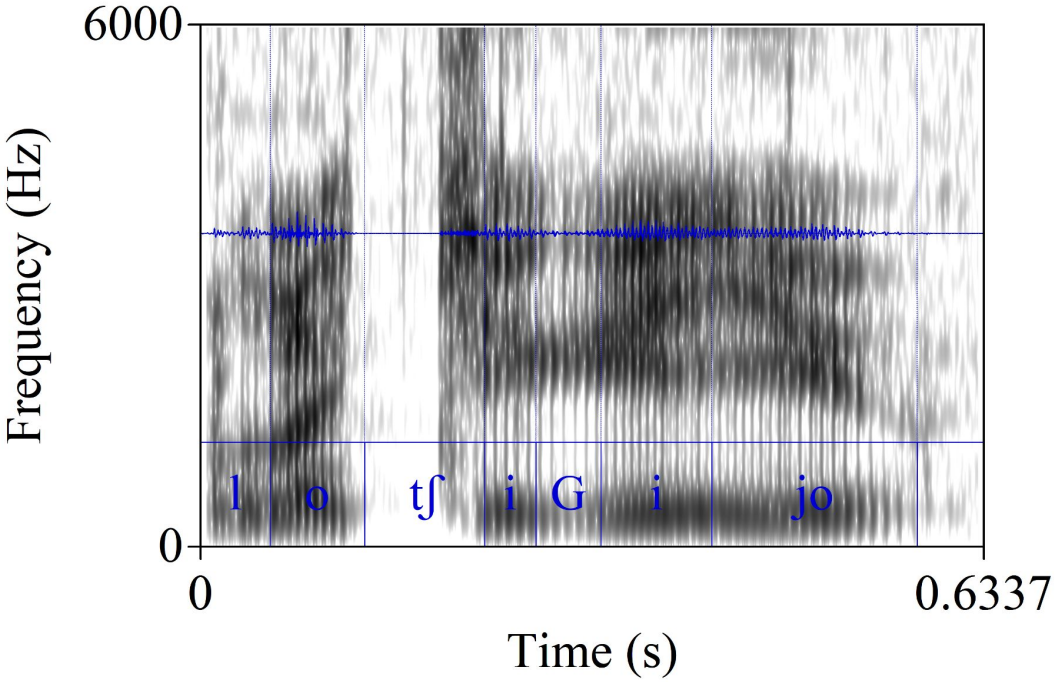
Controlled speech: *chocolates con*



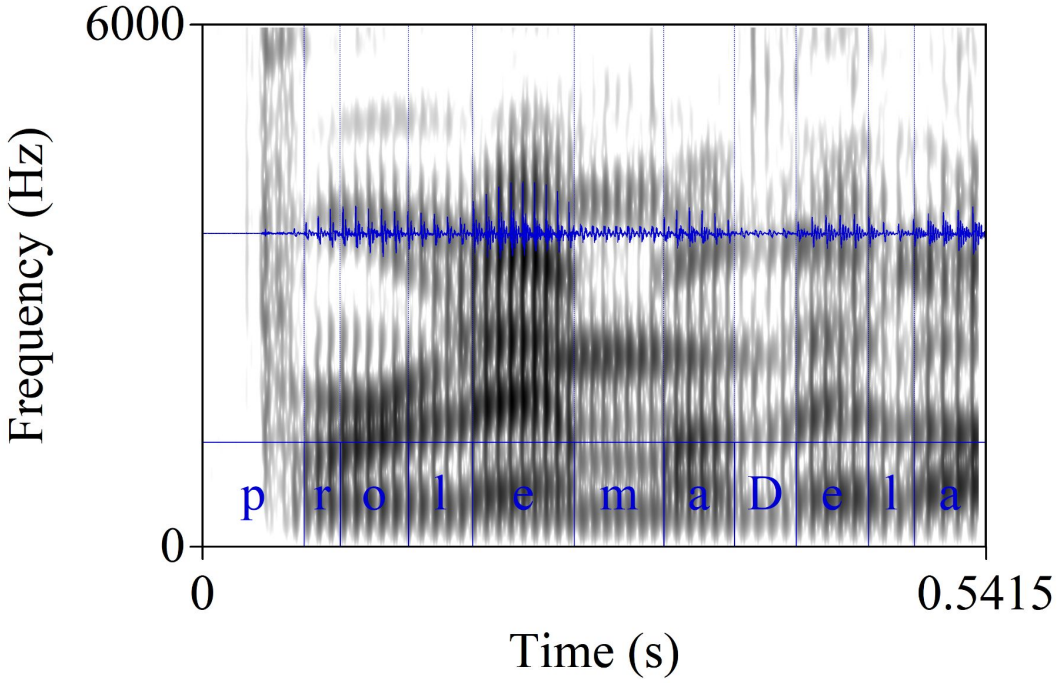
Controlled speech: *croquetas de*



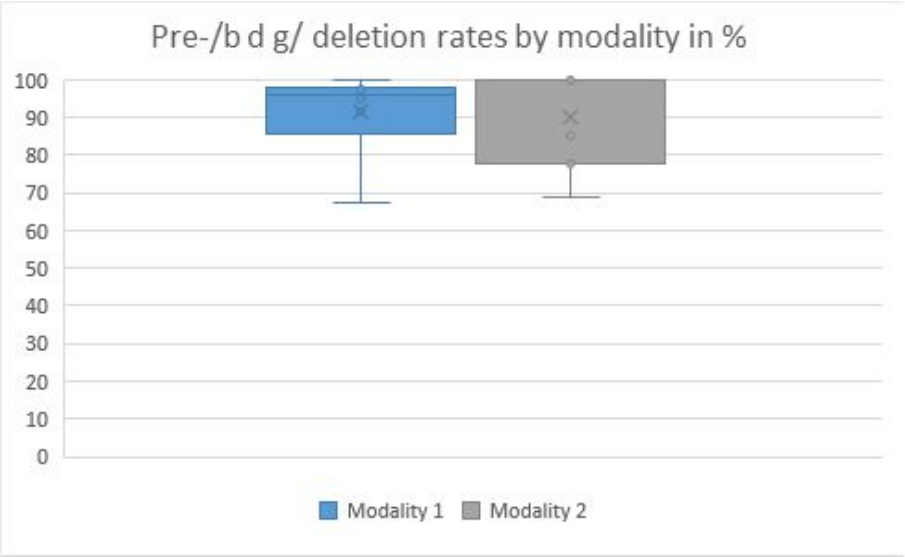
Spontaneous speech: *los chiquillos*

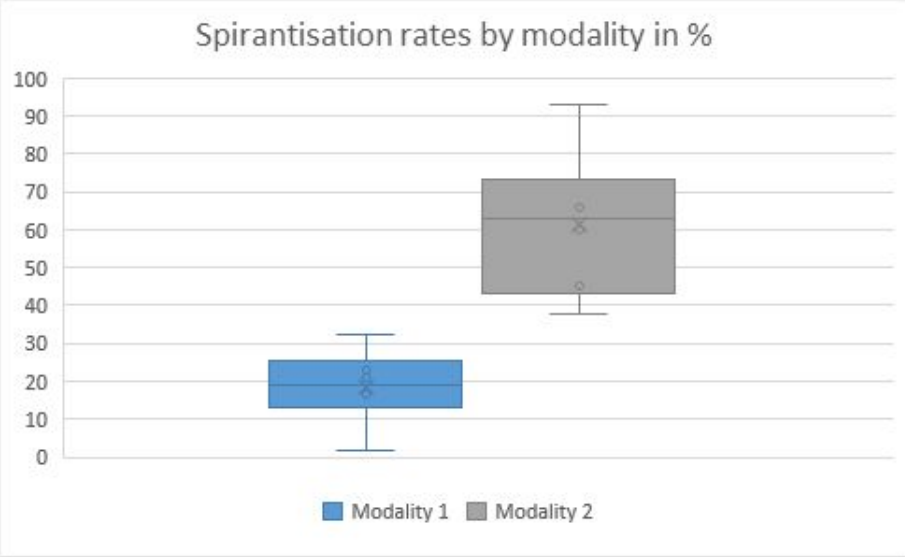


Spontaneous speech: *problemas de la*

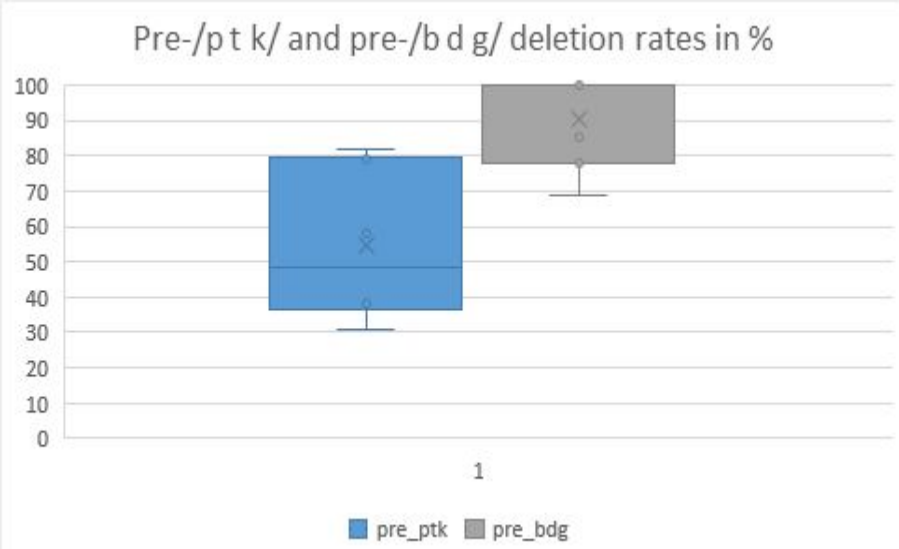


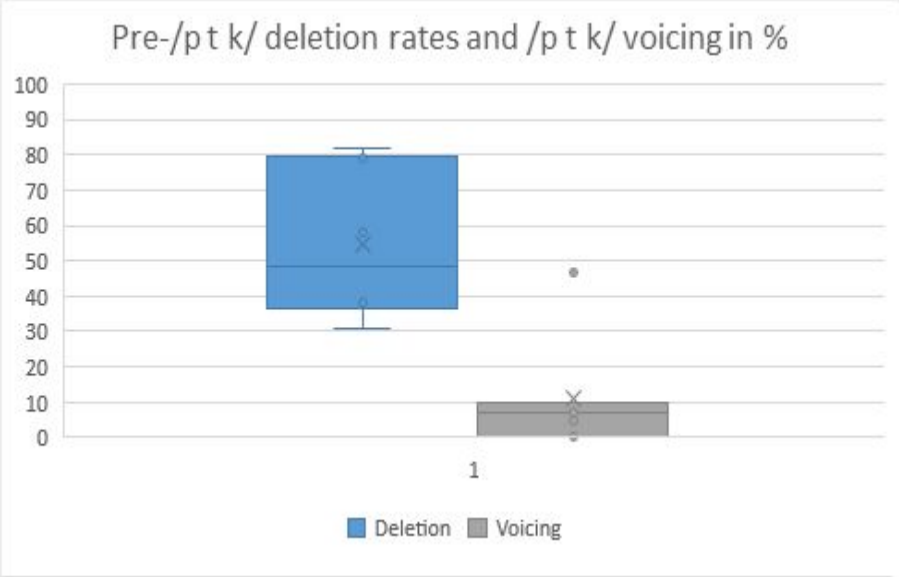
Modality 1 vs Modality 2: /b d g/





Modality 2: /p t k/





Interim summary 2

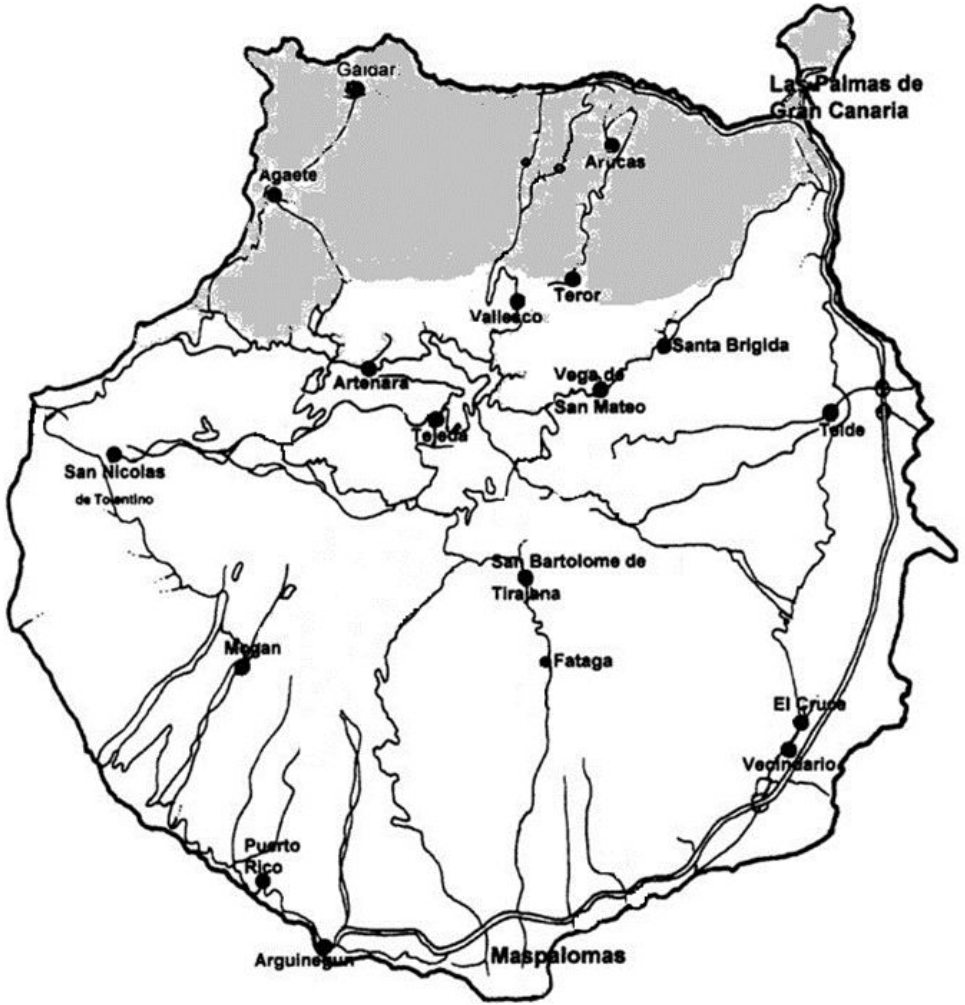
- Individual speaker choices can be **systematic** across different social settings: **different weakening stages**
- Intra-speaker variation can be a reflection of **sound change in progress**
- Variation is situational: **co-phonologies**
- Variation should be modelled by **incorporating external factors into the grammar**
- selective blocking effects (incomplete deletion?)

Using social media in phonetic/phonological analysis

Spanish spoken on Gran Canaria

Sources:

