A longitudinal study of the oral properties of the French-English interlanguage : a quantitative approach of the acquisition of the / /-/i / and / /-/u / contrasts

Adrien Méli

Université Sorbonne
Paris Cité

université PARIS DIDEROT PARIS 7

Thèse de doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

École doctorale "Sciences du langage" (ED 132)

CLILLAC-ARP (EA 3967)

A longitudinal study of the oral properties of the French-English interlanguage

A quantitative approach of the acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts

Adrien MÉLI

Thèse de doctorat en linguistique anglaise

Dirigée par Nicolas BALLIER

Soutenue publiquement à Paris le 4 avril 2018

Directeur de thèse :	Nicolas BALLIER	Professeur à Paris Diderot
Examinateurs :	Emmanuel FERRAGNE	Maître de Conférences à Paris Diderot
	Richard WRIGHT	Professeur à University of Washington
Rapporteurs :	Sophie HERMENT	Professeure à Aix-Marseille Université
	Noël N'GUYEN	Professeur à Aix-Marseille Université

To Luke

Acknowledgements

This work would not have seen the light of day without the infinite, infinite patience of my wife Claire and my supervisor Nicolas BALLIER. I stand in debt to them.

Without fear of repeating myself, I would like to thank Nicolas BALLIER for his infinite, infinite, patience, wisdom, knowledge and *bienveillance*.

For their patient and knowledgeable support, I would like to thank the rest of my thesis committee: Emmanuel FERRAGNE, Sophie HERMENT, Noël NGUYEN, Richard WRIGHT.

For their help on so many issues, Geoff MORRISSON, Aurélie FISCHER; HUMA-NUM and Gérald FOLIOT; Taylor ARNOLD.

And my loving parents for their unwavering faith.

Résumé

Ce travail entreprend d'évaluer l'évolution de l'acquisition phonologique par des étudiants français des contrastes anglais /ɪ/-/iː/ et /ʊ/-/uː/. Le corpus étudié provient d'enregistrements de conversations spontanées menées avec des étudiants natifs. 12 étudiants, 9 femmes et 3 hommes, ont été suivis lors de 4 sessions espacées chacune d'un intervalle de six mois. L'approche adoptée est résolument quantitative, et agnostique quant aux théories d'acquisition d'une deuxième langue (par exemple Flege (2005), Best (1995), Kuhl et al. (2008)). Afin d'estimer les éventuels changements de prononciation, une procédure automatique d'alignement et d'extraction des données acoustiques a été conçue à partir du logiciel PRAAT (Boersma & Weenink (2013)). Dans un premier temps, deux autres logiciels, SPPAS (Bigi (2012a)) et P2FA (Yuan & Liberman (2008)) avaient aligné les transcriptions des enregistrements au phonème près. Plus de 90 000 voyelles ont ainsi été analysées. Les données extraites sont constituées d'informations telles que le nombre de syllabes du mot, de sa transcription acoustique dans le dictionnaire, de la structure syllabique, des phonèmes suivant et précédant la voyelle, de leur lieu et manière d'articulation, de leur appartenance ou non au même mot, mais surtout des relevés formantiques de F_0, F_1, F_2, F_3 and F_4. Ces relevés formantiques ont été effectués à chaque pourcentage de la durée de la voyelle afin de pouvoir tenir compte des influences des environnements consonantiques sur ces formants. Par ailleurs, des théories telles que le changement spectral inhérent aux voyelles (Nearey & Assmann (1986), Nearey (2012), Assmann et al. (2012)), ou des méthodes de modélisation du signal telles que la transformation cosinoïdale discrète (Harrington (2010)) requièrent que soient relevées les valeurs formantiques des voyelles tout au long de leur durée. Sont successivement étudiées la fiabilité de l'extraction automatique, les distributions statistiques des valeurs formantiques de chaque voyelle et les méthodes de normalisation appropriées aux conversations spontanées. Les différences entre les locuteurs sont ensuite évaluées en analysant tour à tour et après normalisation les changements spectraux, les valeurs for-

mantiques à la moitié de la durée de la voyelle et les transformations cosinoïdales. Les méthodes déployées sont les *k* plus proches voisins, les analyses discriminantes quadratiques et linéaires, ainsi que les régressions linéaires à effets mixtes. Une conclusion temporaire de ce travail est que l'acquisition du contraste /ɪ/-/iː/ semble plus robuste que celle de /ʊ/-/uː/.

Mots-clefs : réalisations vocaliques, acquisition phonologique, deuxième langue, approche quantitative, analyses formantiques, méthodes de normalisation, modélisation du signal.

Summary

This study undertakes to assess the evolution of the phonological acquisition of the English /ɪ/-/iː/ and /ʊ/-/uː/ contrasts by French students. The corpus is made up of recordings of spontaneous conversations with native speakers. 12 students, 9 females and 3 males, were recorded over 4 sessions in six-month intervals. The approach adopted here is resolutely quantitative, and agnostic with respect to theories of second language acquisition such as Flege (2005), Best (1995) or Kuhl et al. (2008). In order to assess the potential changes in pronunciations, an automatic procedure of alignment and extraction has been devised, based on PRAAT (Boersma & Weenink (2013)). Phonemic and word alignments had been carried out with SPPAS (Bigi (2012a)) and P2FA (Yuan & Liberman (2008)) beforehand. More than 90,000 vowels were thus collected and analysed. The extracted data consist of information such as the number of syllables in the word, the transcription of its dictionary pronunciation, the structure of the syllable the vowel appears in, of the preceding and succeeding phonemes, their places and manners of articulation, whether they belong to the same word or not, but also especially of the F_0, F_1, F_2, F_3 and F_4 formant values. These values were collected at each centile of the duration of the vowel, in order to be able to take into account of the influences of consonantal environments. Besides, theories such as

vowel-inherent spectral changes (Nearey & Assmann (1986), Nearey (2012), Assmann et al. (2012)), and methods of signal modelling such as discrete cosine transforms (Harrington (2010)) need formant values all throughout the duration of the vowel. Then the reliability of the automatic procedure, the per-vowel statistical distributions of the formant values, and the normalization methods appropriate to spontaneous speech are studied in turn. Speaker differences are assessed by analysing spectral changes, mid-temporal formant values and discrete cosine transforms with normalized values. The methods resorted to are the *k* nearest neighbours, linear and quadratic discriminant analyses and linear mixed effects regressions. A temporary conclusion is that the acquisition of the /I/-/i:/ contrast seems more robust than that of the /U/-/u:/ contrast.

Key-words: vocalic realizations, phonological acquisition, second language, quantitative approach, formant analysis, normalization methods, signal modelling.