1) **Experimental data from 5 young speakers, 128 sentences each**
2) **Spontaneous recordings from WhattsApp, 5 young speakers**



Factor: social setting

lab recordings vs social media

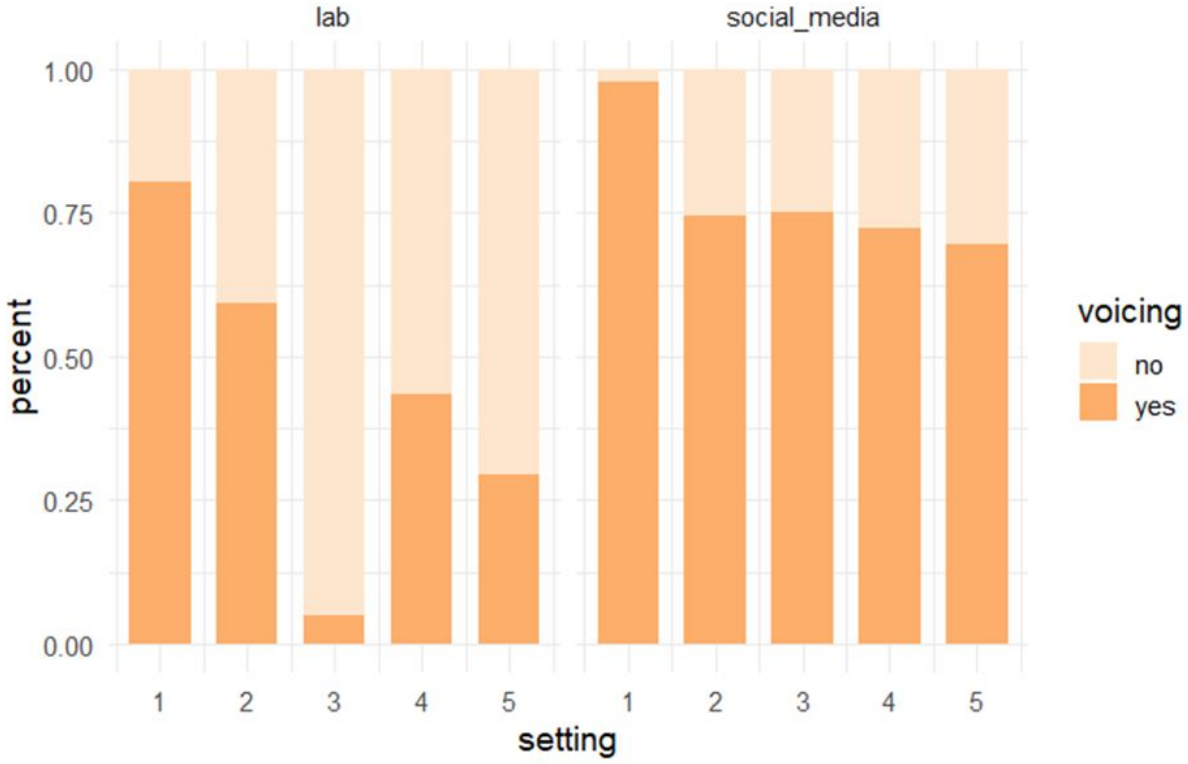
Lab data vs WhattsApp recordings

Speaker	Total recordings	Good quality recordings	Total time (s)	Sounds (social media)	Sounds (lab)[1]
Speaker 1	6	4, 5, 6	73.09	42	77
Speaker 2	2	1, 2	140.27	47	76
Speaker 3	5	1, 3, 4, 5	84.58	40	78
Speaker 4	9	1, 3, 4, 6, 9	172.59	105	78
Speaker 5	3	1, 2	84.08	59	68
Total	**25**	**16**	**554.61**	**293**	**377**

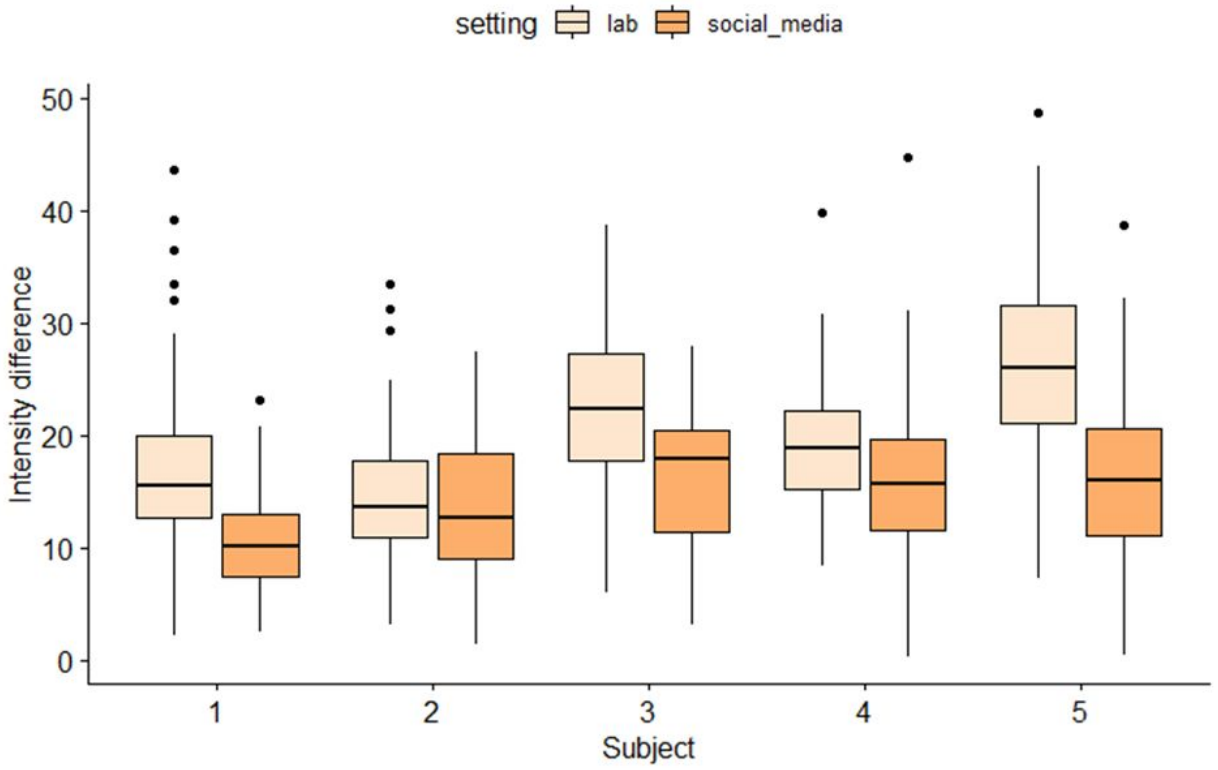
The data

- 670 observations from **5 speakers**
- target: post-vocalic **/p t k/ voicing**
- **43.8% vs 76%** sounds classified as **voiced**
- substantial **interindividual differences** between speakers **in the lab** setting but all speakers seem to be quite **uniform** in the percentage of voicing **in a naturalistic setting**

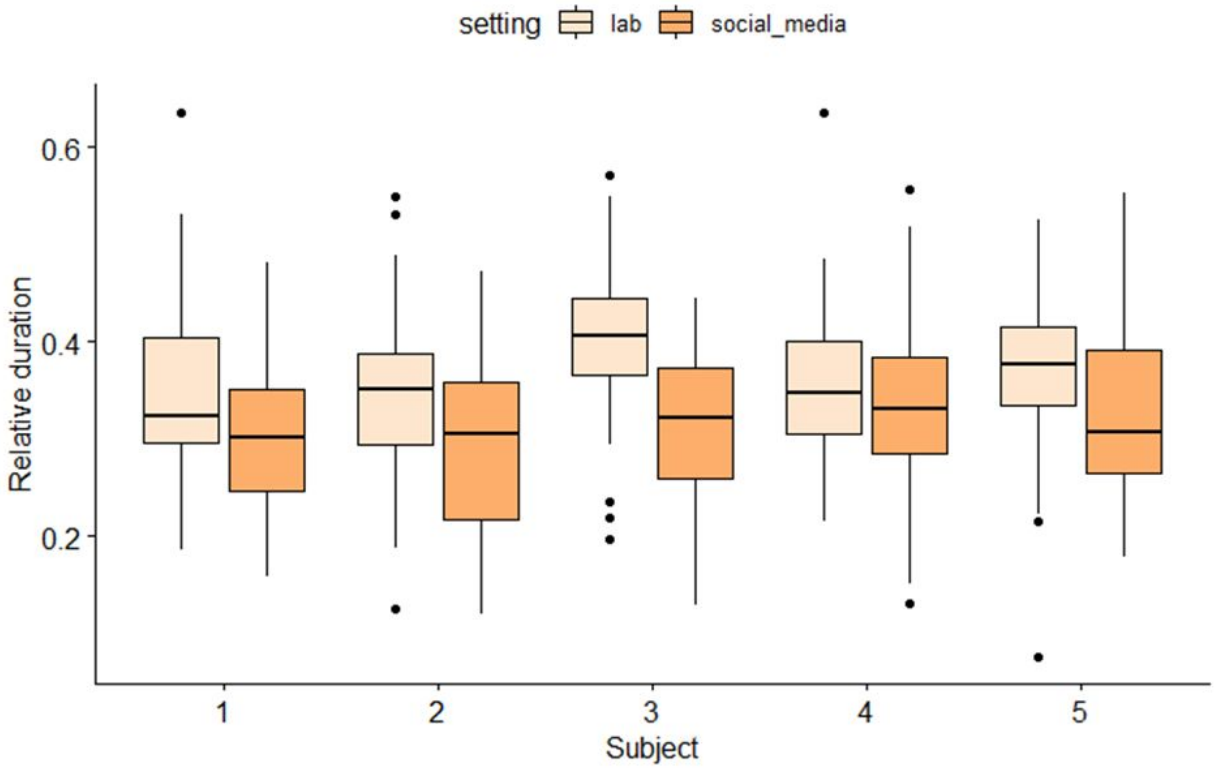
Voicing: lab setting vs the social media



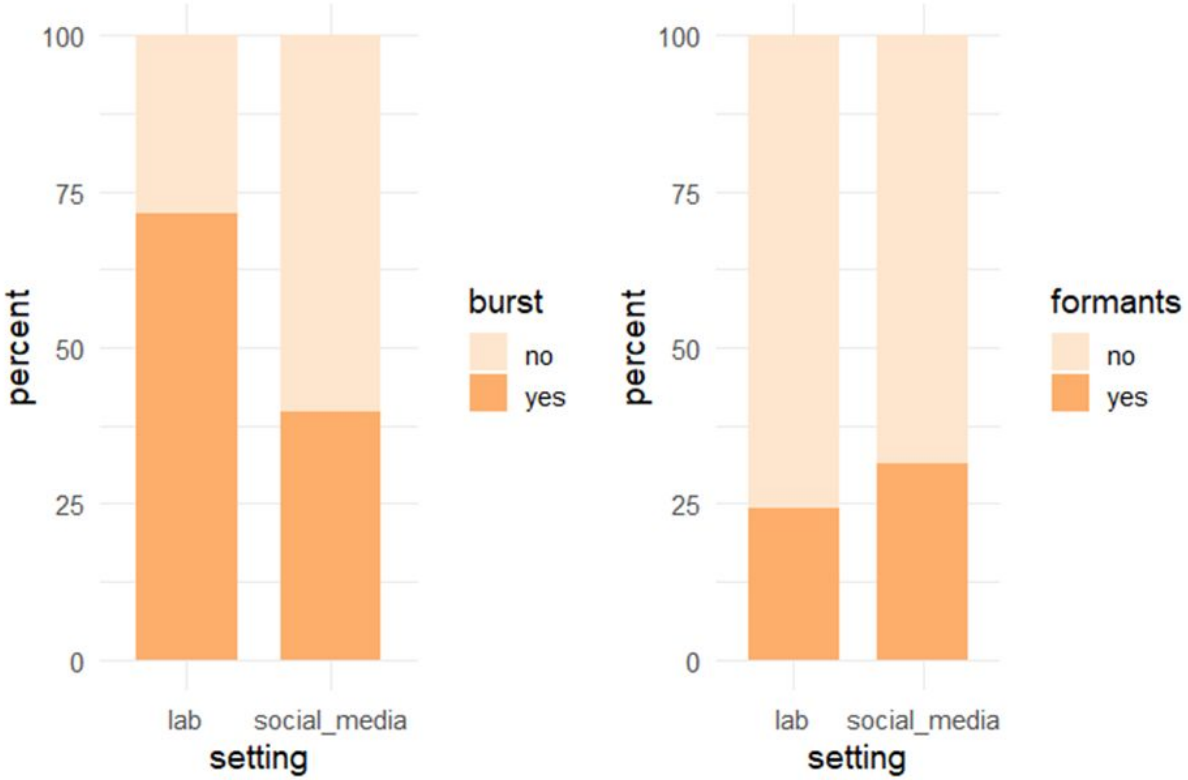
Intensity: lab setting vs the social media



Duration: lab setting vs the social media



Burst and formants



Interim summary 3

- social setting affects the **naturalness** of speech in a particular way, i.e. both **inter- and intra-speaker variation**
- speakers in the same age range speak in a similar fashion, with similar rates of lenition
- **speaker strategies** pertaining to supervised speech differ
- how we access the data affects our **generalisations**

Motion capture study on /p b/

Factors tested:
prosodic and phonological effects

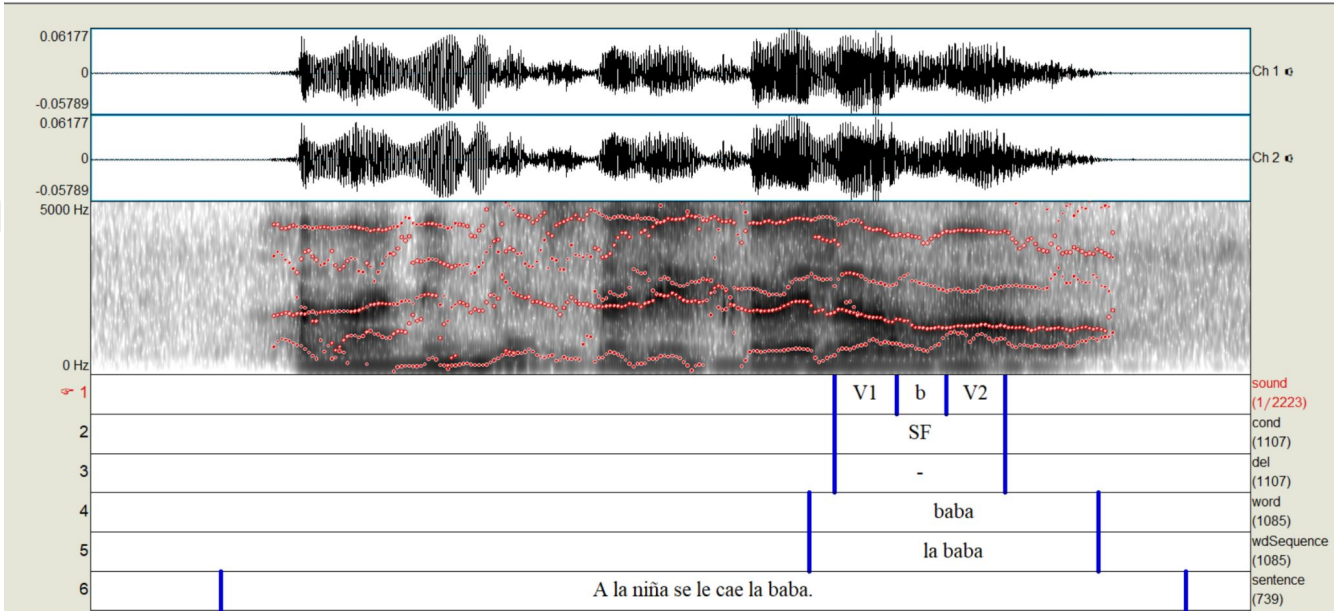
- post-vocalic /p b/ tested for lip aperture and lip area measurements
- to be correlated with acoustic markers of lenition
- 376 sentences, a total of 560 target words
- Conditions:
  - stressed syllable (S)
  - unstressed syllable (US)
  - stressed syllable in focus (SF)
  - deletion context (del)

Examples of sentences used

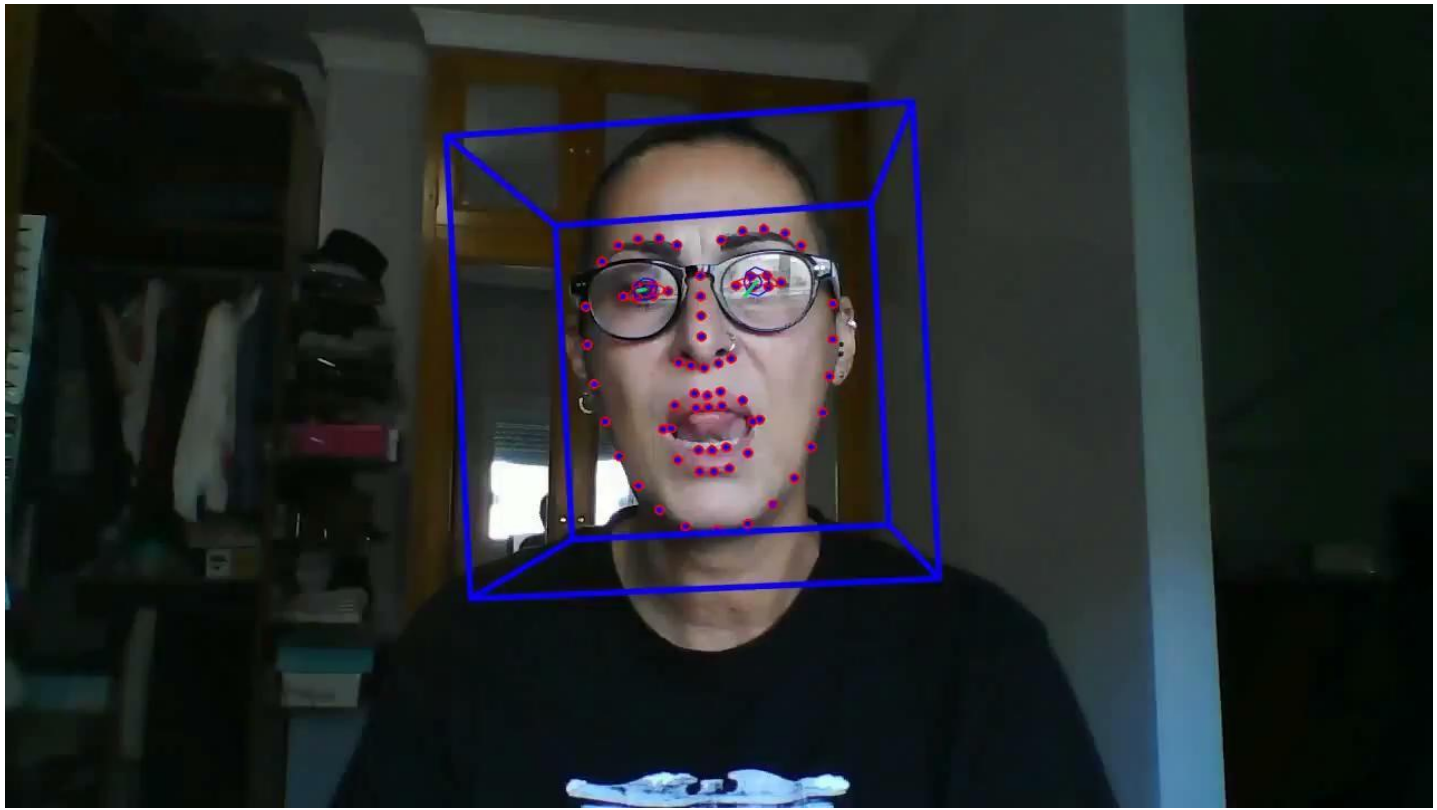
La barrera estaba mal colocada y el portero no veía.	US /b/
La paciencia de esa mujer me tenía impresionado.	US /p/
La banda de música empezó el concierto con **la bamba**.	S /b/, SF /b/
La paga mensual es **más baja** de lo que **pensaba Paco.**	S /p/, DEL, SF /p/
La vaca de Juan cuesta **mucha pasta**.	S /b/, SF /p/
Las Vacas Locas es una banda de música de Tenerife.	DEL /b/
Soy de Gáldar, pero vivo en **Las Palmas**.	DEL /p/