Table of contents

List of figures

List of tables

Nomenclature

Roman Symbols

m	Median
$\hat{y}$	Predicted value

Greek Symbols

μ	Mean
σ	Standard deviation

Acronyms / Abbreviations

AIC	Akaike Information Criterion
BDM	Bark Difference Metric
CCI	Control/Compensation Index
CMUPD	Carnegie Mellon University Pronouncing Dictionary
DNV	Distance to Native Values
EPD	English Pronouncing Dictionary
GCOP	Global Cut-Off Point

HSD Honest Significance Difference

IPA International Phonetic Alphabet

KNN *k*-Nearest Neighbours

LDA Linear Discriminant Analysis

LMER Linear Mixed Effects Regression

LPD Longman Pronunciation Dictionary

MCP Mean Centile Product

MoA Manner of Articulation

NLMe Native Language Magnet Theory expanded

nPVI normalized Pairwise Variability Index

NSS Native Speakers' Subcorpus

OC Optimal Centile

OOD Onset-to-Offset Distance

P2FA Penn Phonetics Lab Forced Aligner

PoA Place of Articulation

PVI Pairwise Variability Index

QDA Quadratic Discriminant Analysis

RP Received Pronunciation

rPVI raw Pairwise Variability Index

SAMPA Speech Assessment Methods Phonetic Alphabet

SCP Sigma Centile Product

SL Source Language

SPPAS SPeech Phonetization Alignment and Syllabification

TAD Theory of Adaptive Dispersion

TL Target Language

TRP Time Reference Point

TTR Type/Token Ratio

VISC Vowel-Inherent Spectral Change

VOT Voice Onset Time

Introduction

The main purpose of this study is to compare the acquisition over time by French students of two English phonological contrasts, /iː/-/ɪ/ and /uː/-/ʊ/. The data used consist of recordings of task-driven spontaneous conversations between French learners of English from Université Paris Diderot and American or British assistants as part of the LONGDALE project (Goutéraux (2013)). 11 learners were recorded over three sessions at six-month intervals, and 12 learners over four sessions, also at six-month intervals. The approach adopted is resolutely quantitative and data-driven. An automatic process, based on two aligners, the SPeech Phonetization Alignment and Syllabification (SPPAS, Bigi (2012b), Bigi & Hirst (2012)), and the Penn Phonetics Lab Forced Aligner Toolkit (P2FA, Yuan & Liberman (2008)), was designed to extract information for *all* the vowels pronounced by the learners, not only the four phonemes, /ɪ/, /iː/, /ʊ/ and /uː/, under study. The extracted data for each vowel contains extra-linguistic, linguistic and acoustic information, available in two 92,000 $\times$ 542 spreadsheets (one for each aligner). The procedure of data extraction was also applied to three subcorpora: two subcorpora from recordings of the 12 learners who took part in all four sessions, with read lists of English words on the one hand, and a text read in French on the other; and another subcorpus of spontaneous conversations of native English speakers.

This brief introductory chapter has three aims: *(i)* to situate our investigation among the dominant frameworks in Second Language Acquisition (SLA); *(ii)* to discuss the perspective

followed to analyse our data; *(iii)* to outline how the chapters of this work partially contribute to the analysis of the interlanguage.

Theories in SLA

The source language (SL) of this study, French, features vowels, /i/ and /u/, which theories in Second Language Acquisition call "similar" (*c.f.* in particular Flege (1995), Flege (2005)) to these two contrasts. The French learners's task, represented in figure 1, therefore consists in dissassociating the two contrasts in the Target Language (TL). Such a parallel process, such a phonological symmetry between the SL and the TL make it possible to

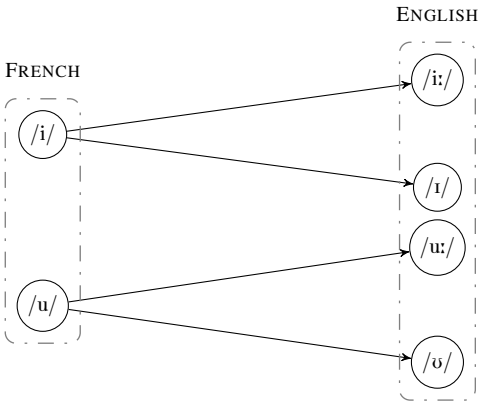


Fig. 1: Parallelism in the phonemic structures of the two contrasts

validate, or invalidate, the predictions of most SLA models, which only factor in phonemic structures when assessing the difficulty of acquisition. These models traditionally posit prosodically bijective predictions, whereby acquiring a given prosodic level in a target language is correlated to the structures of that same prosodic level already accessible to the learner. For phonemes, this is the case with models such as Kuhl et al. (2008) Native Language Magnet Theory expanded (henceforth, NLMe), or Flege (1995) Speech Learning Model (SLM), or Best (1995) Perceptual Assimilation Model (PAM). In the case at hand here, the predictions of such models form the Null Hypothesis, and can be formulated in the following way:

$\mathscr{H}_0$: *no differences exist in the acquisition of the two contrasts /iː/-/ɪ/ and /uː/-/ʊ/.*

The potential influence of extra-phonemic parameters such as phonemic or lexical frequency, syllabic structure, phonological neighbourhood, the existence and number of minimal pairs, etc., is therefore generally not taken into account. However, outside the field of SLA, formalizations of inter-level interactions exist: for instance, exemplar theories (Pierrehumbert (2001), Bybee (2007), Bybee (2010)) relate phonemic pronunciation to frequency of use; prosodic positions have been shown to influence the realization of phonemes (Keating et al. (2004)); syllabic structure and places of articulation have been shown to be connected (Tabain et al. (2004)); phonemic processing and speech-errors likewise depend upon phonological neighbourhood density and clustering coefficients (the similarities between phonological neighbours, Chan & Vitevitch (2010)).

Main purposes

The original goals of the study were therefore twofold: to establish whether the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts followed the same patterns of acquisition; and to establish whether extra-phonemic parameters might play a role in that acquisition. It is particularly in order to provide an answer to this second question that the nature and the purpose of this work evolved. In the process of fine-tuning the `PRAAT` (Boersma (2001)) scripts that generated the `TextGrids` from which the data was extracted, and as the quantity of collected information kept growing and growing, the nature of the research evolved from a perhaps more classic, results-driven, purpose-oriented study to one concerned with methods of processing and visualizing information. The unique nature of the extracted data, being altogether longitudinal, conversational and focused on vocalic realizations, demanded that specific methods of treatment and visualization be devised.

By being longitudinal, the collected information makes it possible to trace the various steps of the evolution of the learners' interlanguage, and more specifically of the quality of their vocalic realizations. Positing the existence of an "interlanguage" implies the existence of a transition between multiple states. These states have been investigated for consonants (*c.f. e.g.* Strik et al. (2007) for an example with computer assisted language learning), disfluences (Brand & Götz (2013)) or prosody (*c.f. e.g.* Trouvain & Barry (2007)), but less frequently for vowels (*c.f.* Gnevsheva (2015)). Longitudinal studies are particularly appropriate to unveil the properties of interlanguage, but such studies on learners' pronunciations are rare: Abrahamsson (2003) investigated the evolution of the production of Swedish codas by three Chinese learners in conversational speech, and demonstrated a U-shaped curve of acquisition. The present focus on vocalic realizations, also in conversational speech, challenges the possibility of resorting to the same methods of acoustic analyses as those used in experiments based on recorded lists of words. A lot of production studies focusing on vowel realizations thus resort to embedding the vowels in controlled consonantal environments such as /hVd/ (*c.f.* e.g. Hillenbrand et al. (1995), Ferragne & Pellegrino (2010), Clopper et al. (2005)) in order to minimalize and predict the potential influence of the consonants on the vowels' formants values.