Data extraction and video output analysis

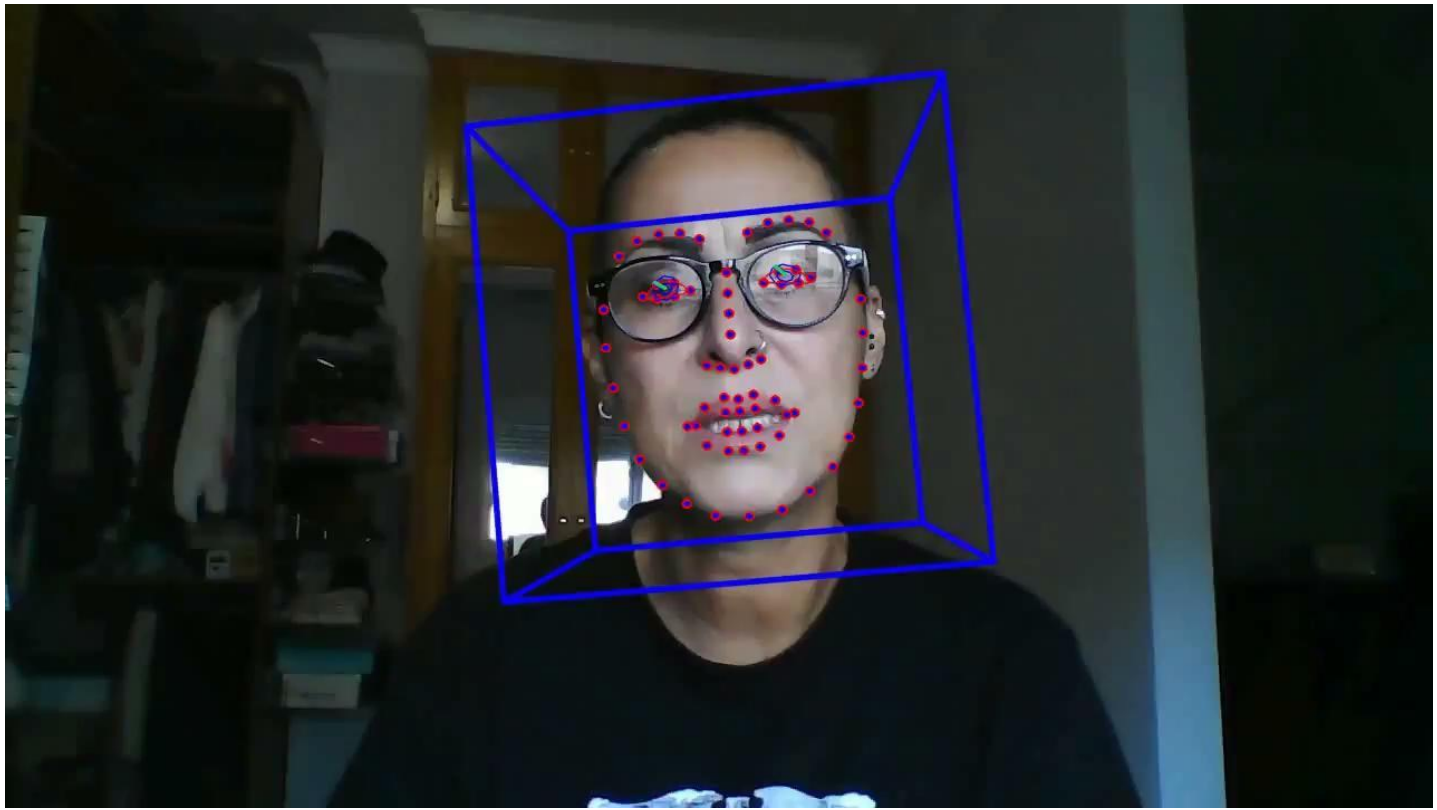
Temporal marks for the target words and their critical VCV segment sequences were annotated to Praat TextGrids



Example: /aba/

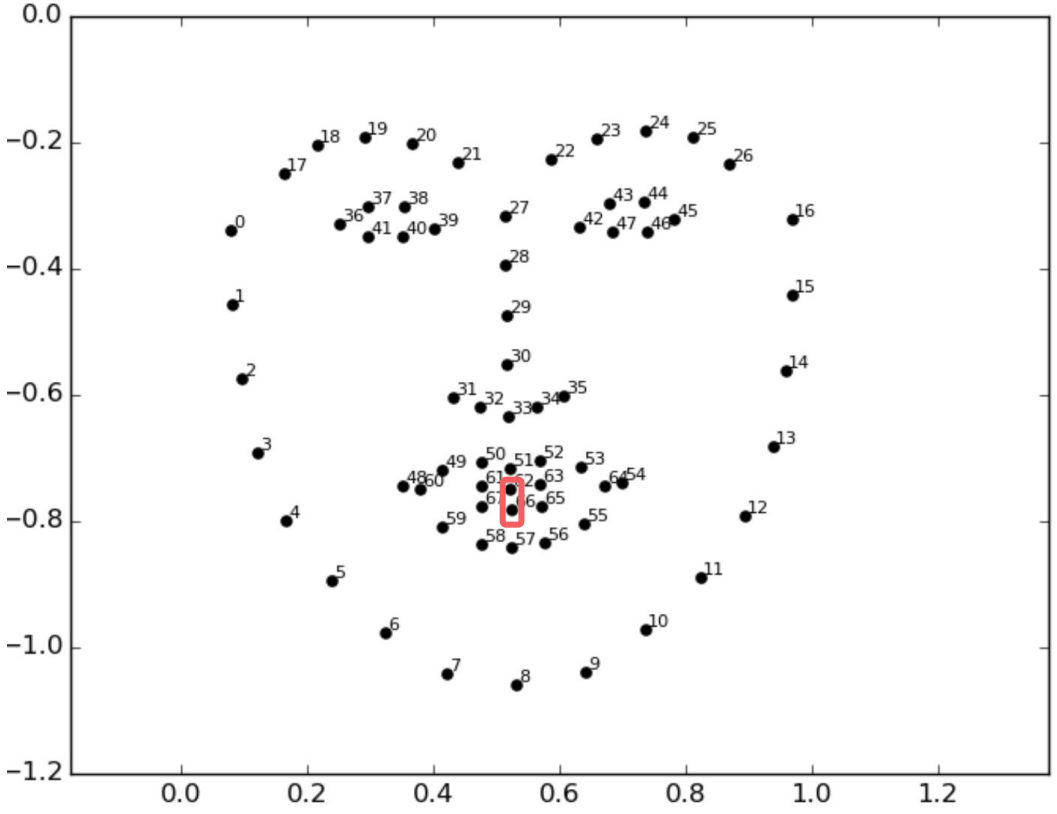


Example: /apa/



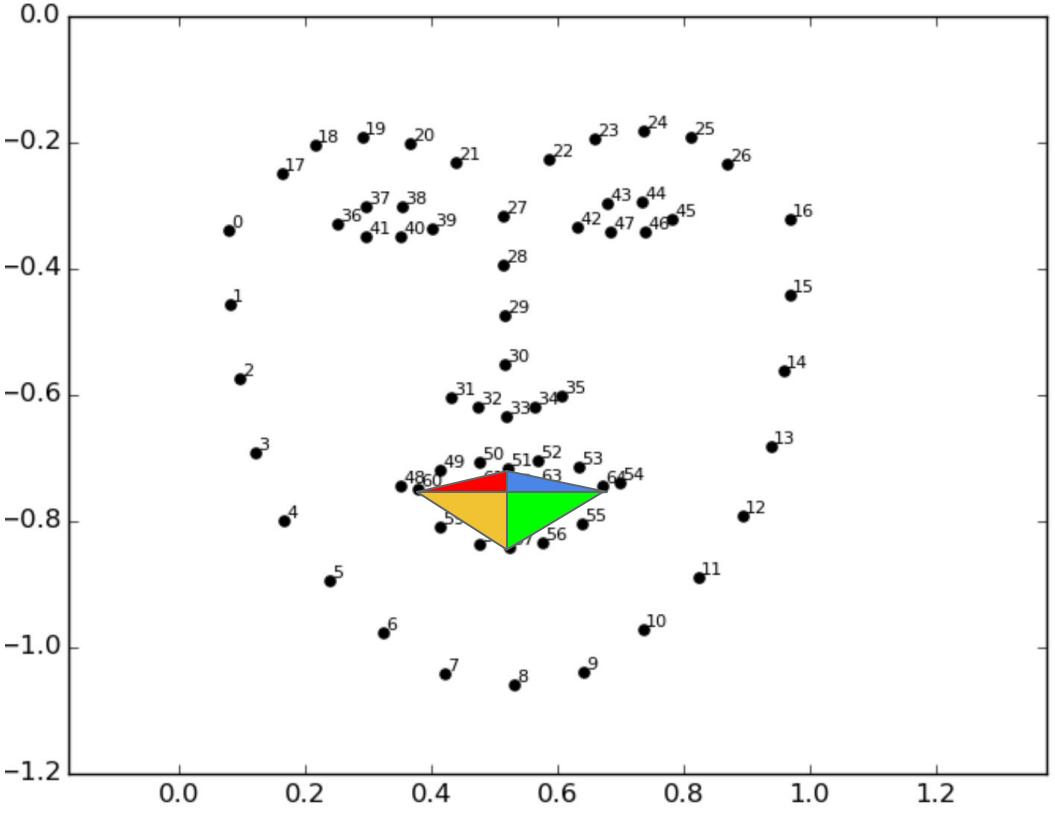
Data extraction and video output analysis

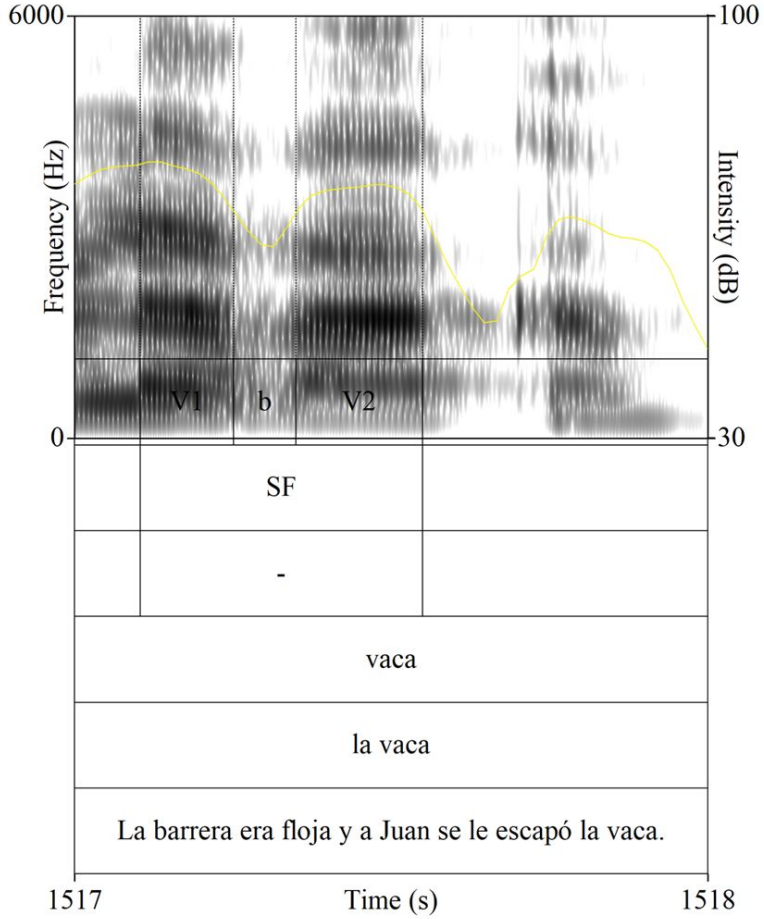
- For each frame of each trial, a custom Python script determined…
  - **Vertical Lip Aperture -** euclidean distance here