Along the way, and because of the unique combination of features (longitudinal, conversational and on vocalic realizations), the purposes of the study therefore mutated, from an SLA contribution to proposals on how to visualize, process and analyze the complex, multi-layered data. Some of the major concerns this study of learners' vocalic realizations tries to address are the following: to provide reproducible protocols for the investigation of vocalic realizations on the basis of recordings; to determine the best processing treatments of acoustic data that make it possible to retain the maximum amount of information while preserving the specificities of conversational speech; to design methods of concisely and effectively representing longitudinal data for several speakers.

The philosophy of this work is resolutely neutral and unassuming, and based on a quantitative, data-driven approach. The methods used to process the information are compared to one another with the sole purposes of clarity and computational efficiency – not of obtaining results, not of rejecting or accepting the Null Hypothesis.

Chapter content

Chapter 1 details the procedures used to extract information from the original recordings; specifies what sort of information was collected; provides the explanations why these types of data were collected; and attempts to assess the quality of the extraction for the parts of the data which are not directly related to the phonemic categories of vocalic realizations. Chapter 1 is the answer to the original question whether extra-phonemic parameters might play a role in the acquisition of the phonemic contrasts /ɪ/-/iː/ and /ʊ/-/uː/. Do different prosodic categories, such as syllable or words, permeate interlanguage? To find out, syllables and words also had to be aligned by the two aligners used, and collected for each vocalic realization collected. With the potentially infinite variety of consonantal environments pertaining to conversational speech, some sort of control had to be introduced too. This goal led to the retrieval of formant values at each centile of the vowels' durations, in keeping with theories such as vowel inherent spectral change (VISC, Nearey & Assmann (1986), Morrison & Nearey (2006), Hillenbrand (2012), Morrison (2012)) or mathematical transformations of the raw signal in Hertz such as discrete cosine transforms (DCT, Harrington (2010)). But in order to distinguish, within the formant values, what exactly pertained to natural formant transitions from what might pertain to interlanguage, bases for comparison, *i.e.* native references, were needed. This realization led in turn to the creation of the native subcorpus, using the same procedure as the one applied to the main corpus and the two LONGDALE subcorpora. In an attempt to assess the quality of the automatic alignment and of the automatic extracted data – a recurring concern in this work –, missing values in

the datasheets are investigated, along with the natures of the syllabic structures. These are verified and compared to the pronouncing dictionaries used by the aligners and the algorithm designed in this work. Finally, a study of the durations of the vowels and of speech rate aims at assessing the quality of both the extraction and the learners' discourses.

Chapter 2 sets out to determine whether certain aspects of interlanguage independent from speakers' idiosyncrasies exist – more specifically, whether cross-speaker patterns of acquisition of the two /ɪ/-/iː/ and /ʊ/-/uː/ contrasts exist. It begins by ensuring that the formant values on all centiles are within reasonable ranges. In keeping with studies of vowel inventories, such as Al-Tamimi & Ferragne (2005) or Gendrot & Adda-Decker (2007), how the phonemes are distributed in the vocalic trapezoid is investigated, and compared with native both French and English native values. The length of the /ɪ/-/iː/ and /ʊ/-/uː/ vectors is measured against the convex hulls linking the outermost vowels in the F_1/F_2 space. The skewness in the distribution of the phonemic categories is then also surveyed, based on the assumption that the gaps in the frequencies of occurrences between the different vocalic categories is very likely to exert influence on the learners' interlanguage. This observation led to a comparison of the various methods of normalization, in order to find out which suits best a dataset with uneven number of occurrences across the phonemic categories.

The acquisition of a language being often a very different experience from one learner to another, chapter 3 focuses on trying to specify the evolution of the interlanguage of each of the 12 speakers who took part in all four sessions of the LONGDALE project. The theory of VISC is applied to the main corpus and to the native subcorpus, and the lengths of the learners' vectors starting at 20% of the vowels' durations, and ending at 80% in the F_1/F_2 vocalic space are compared to their native counterparts. In a further attempt to assess the states of acquisition, the robustness of the findings are tested by looking at the standard deviations of the vectorial values. With a growing body of evidence pointing to actual differences in the acquisition of the four phonemes under study, a machine learning classification method,

the *k*-nearest neighbours, was run on the main corpus, with a native dataset from Peterson & Barney (1952) used as the training set. Confusion matrices are then investigated, more specifically the phonemic distributions of the predictions for /ɪ/, /iː/, /ʊ/ and /uː/. This experiment making it hard to visualize the longitudinal evolution of the interlanguage, a study using linear mixed-effects regressions was then carried out. The formant values and their standard deviations served as response variables, and the effect of session, *i.e.* the evolution over time, was investigated. Models predicting several sorts of changes were compared, and the acquisition of the four phonemes showed different evolutions. Finally, the entire signals for the first three formants were modelled using DCTs, and comparisons were there again made with native values. The dispersions of /ʊ/ and /uː/ were greater, and the acquisition of /uː/ in particular seemed to be less robust than those of the other phonemes. The chapter ends on an ultimate comparison of models based on mid-temporal formant values on the one hand, and on DCTs on the other. From this comparison, one of the strongest recommendations of this work is formulated – that DCTs are particularly appropriate for the study of conversational data.

Chapter 1

Corpus and Data

This chapter details the procedure implemented to obtain the data which is analyzed in chapter 2.

1.1 Corpus

Subsection 1.1.1 describes the profile and background of the participants of the study. Subsection 1.1.2 details the content and characteristics of the recordings.

1.1.1 Participants and metadata

25 participants, 20 women and 5 men, were recorded between September 2009 and May 2013 as part of the LONGDALE project (Goutéraux (2013)). Metadata was collected from a form the participants filled in themselves. There are two sets of participants, all of them students from Université Paris Diderot.

The first set comprises students, 8 females and 2 males, who completed three sessions. The second set is made up of the students, 10 females and 3 males, who attended all four recording sessions. Of lesser interest perhaps, but still worthy of note, is the fact that students whose ID number are inferior to 110 were recorded in September 2009 for Session 1, June

Table 1.1: Summary of the participants' metadata

Student ID	Sex	Number of sessions	Native languages	Days spent in ESC
DID0014	Male	4	French, Vietnamese	37
DID0020	Female	3	French	380
DID0024	Female	4	French	7
DID0035	Female	4	French	7
DID0039	Female	4	French	0
DID0062	Female	4	French	14
DID0068	Male	4	French	14
DID0071	Female	4	French	44
DID0096	Female	4	French	35
DID0106	Female	4	French	44
DID0108	Female	4	French	3
DID0119	Female	3	French	30
DID0126	Female	3	French	400
DID0127	Female	3	French	7
DID0128	Female	4	French	7
DID0129	Female	3	French	7
DID0135	Female	4	French	7
DID0138	Female	3	French	270
DID0145	Female	3	French	30
DID0146	Female	3	French	30
DID0156	Male	3	French, Greek	30
DID0168	Male	4	French	14
DID0213	Male	3	French	450

or November 2010 for Session 2, April 2011 for Session 3, and May 2012 for Session 4. Students with ID numbers superior to 110 were recorded one year later: in September 2010 for Session 1, October 2011 for Session 2, April 2012 for Session 3 and finally May 2013 for Session 4.

All students were beginning a three-year course in English at Université Paris Diderot at the time of recording of their first session. This was a second course or a minor for three participants: student DID0213's major was history; student DID0138 had obtained a Bachelor in biology; and student DID0035, a Master in sociology.