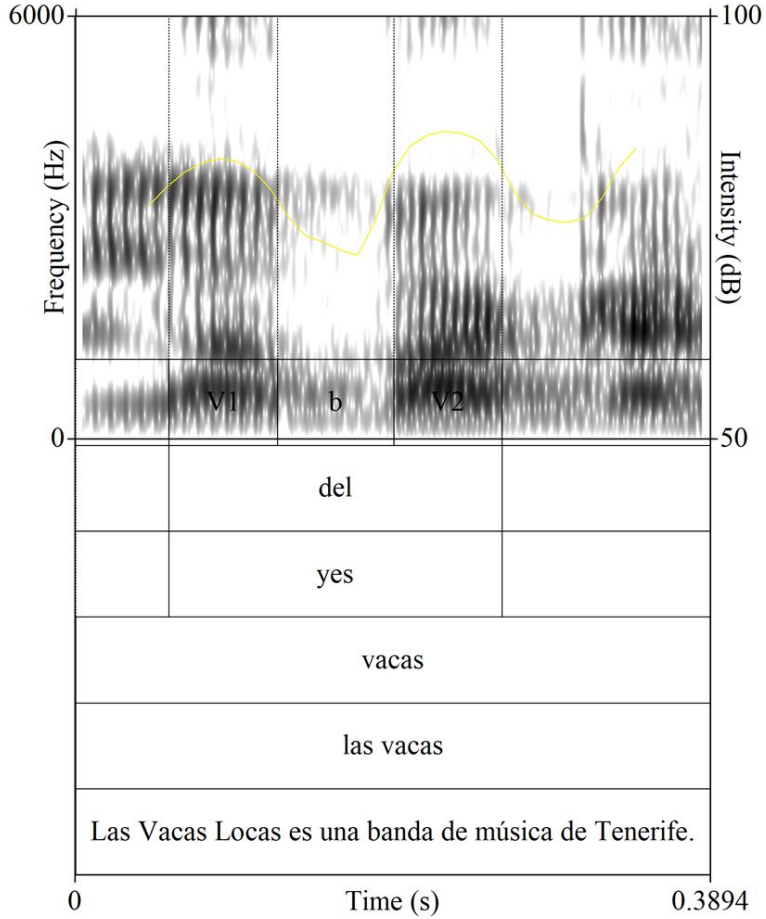


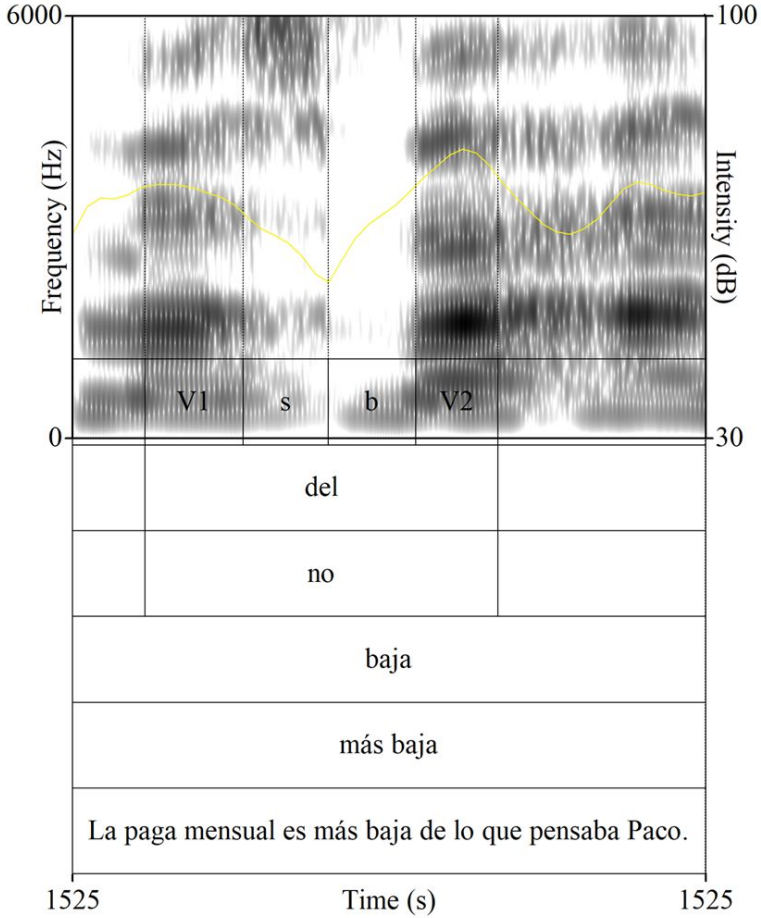
Data extraction and video output analysis

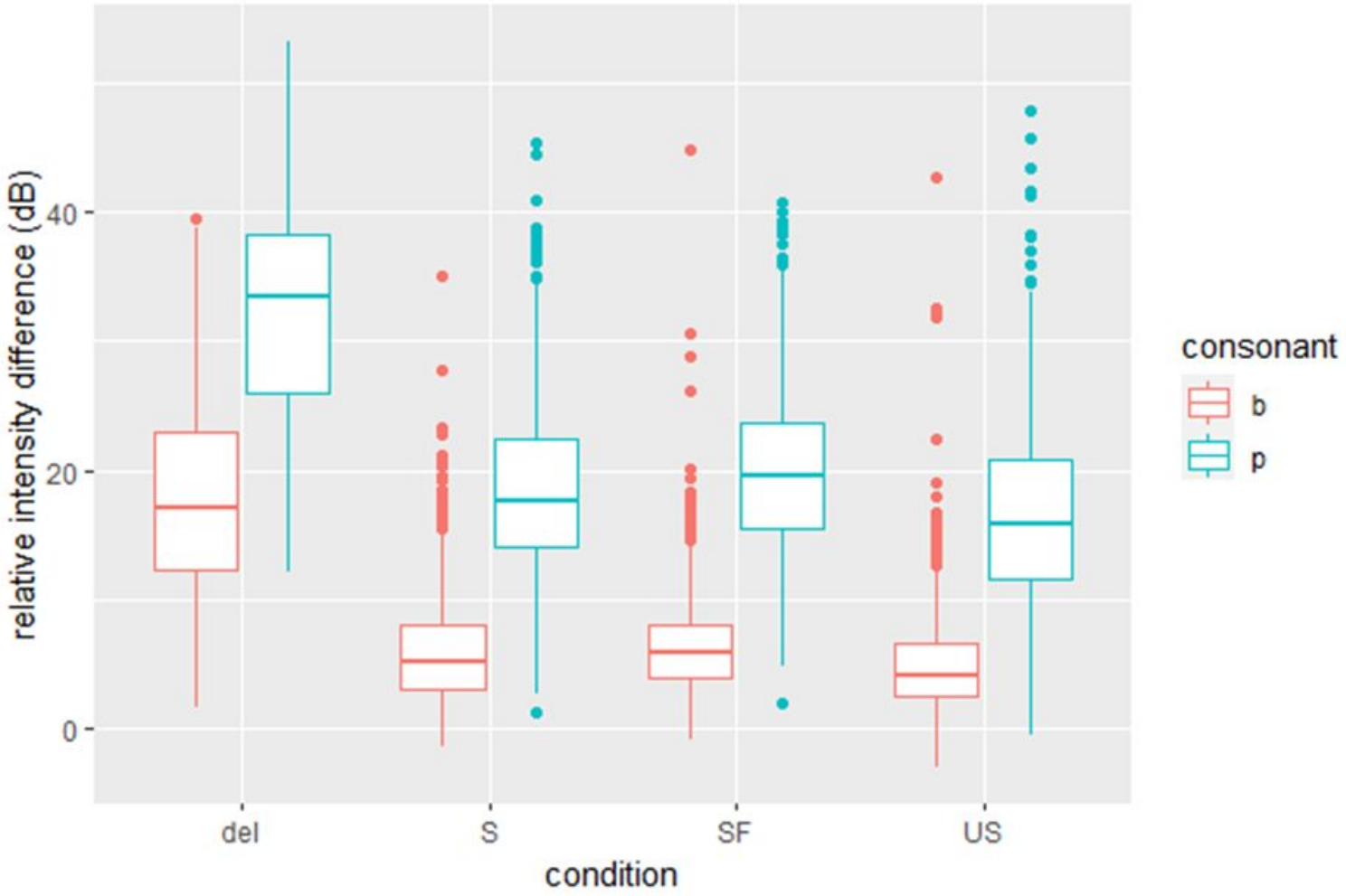
- For each frame of each trial, a custom Python script determined…
  - **Vertical Lip Aperture -** euclidean distance here
  - **Lip Area -** areas of these triangles (plus central rectangle, which here has area 0)