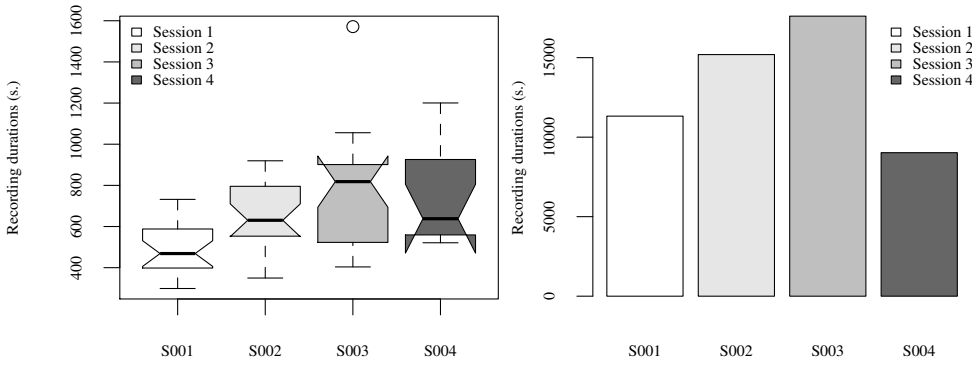
None of the participants reported proficiency in any other language than French or English, except student DID0156 who reported Greek as her native language, along with

French; and student DID0014, who reported Vietnamese and French as native languages, and said he spoke French, Vietnamese and English at home. The number of days spent in an English-speaking country (ESC) was also collected after each session, but reported numbers did not change between the first and the last session.

Two students, DID0128 and DID0213, took the Test of English as a Foreign Language (TOEFL), and reported a score of 107/120 and 111/120 respectively. Other students (DID0014 and DID0024) reported scores without mentioning what test or examination they had been taking. Table 1.1 summarizes the participants' metadata.

1.1.2 Recordings and tasks

All 82 interviews were recorded in an individual stereo 16-bit resolution sound file at a sampling rate of 44100 Hz captured in an uncompressed, pulse code modulation format using an Apex435 large diaphragm studio condenser microphone with cardioid polar pattern. They all begin with an interview of the learner conducted by a native speaker. The learner was then presented with a task which changed with the session (*cf.* below). The native speaker and the learner each had a microphone, and were recorded on a separate channel, although some crossover between the two channels happened (e.g. the interviewer's utterances were recorded on the interviewee's channel). The interviews were not conducted in a deaf room: the quality therefore varied greatly from one recording to another, or from one moment during the interview to another, with background noises such as footsteps, cars or distant chatter sometimes audible. The recordings lasted 656 seconds on average, with great per-speaker and per-session variability, as shown in figure 1.1. The comparatively shorter aggregated duration for Session 4 displayed in figure 1.1b can be explained if we recall the lower number of participants for that session: 13 students, against 25 for all other three sessions. However, too much importance should not be granted to total recording durations. Section 1.2 will present a more accurate assessment of learners' actual speaking time and speech rate.



(a) Medians and quartiles of recording durations.

(b) Recording durations.

Fig. 1.1: Aggregated per-session summary of recording durations.

Tasks and questions asked changed at each session. The design of the tasks and questions replicated those of the Louvain International Database of Spoken English Interlanguage (LINDSEI, Brand & Kämmmerer (2006)).

Sessions 0 & 2

In session 1 and 2, participants were to answer one of the three following questions:

1. *Suppose you have time and money to travel or move to a different country/city, where will you go? Why? How will you organise your new life?*
2. *Can you tell me about an important event, experience or meeting which has made a difference or changed your life in the past six months?*
3. *Do you feel creative? Tell me about a work of art you would like to create or participate in: a play, a film, a musical event, a book, a painting, a computer game, etc. How would you go about it?*

Figure 1.2 shows how many students selected one of the three possible subjects in the two sessions. Question 2 was overwhelmingly chosen in Session 1, only to be discarded in

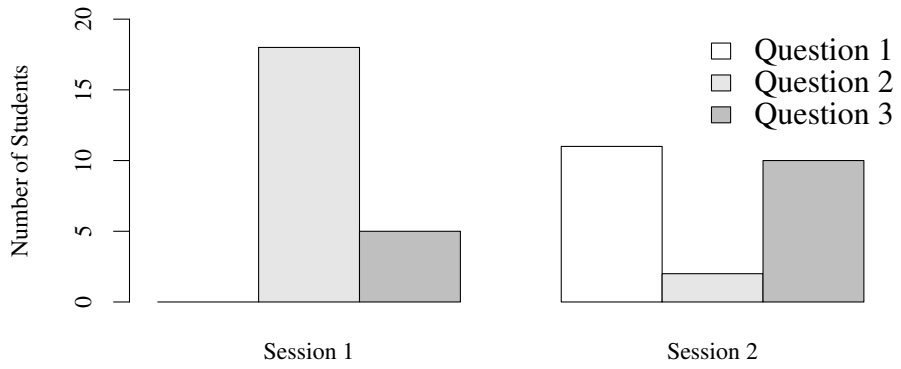


Fig. 1.2: Number of students who chose to answer questions 1, 2 or 3 in Sessions 1 & 2.

Session 2. Besides, out of the 5 students who chose task 3 ("*Do you feel creative?*") in Session 1, only one, DID0062, answered the question. The other four spoke of a film they had seen, and therefore failed to answer the question in a relevant manner. The same mistake did not happen again with *any* of the 10 students (who sometimes turned out to be the same, in the case of speakers DID0035, DID0106, DID0119 & DID0128) whose chose to answer that question in Session 2. This fact may serve as an indication of a certain improvement in understanding tasks formulated in English after a year at university studying the language.

Session 3

In session 3, the interviewers were requested to read the following prompt:

You are going to see four works of art (paintings), one after the other. I'd like you to react to each of them quite spontaneously and tell me how you feel about them.

They were also given the following optional additional questions:

- *Can you justify, explain why you like or dislike picture one, two, three, four?*
- *Which of these four pictures would you like to have at home, in your room?*
- *If you were to take one of those pictures to illustrate a book you want to write, which one would you choose?*

The four paintings they were to describe were shown to the participants in the following order[1]:

1. *Carnation, Lily, Lily, Rose*, by John Singer Sargent (1885-1886).
2. *Nude, Appledore, Isle of Shoals*, by Childe Hassam (1913).
3. *Carcass of Beef*, by Chaim Soutine (1925).
4. *The Garden*, by Andreas Schulze (2009).

Recordings of Session 3 were the longest on average: they lasted 704 seconds, against 452s., 607s. and 694s.[2] for Sessions 1, 2 and 4 respectively. Student DID0020's interview in Session 3 was by far the longest (1571s.): student DID0024's fourth session, the second longest recording in the corpus, lasted 1200s., *i.e.* it was 5 minutes shorter.

Session 4

In session 4, the participants had to perform a map task as designed by Anderson Anderson et al. (1991). Figure A.1 in section A.1 shows the two maps that were given to the learner and the native speaker. The maps share common landmarks, but some of these landmarks are unique to each map. The map that was given to the learner contains an itinerary, with a starting point and a finishing point. The native speaker was given the map without the itinerary. This informational gap aimed at eliciting questions from the learner.

Of interest also for this study are the extra reading tasks the learners were given at the end of this session. The students were asked to read lists of words featuring all the vowels in English. These words were grouped according to the vowels they contained. They were also asked to read a short text in French. Both this list and the text in French can be found in Appendix A. The recordings of the 13 students who completed these tasks have also been analyzed, and the acoustic information extracted from the text in French and the list of

[1] Three of the four paintings that were presented are copyrighted and may not be reproduced here.

[2] All means were calculated using the respective number of participants in each session, *i.e.* 25 in all sessions but Session 4, which had 13 participants.

Table 1.2: Summary of the participants' chosen tasks for Sessions 1 & 2 and individual recording durations.

Student ID	Question answered in:		Total recording time (in s.)			
	Session 1	Session 2	Session 1	Session 2	Session 3	Session 4
DID0014	2	3	544	628	1055	849
DID0020	2	2	439	465	1571	NA
DID0024	2	1	298	607	758	1200
DID0035	3	3	589	754	858	530
DID0039	2	3	648	755	887	581
DID0062	3	1	605	491	621	622
DID0068	2	2	586	624	915	520
DID0071	2	1	327	630	974	652
DID0096	2	1	574	483	758	1000
DID0106	3	3	317	601	1003	289
DID0108	2	3	326	631	879	536
DID0119	3	3	462	874	403	NA
DID0126	2	3	468	828	524	NA
DID0127	2	1	367	533	520	NA
DID0128	3	3	411	919	869	718
DID0129	2	1	652	349	413	NA
DID0135	2	1	731	431	513	NA
DID0138	2	3	504	868	935	NA
DID0145	2	1	429	803	827	NA
DID0146	2	1	407	787	437	NA
DID0156	2	1	522	658	556	NA
DID0168	2	3	388	572	818	834
DID0213	2	1	714	891	505	NA

English words serves as reference for native formant values and phonological knowledge of the target language respectively.

Finally, in all sessions, the students were asked questions about their personal and academic projects, and, from Session 2 onwards, what they thought about the course they had been following.

Table 1.2 lists the questions the learners chose to answer in sessions 1 & 2, as well as the duration of each recording. A per-speaker graphical representation of these durations can be found in figure 1.3.

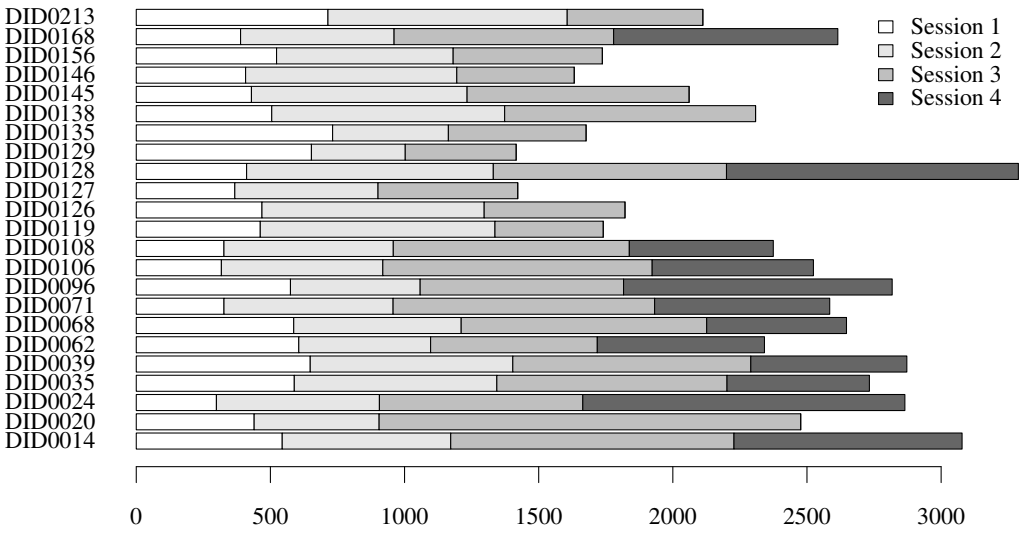


Fig. 1.3: Per-speaker recording durations.

1.2 Workflow

The purpose of this section is to detail the method that was implemented in order to obtain the final database. Section 1.2.1 presents the procedure from the original sound files to the final multitier TextGrids. The main script to generate the data is described in section 1.2.2;

the resulting TextGrids are presented in section 1.2.3; details for the final `.csv` spreadsheet can be found in section 1.2.4.

1.2.1 Global procedure

The goal to reach when designing the alignment process was to obtain a high number of automatically aligned transcriptions in an efficient way. One key aspect was that aligners naturally work best with native speech. It was therefore critical that alignment errors due to learners' mispronunciations should be contained. It was decided that the best course would be to feed the aligners as short extracts as possible in order to minimize the risk of a domino effect, whereby an alignment error might spread and create other errors in the extract.

The initial documents consisted of the original recordings and their transcriptions. The transcriptions were compliant with the requirement of the LONGDALE project, and contained XML-like tags that flagged events such as speakers' turns, overlapping speech, whispers, laughter. Transcribers had also been instructed to mark certain aspects of pronunciation such as pauses or "the" pronounced /iː/. The exact guidelines can be found in Appendix B. The following paragraphs present the procedure for a single recording. The procedure is also summarized visually in figure 1.4.

First a Python script (later, a more efficient Perl script) edited out all tags, punctuation marks and pronunciation-related flags, and formatted the text so that lines contained no more than 80 characters. The corresponding recording was then downsampled to 16kHz. The reason is that the two aligners used, the SPeech Phonetization Alignment and Syllabification (SPPAS, Bigi (2012b), Bigi & Hirst (2012)), and the Penn Phonetics Lab Forced Aligner Toolkit (P2FA, Yuan & Liberman (2008)), both require 16kHz sampling (P2FA recommends 11kHz but accepts 16kHz). The sound file was therefore downsampled from 44kHz to 16kHz using Sox (Bagwell (2018)), and then opened in Audacity (version 2.1.0), where a Label Track was created. It was then played in Audacity, and a label was added at the end of