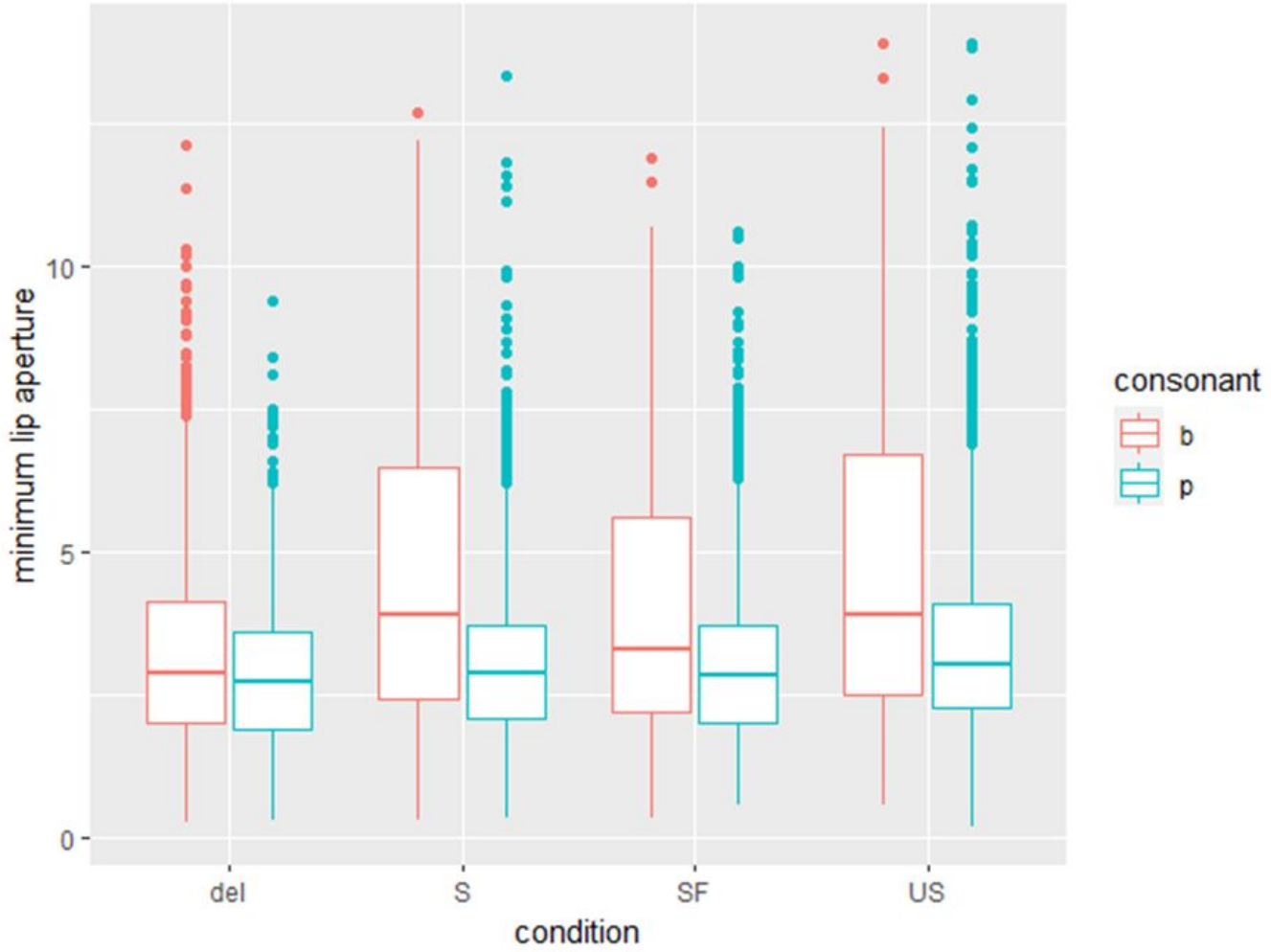


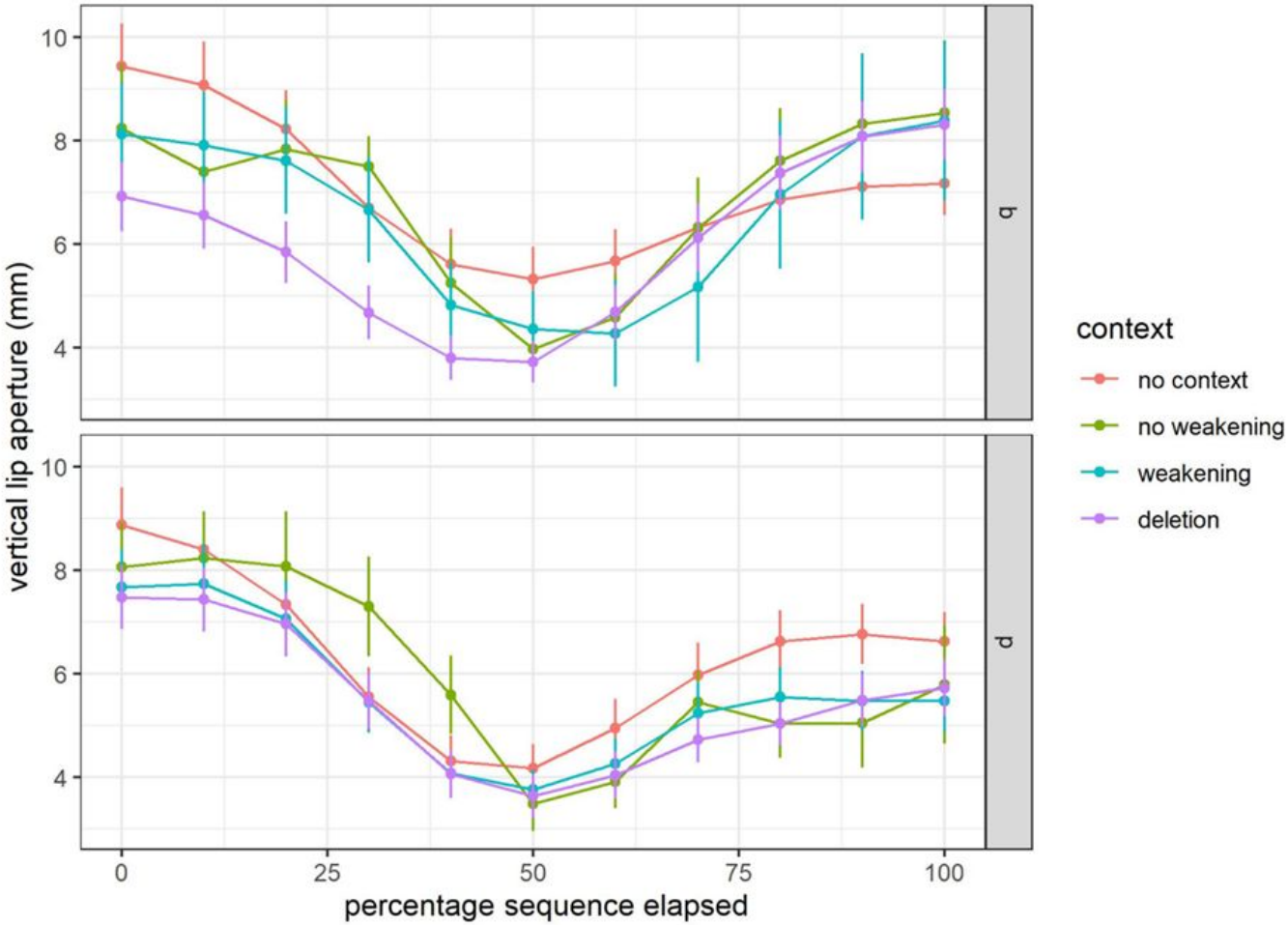

6000
100
Frequency (Hz)
Intensity (dB)
V1
b
V2
0
30
SF
-
vaca
la vaca
La barrera era floja y a Juan se le escapó la vaca.
1517
Time (s)
1518

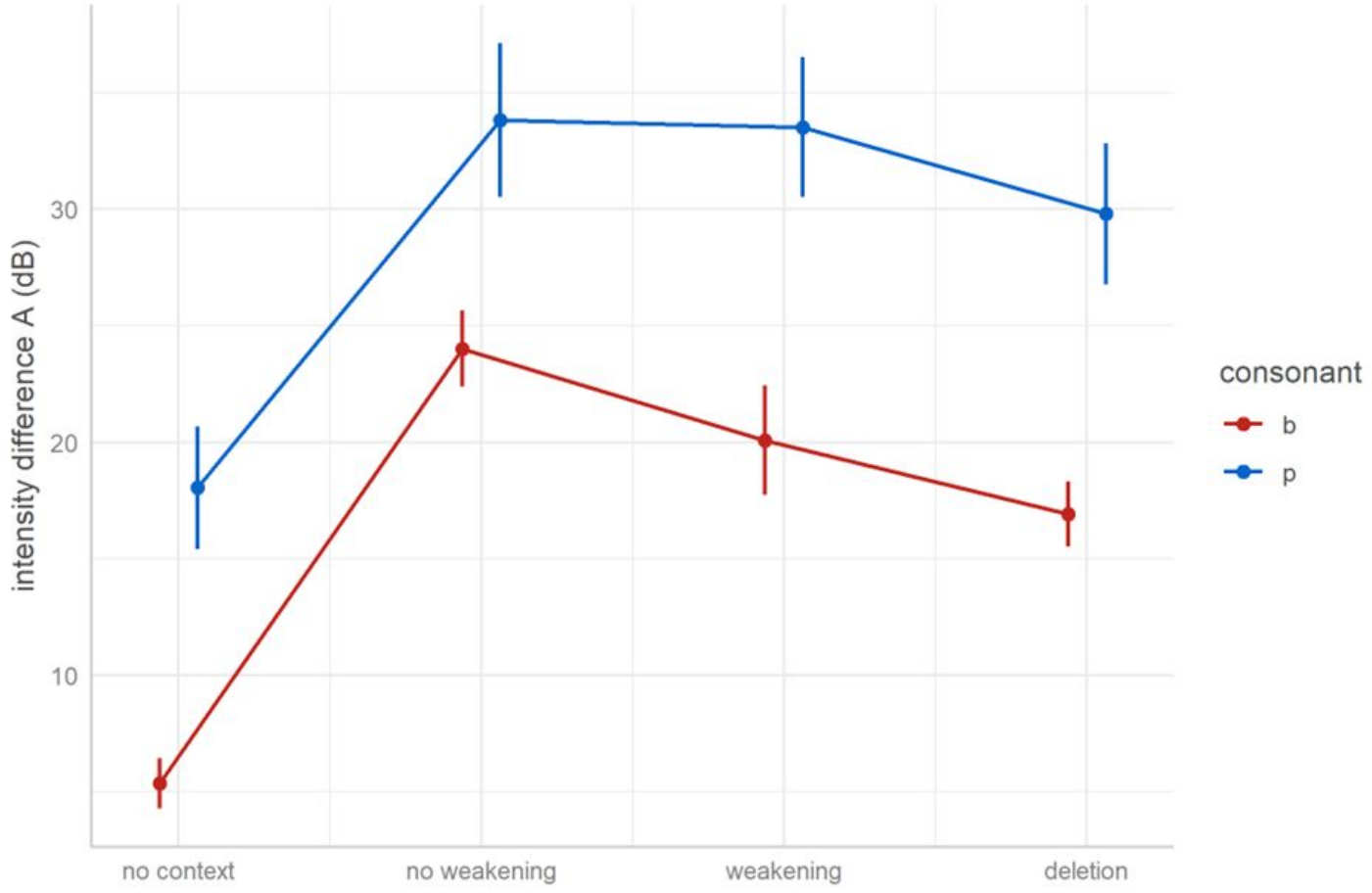

6000
100
Frequency (Hz)
Intensity (dB)
V1
b
V2
0
50
del
yes
vacas
las vacas
Las Vacas Locas es una banda de música de Tenerife.
0
Time (s)
0.3894

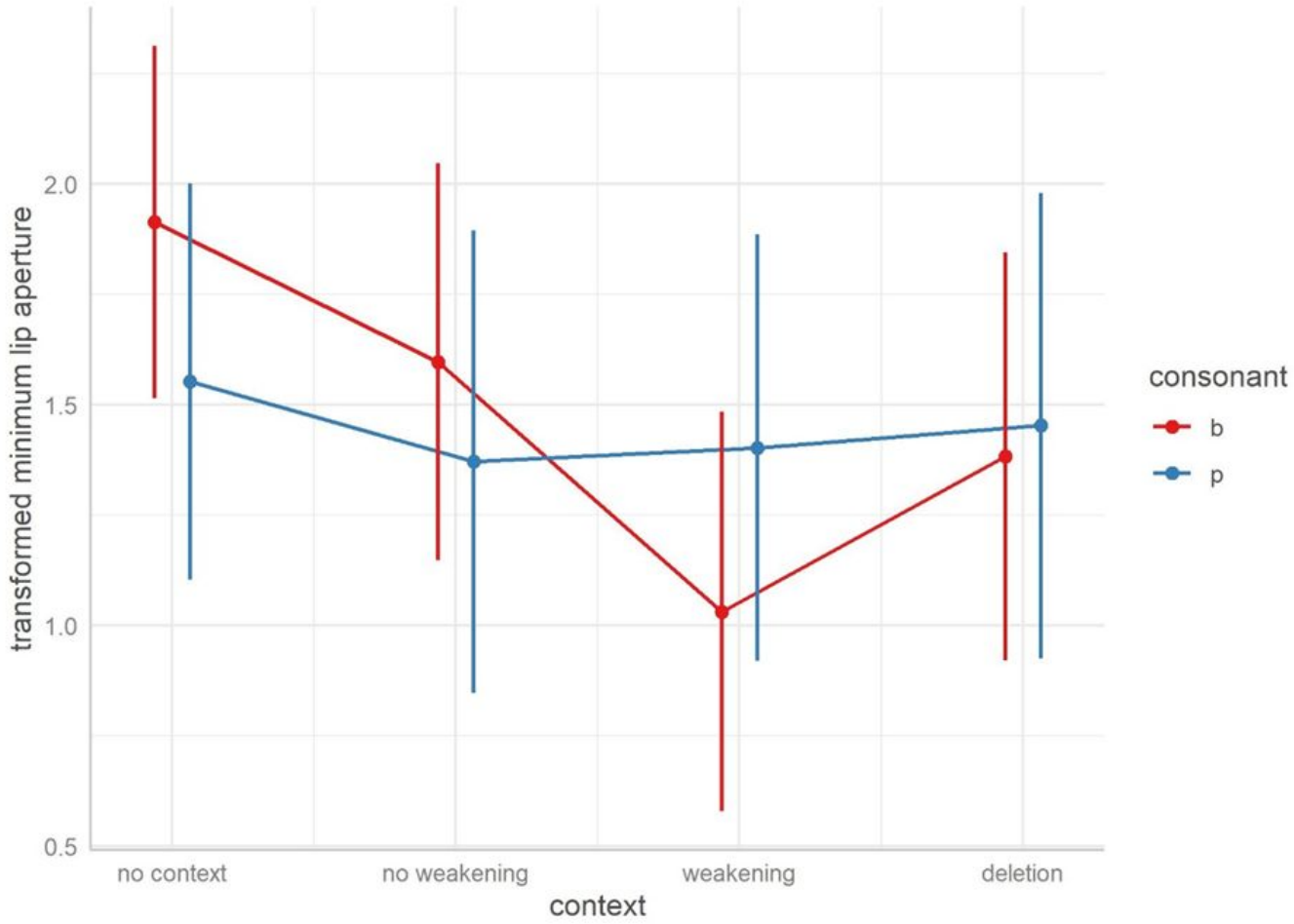

6000
100
Frequency (Hz)
Intensity (dB)
V1
s
b
V2
0
30
del
no
baja
más baja
La paga mensual es más baja de lo que pensaba Paco.
1525
Time (s)
1525


relative intensity difference (dB)
40
20
0
del
S
SF
US
condition
consonant
b
p


minimum lip aperture
10
5
0
del
S
SF
US
condition
consonant
b
p


vertical lip aperture (mm)
10
8
6
4
0
25
50
75
100
percentage sequence elapsed
q
d
context
no context
no weakening
weakening
deletion


intensity difference A (dB)
30
20
10
no context
no weakening
weakening
deletion
consonant
b
p


transformed minimum lip aperture
2.0
1.5
1.0
0.5
no context
no weakening
weakening
deletion
context
consonant
b
p

Interim summary 4

- ❑ **an intermediate category** in deletion contexts?
- ❑ possibly, **incomplete deletion** or gestural masking
- ❑ independent evidence for lenition, and **opacity**
- ❑ **how to disentangle phonology from variation?**