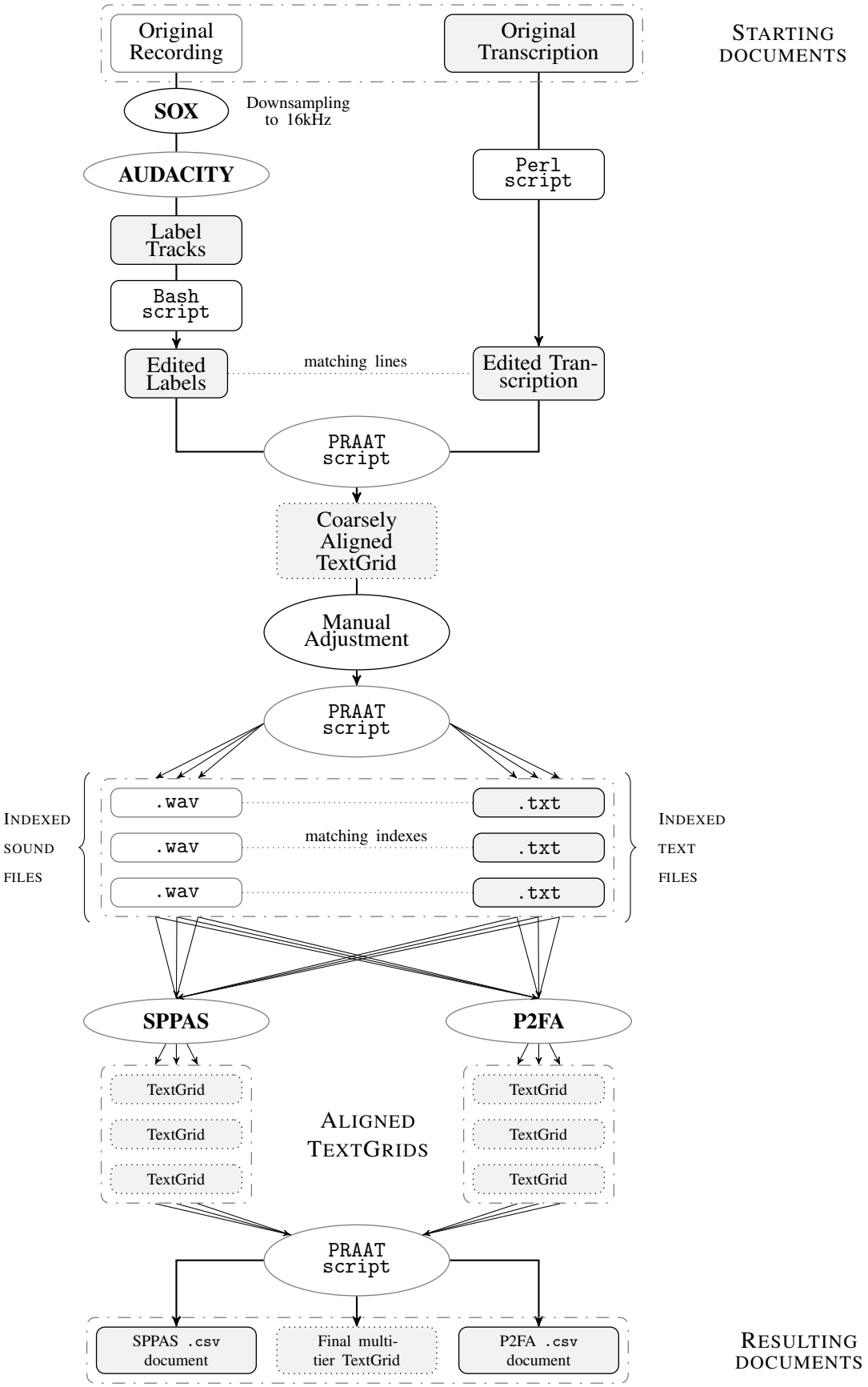


Fig. 1.4: Flow chart from the original LONGDALE recordings and transcription files to the final `.csv` spreadhseet and multitier TextGrid.

every 80-character line. The labels were then exported in Audacity to a two-colum textfile containing the time stamps corresponding to each line of the transcription file. Another Python script (later a far simpler bash line) deleted one column along with any labels that may have been added while listening to the file. At that stage, two files had been obtained: the 80-character transcription file, and the label file with the time stamp. The latter contained exactly one line less than the former. Both files were then fed to a Praat script (Boersma & Weenink (2013)), which created a TextGrid. Version 5.4.08 was used to perform the calculations. The script (`PRAAT01`) added boundaries at times matching the time stamps from the label file, and labeled the newly created interval with the corresponding line from the transcription file. This yielded a coarsely aligned TextGrid, which was then manually edited: the recording and the TextGrid were opened in Praat, and boundaries were added or adjusted manually. Boundaries were added with the two following concerns: *(i)* to select speech which was likely to be correctly aligned; *(ii)* to obtain reasonably short sequences. Noisy parts, extreme overlapping speech, non-existing expressions, grunts, coughs, fits of

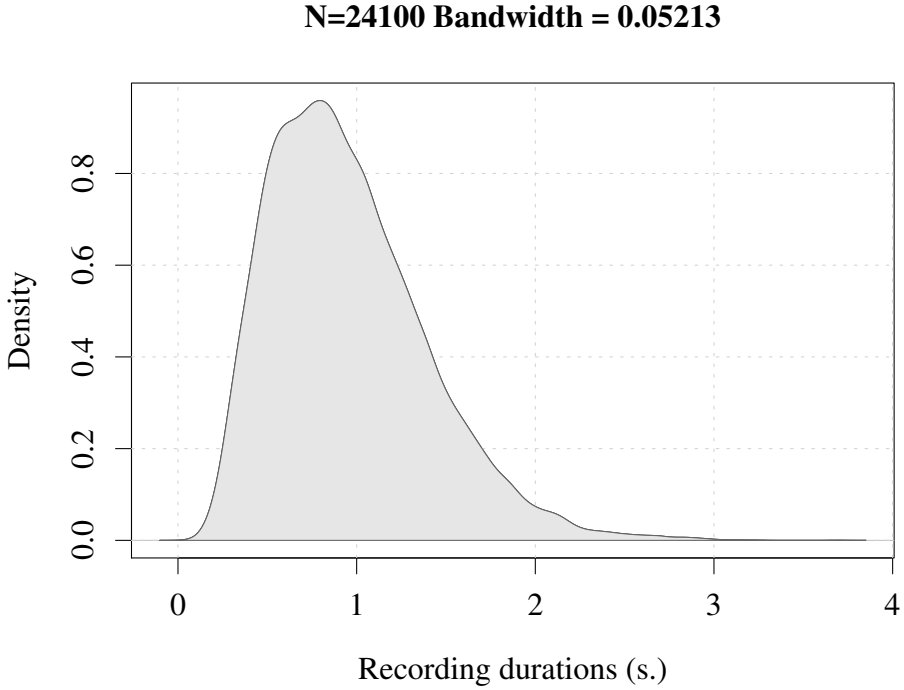


Fig. 1.5: Kernel density plot of short interval durations.

laughter, were therefore labeled as non-exploitable in the TextGrid, and boundaries were added, sometimes in the middle of sentences. When adding a boundary, the spectrogram was visible. More often than not, mid-sentence boundaries mark the beginning of easily recognizable phonemes, especially voiceless plosives or fricatives. Boundaries were added with no considerations of meaning or syntax. The average duration of these intervals over the whole corpus is 0.95 second, and 24,398 such intervals were created and processed. The maximum duration of their corresponding sound files is 3.69 seconds. Figure 1.5 shows the kernel density plot of the durations of these short manually aligned intervals. Figure 1.6

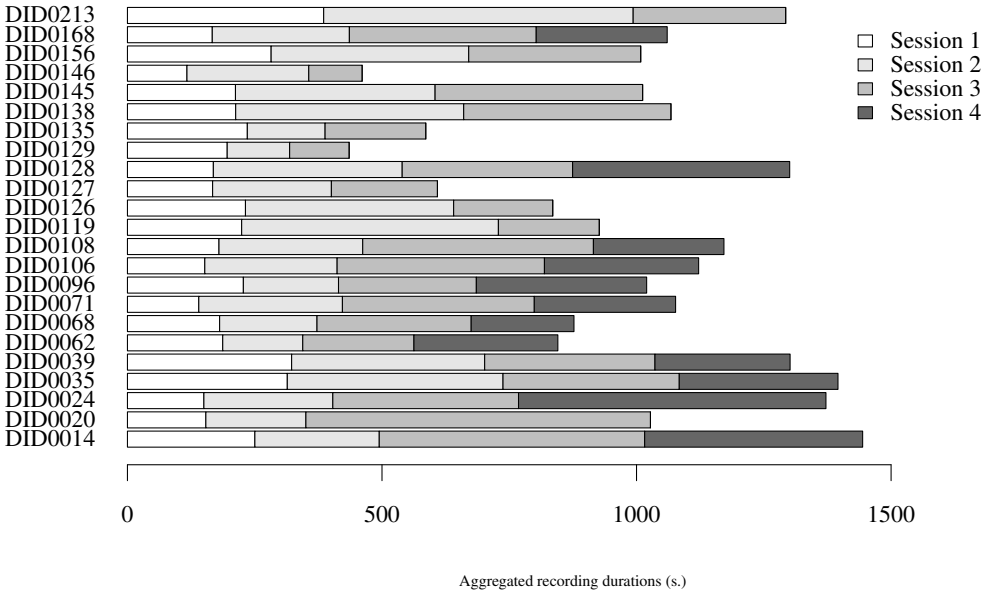


Fig. 1.6: Per-speaker aggregated durations of extracted speech.

shows the aggregated duration of the learners' speech actually extracted and analyzed, *i.e.* with pauses, backchanneling, grunts, laughter and the interviewer's own speech, removed. These durations correspond to the sum of the length of each short `.wav` file for each speaker in each session. At that point, the TextGrid has three tiers: one for the learner's transcription, one for the native speaker's, and an empty tier. Once this lengthy procedure was over, the newly adjusted TextGrid and the original recording were fed into another Praat script

(PRAAT02), which did two things: *(i)* it created a sound file and a corresponding transcription file for each selected interval in the main file. The transcription file is a simple `.txt` file containing the label of its matching interval in the main TextGrid. Both the sound file and this transcription file were put in a subfolder. Intervals corresponding to non-exploitable passages were left aside; *(ii)* each pair of `.txt` and sound file was indexed with the number of the interval in the main file they corresponded to. After the completion of the script, the subfolder therefore contained as many short `.txt` and sound files as intervals selected for analysis in the main TextGrid.

The next step was to align these transcriptions with their sound files automatically. The first aligner used was SPPAS. The version used for processing the files was v1.7.0. One setting was modified from default, `Syllabification`, in which an interval tier was used, "PhnTokAlign", and not the default "TokensAlign". This was to ensure syllabic alignment on the phonemic tier, not on the word tier. SPPAS takes `.txt` and `.wav` files as input and returns PRAAT TextGrids with tier intervals aligned on phonemic and syllabic boundaries. For syllabic alignment, SPPAS uses an algorithm based on a set of rules which ranks phoneme classes according to their likelihood to be in onset or coda position. However, as v1.7.0 did not ship with an algorithm for English syllables[3], the built-in list of rules for French was adapted for English phonemes. The modification can be checked in section C.3 in appendix C. The reasoning underlying these changes is that French syllabic structures may have an effect on English realizations at the phonemic level. Having SPPAS syllabify learners' utterances using its own built-in algorithm for French provides the means to test this assumption. The syllabifying processes are explained in more detail in section 1.6. Once the subfolder was processed, it contained a merged TextGrid with tiers for phonemes, syllables and words. The second aligner used in this study was P2FA (v1.002), which is based on version 3.4 of the Hidden Markov Model Toolkit (HTK, Young et al. (2006)). Just like SPPAS, P2FA uses a

[3] Earlier versions did, but unfortunately syllabification in English cannot work with rules that only access the phonemic level. One simple example can show this: "present" *(v.)* ⇒ /pri.ˈzent/ – "present" *(adj.)* ⇒ /ˈprez.ənt/.

`.txt` file (but with a capitalized transcription) and a `.wav` file as inputs. A simple bash script: *(i)* capitalized the transcription files; *(ii)* downsampled the sound files; and *(iii)* ran P2FA in the entire subfolder. P2FA returns a TextGrid with word and phoneme alignment for each sound file.

Finally, another homemade[4] PRAAT script (`PRAAT03`) performed the following things: *(i)* it reintegrated the merged SPPAS TextGrid and the P2FA TextGrid into the original main TextGrid (from which the short intervals had been extracted); *(ii)* parsed the SPPAS and P2FA phonemic tiers and collected acoustic information (*cf.* section 1.2.3 below) about each phoneme; *(iii)* retrieved the pronunciation of each word from the Longman Pronunciation Dictionary (LPD, Wells (2008)); *(iv)* created dedicated syllabic tiers, a process described in the next section 1.2.2.

This workflow was applied to all 82 recordings. 92,332 SPPAS-aligned and 92,059 P2FA-aligned vowels[5] were automatically extracted. The next subsections detail the process of extraction: section 1.2.2 describes the structure of `PRAAT03`; the structure of the `Praat` multitier TextGrid obtained is described in section 1.2.3; the dataframes and their headers are detailed in section 1.2.4.

1.2.2 `PRAAT03`

All the data collected for analysis in the following chapters comes from script `PRAAT03`. This section explains how the script works in detail.

`PRAAT03` has three[6] main loops: *(i)* loop 1 parses the TextGrid file names in the subfolder; retrieves the index, contained in those names, which matches the interval number in the main TextGrid; selects the SPPAS merged TextGrids and the P2FA TextGrids; copies and pastes their boundaries and labels to the main file at the indexed interval number; deletes

[4] "Homemade" is not perfectly accurate. A lot of inspiration was drawn from Mietta Lennes's scripts, especially in terms of what *could* be done.

[5] The reasons why the total count of vowels differ between SPPAS and P2FA are explained below.

[6] Technically, there is a fourth loop, as the script is able to process several main files in the same folder.

micro-intervals (*i.e.* intervals with durations inferior to 0.001 second caused by tiny variations in interval timing when extracting and concatenating the short files from and to the main file). *(ii)* the second loop scans the SPPAS-aligned phonemic tier of the multitier TextGrid created in the first loop; retrieves its pronunciation in the LPD for each phoneme at the beginning of a word; looks for English syllable boundaries in the succeeding phonemes; adds a syllable boundary on the specifically added dedicated tier if the phoneme is the syllable coda; identifies whether the current phoneme is the syllable nucleus, *i.e.* whether it is a vowel; if so, a `.csv` textfile is appended with information of a form extensively presented in Appendix D and explained in section 1.2.4. *(iii)* the third loop is the P2FA version of the second loop. The only significant difference is that the French syllable tier had to be inferred from its SPPAS-generated counterpart.