Remote data collection and comparative analysis of /p b/ productions

Factors studied:

self-recordings vs social media vs lab speech

Data samples

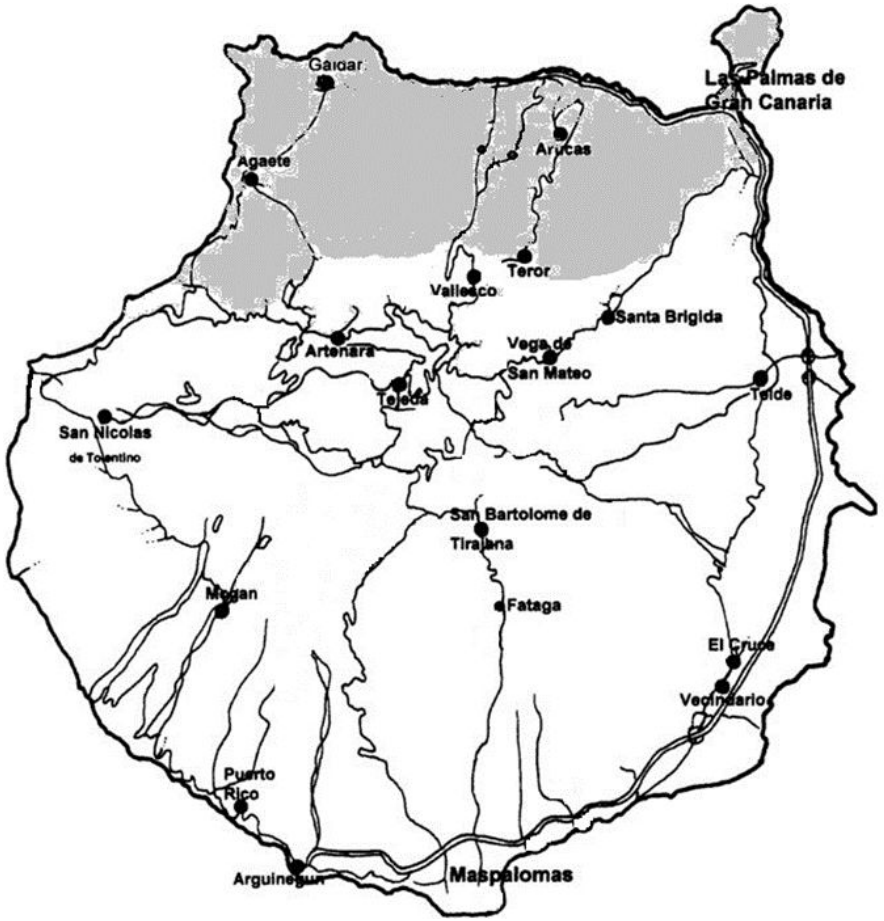
4 data samples:

a) 8 field experiment recordings of read sentences.

b) 8 self-recordings of the same sentences in a repeating condition.

c) 8 self-recordings of spontaneous speech (monologue).

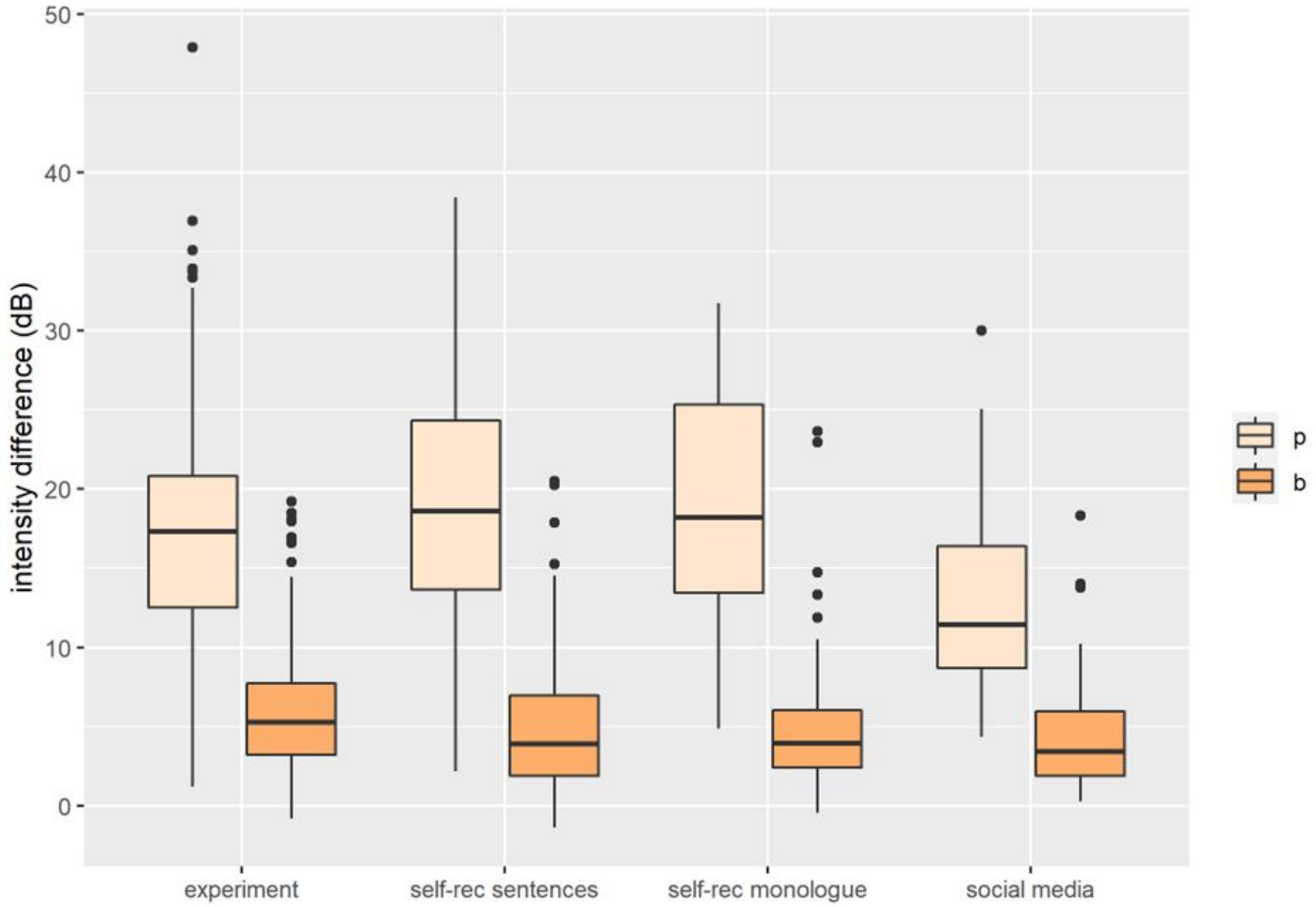
d) 4 social media recordings made via WhatsApp.

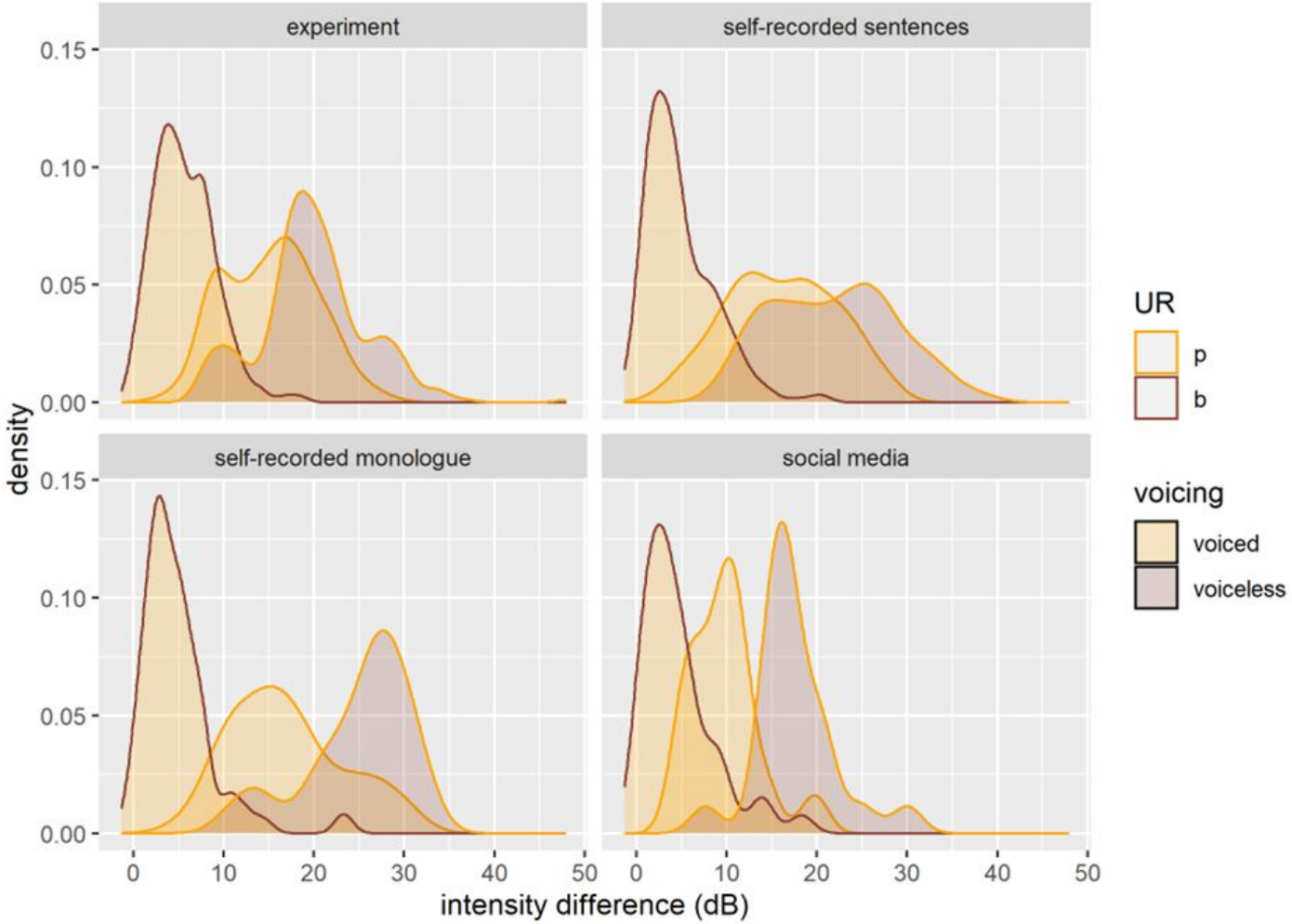


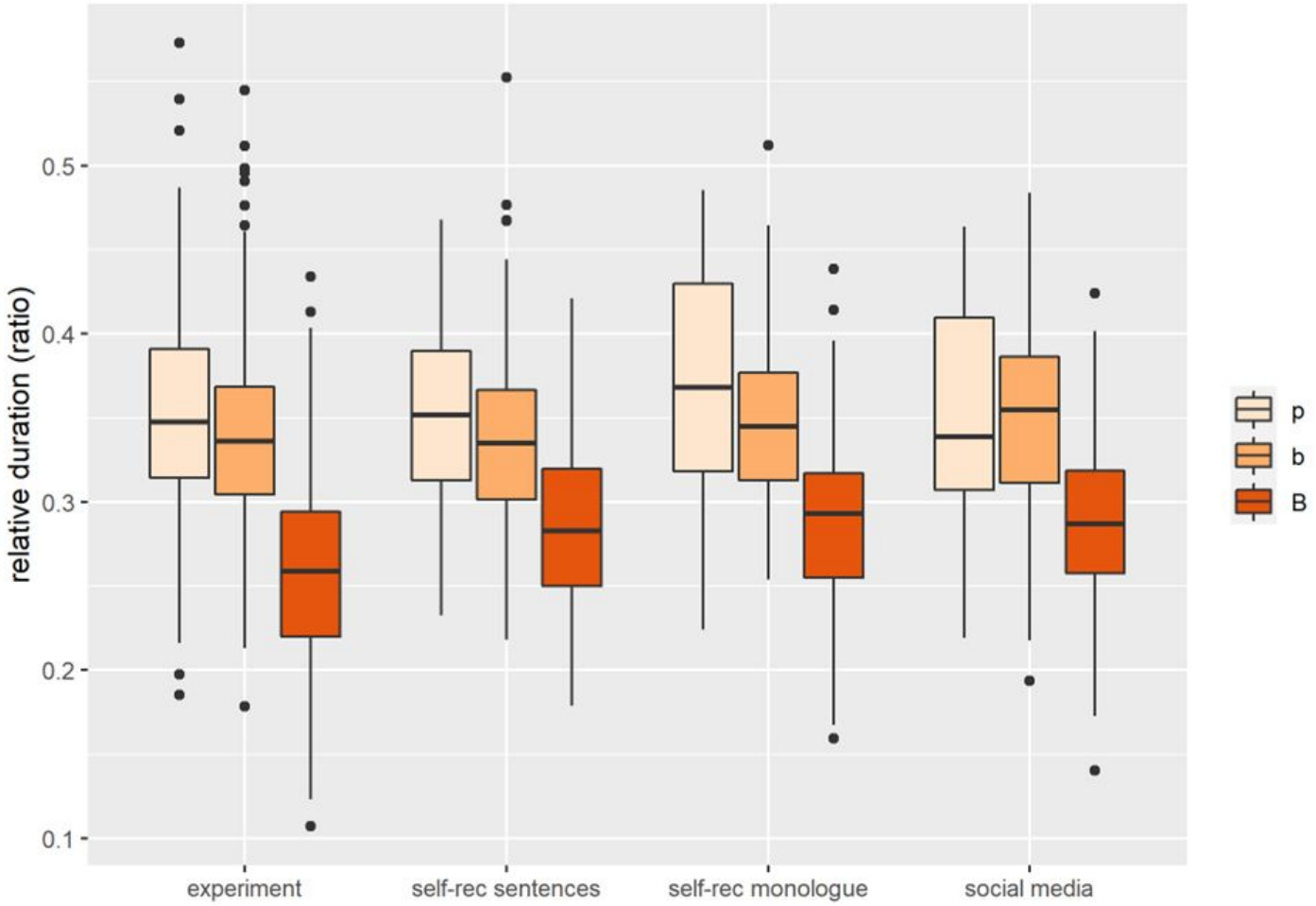
Data samples

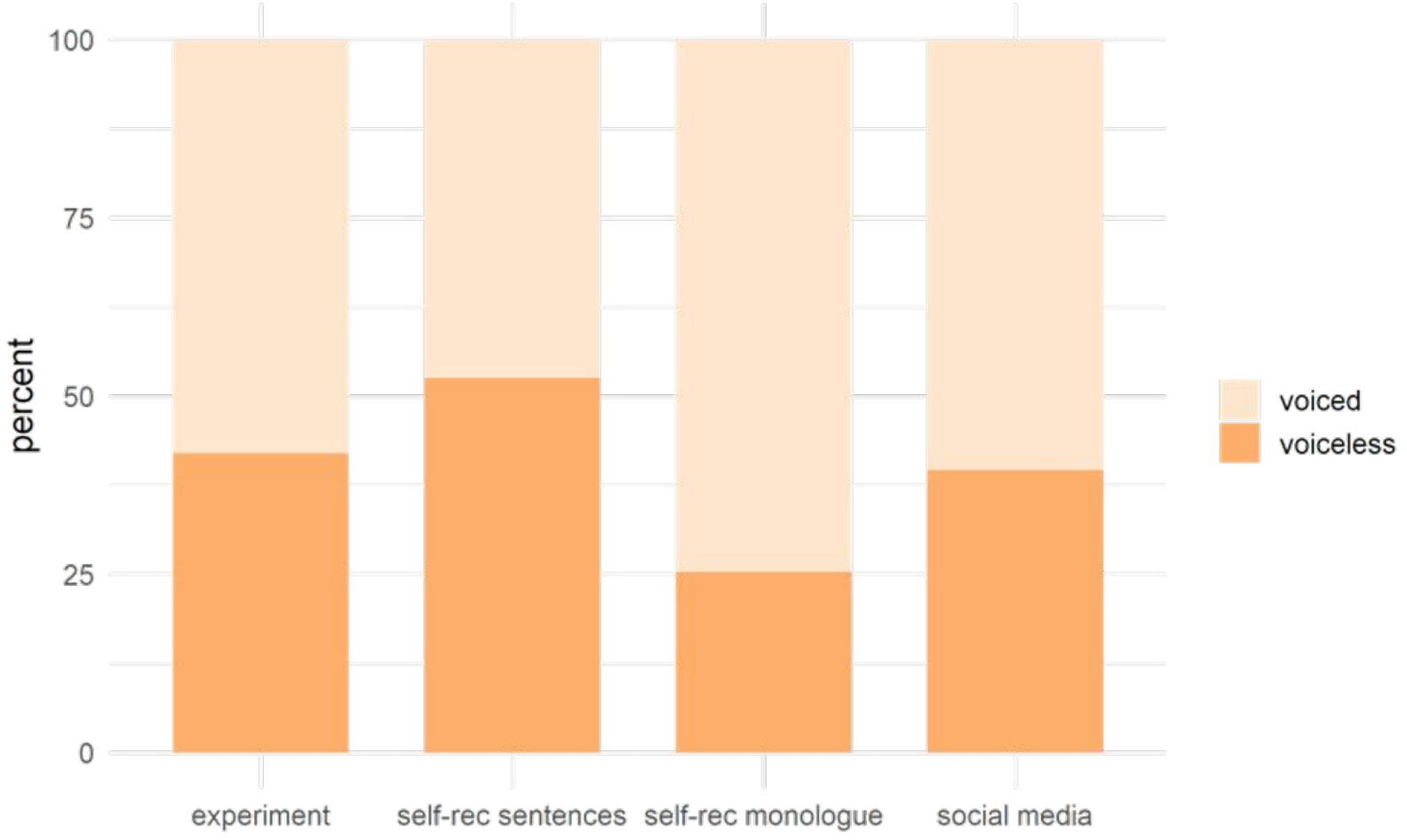
Table 1. Number of intervocalic /p b/ sounds analysed per participant

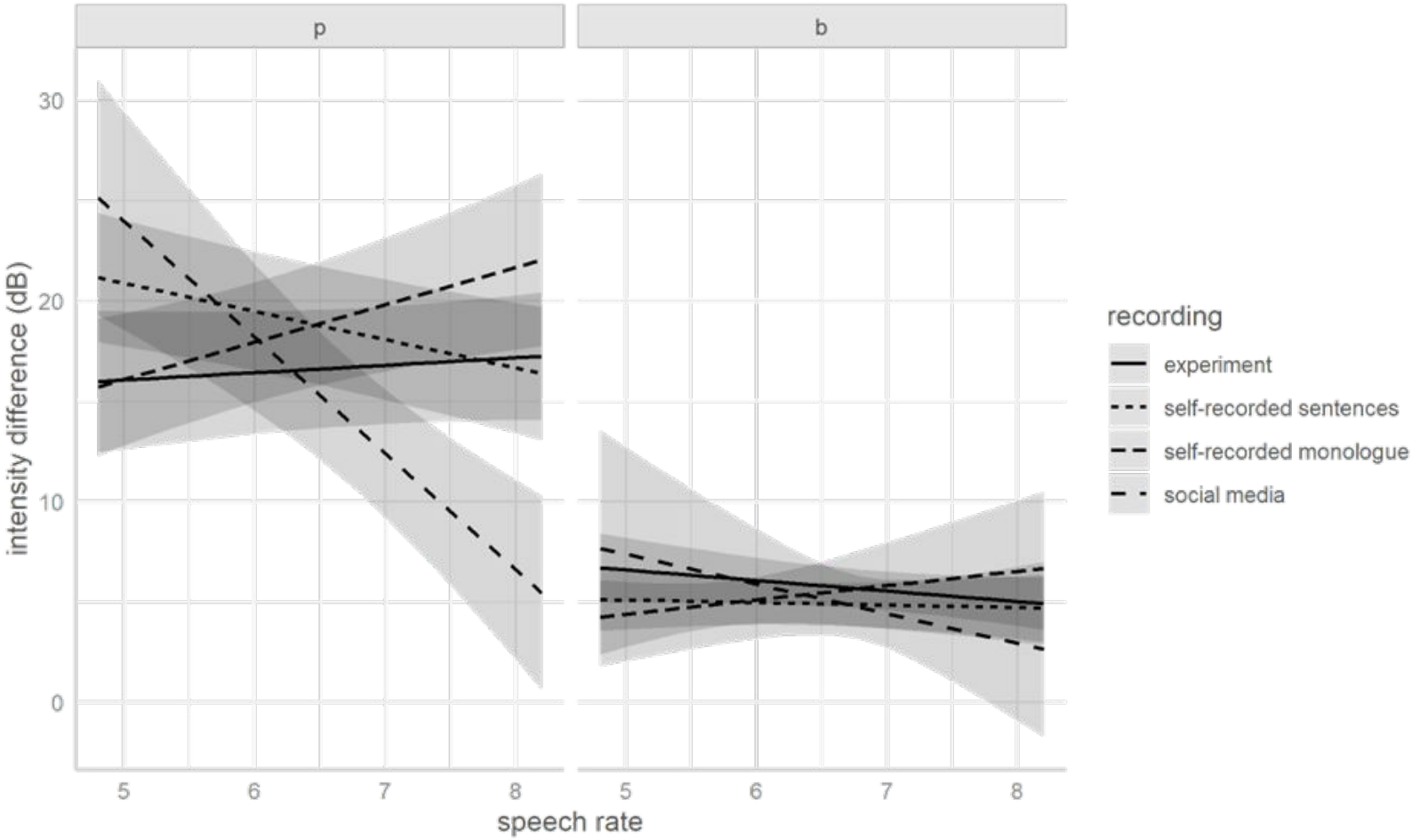
Participant	experiment	Self-recorded sentences	Self-recorded monologue	Social media	Total
P6	179	62	17	0	258
P7	195	58	11	0	264
P10	180	61	53	0	294
P14	179	58	18	24	279
P17	181	61	15	0	257
P19	157	57	23	8	245
P20	176	59	14	15	255
P21	170	62	33	45	310
Total	**1417**	**478**	**184**	**93**	**2172**


intensity difference (dB)
50
40
30
20
10
0
experiment
self-rec sentences
self-rec monologue
social media
p
b


experiment
self-recorded sentences
self-recorded monologue
social media
density
0.15
0.10
0.05
0.00
0
10
20
30
40
50
intensity difference (dB)
UR
p
b
voicing
voiced
voiceless


relative duration (ratio)
0.5
0.4
0.3
0.2
0.1
experiment
self-rec sentences
self-rec monologue
social media
p
b
B


100
75
50
25
0
percent
experiment
self-rec sentences
self-rec monologue
social media
voiced
voiceless


p
b
30
20
10
0
intensity difference (dB)
5
6
7
8
speech rate
recording
experiment
self-recorded sentences
self-recorded monologue
social media

Interim summary 5

- taking out **the experimenter** is not as powerful as changing **the task**
- reading or repeating sentences inhibits natural language processes to some extent
- more spontaneous productions may be related to **different cognitive mechanisms, speech planning and motor coordination** regardless of the speech rate
- **comparison with more authentic speech data samples** is necessary to confirm generalisations

Do the data help or not?

Too much detail vs

the trap of the incomplete picture

Compare results from the different quantitative studies mentioned

- percentages and generalisations often depend on **(sub)database** and **type of comparisons**….
- how much voicing/approximantisation makes the process optional but categorical and not gradient?

How reliable is making generalisations based on auditory analysis?

my own work (2016, 2018)

Weak minimal pairs in Gran Canarian

la cama	[lagáma]	'the bed'	la gama	[laɣáma]	'the range'
cuatro	[kwádro]	'four'	cuadro	[kwáðro]	'painting'
paca	[pága]	'pack/alpaca'	paga	[páɣa]	'pays'
grato	[grádo]	'pleasant'	grado	[gráðo]	'degree'
la poca	[labóka]	'the little'	la boca	[laβóka]	'the mouth'

General conclusions

What does working with different types of databases give us?

- helps elucidate factors affecting **sound change**
- helps get **the whole truth** about the studied processes
- helps identify **gradient vs categorical** changes (true categoricity?)
- helps identify **co-phonologies** by looking at intra-speaker differences
- helps overcome some aspects of **the observer's paradox**

In general:

It's good to have **comparative data,** ideally from the **same speakers** and make sure that our **generalisations** are not based on **incomplete data** or **false assumptions**

Thank you!

Slides and publications at www.karolinabros.eu