The script adds 9 tiers – *i.e.* tiers which were not created by the aligners and imported from the shorter TextGrids: for both aligners, the English syllable tiers, the LPD pronunciation syllable tiers and the stress tiers (referred to below as the "English syllable tiers"). For P2FA, the French syllable tier is inferred from the SPPAS French syllable tier. Finally, the Pairwise Variability Indices (PVI) tiers, which fuse adjacent consonants and vowels together regardless of syllable or word boundaries. The order of creation varies: the English syllable tiers are created at run-time, and so is the P2FA French syllable tier. However, this tier requires SPPAS tiers to have been created prior to its creation, which explains why the loop dedicated to SPPAS tiers must take place before the loop dedicated to P2FA tiers. The consonantal and vocalic intervals for the PVI tiers are calculated for each main TextGrid after completion of the two aligners' loops. Table 1.3 summarizes the origin of each tier and what their dependencies are. Word and phoneme tiers were aligned by the aligners' internal algorithms. English syllable and stress tiers are created in `PRAAT03`, with boundaries aligned on the LPD syllabic transcriptions. PVI tiers essentially consist in duplicating the phonemic tiers of the two aligners and then merging adjacent intervals featuring the same manner

Table 1.3: Summary of tier names, sources and dependencies

Tier Number	Tier Name	Boundary Source	Dependencies
1	Student	Manual	None
2	Assistant	Manual	None
3	Empty	N/A	N/A
4	P2FA: phonemes	P2FA	Student
5	P2FA: words	P2FA	Student
6	P2FA: FS	PRAAT03	SPPAS: FS
7	P2FA: ES	PRAAT03	P2FA: phonemes
8	P2FA: LPDS	PRAAT03	P2FA: ES
9	P2FA: Stress	PRAAT03	P2FA: ES
10	SPPAS: phonemes	SPPAS	Student
11	SPPAS: words	SPPAS	Student
12	SPPAS: FS	SPPAS	SPPAS: phonemes
13	SPPAS: ES	PRAAT03	SPPAS: phonemes
14	SPPAS: LPDS	PRAAT03	SPPAS: ES
15	SPPAS: Stress	PRAAT03	SPPAS: ES
16	PVI: SPPAS	PRAAT03	SPPAS: phonemes
17	PVI: P2FA	PRAAT03	P2FA: phonemes

of articulation (reduced to vowels or consonants only). The French syllable tiers feature crucial differences, from one aligner to the other, in the way they were generated. Compared to P2FA, SPPAS features an extra syllabifying algorithm based on a set of user-definable rules (*c.f.* section C.3). The SPPAS-aligned French syllable tier was generated from the default French rules provided with SPPAS, but modified to remove sounds specific to French (*e.g.* nasal vowels), and to include specifically English sounds (*e.g.* interdental fricatives or /h/). The list of English phonemes, converted from the list of French phonemes natively provided by SPAAS, can be found in section C.3. The P2FA-aligned French syllable tier was generated by PRAAT03 from the SPPAS-aligned French syllable tier. To understand how it was generated, a closer look at how the data was collected is necessary. The rough outlines of the process and the obstacles that were encountered are explained in the paragraph below.

The acoustic parameters used to extract the pitch, the formants and the intensity are standard: they follow the recommendations of the PRAAT manual. The code (*c.f.* section C.4) used to generate pitch, formant and intensity values is sex-dependent. For pitch analysis, the

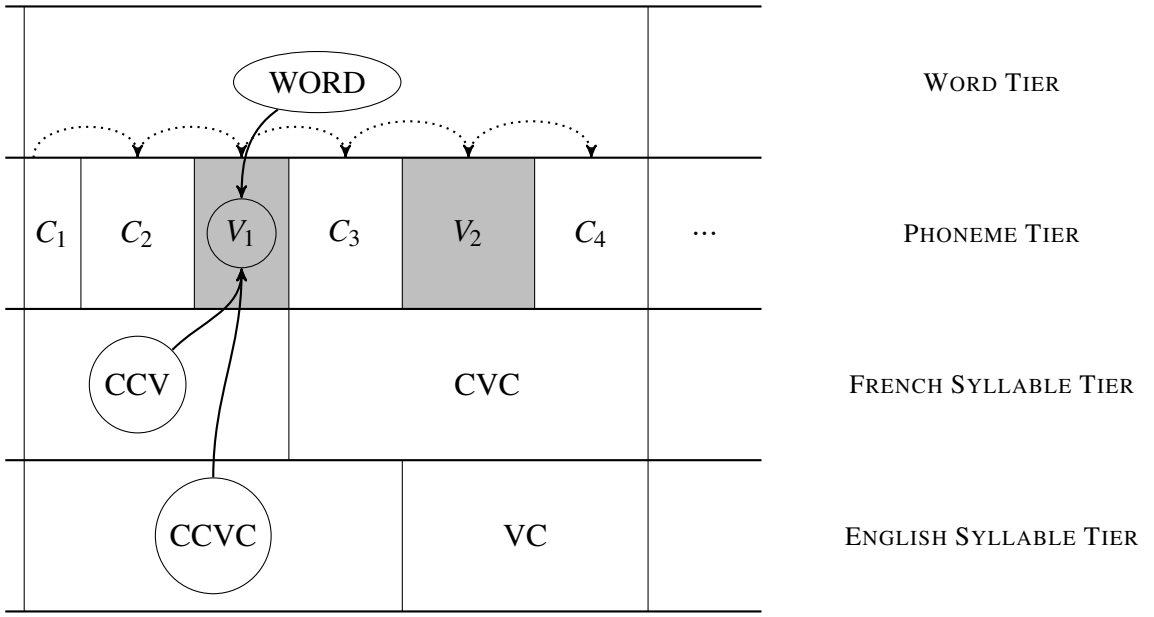


Fig. 1.7: Representation of the intervals scanned (dotted arrows) by PRAAT03 on the phonemic tier for either aligner (SPPAS or P2FA). When a vowel is parsed (grey rectangles), the corresponding labels of the other tiers are retrieved (circled nodes), and the aligner's dataframe is appended. C_x and V_x index consonants and vowels respectively.

time step was set to 0; the pitch-floor, to 75 Hz for men, 100 for women; the pitch ceiling was set to 300 Hz for men, 500 Hz for women. In formant analyses, the time step was also set to 0; the maximum number of formants per frame was 5; the maximum frequency was set to 5,500 Hz for women, and 5,000 Hz for men; the window length was kept at its default value of 0.025 second, with a pre-emphasis of 50 Hz.

This paragraph[7] focuses on the processes taking place in loops *(ii)* (for SPPAS) and *(iii)* (for P2FA), briefly described above, and presents some of the coding obstacles that were encountered. The procedure common to both aligners is symbolically represented in figure 1.7. Each interval on a given aligner's phonemic tier is parsed by PRAAT03. The moment when a given phoneme is scanned in either loop is defined as the Time Reference Point (henceforth, TRP). Data collection, *i.e.* the appending of the aligner's `.csv` dataframe, takes place when the currently parsed phoneme is a vowel. This entails that a lot of calculations are made on the first phoneme of a given word, *e.g.* its duration, its CELEX frequency, its

[7] What follows will probably suit the more technically inclined readers best.

numbers of syllables in French, in the LPD or in the aligner's dictionary. Syllabification in particular has to be carried out before the nucleus is parsed, even though syllable boundaries have not yet been created in the TextGrid: each vowel, *i.e.* each datapoint, must include the total number of syllables contained in the word it appears in (columns `LPDSC` and `SC`, *c.f.* section 1.2.3). As the word's phonemes get parsed, *i.e.* as the TRP moves forward from one

Table 1.4: Correspondences between the different transcription systems

IPA	LPD	SPPAS	P2FA
/æ/	&	{	AE*x*
/e/	e	E	EH*x*
/ɪ/	I	I	IH*x*
/ə/	@	@	AH0
/ɒ/	Q	A	AA*x*
/ʌ/	V	V	AH*x*
/ʊ/	U	U	UH*x*
/iː/	i:	i:	IY*x*
/uː/	u:	u	UW*x*
/ɜː/	œ:	3:r	ER*x*
/ɑː/	A:	A	AA*x*
/ɔː/	O:	O:	AO*x*
/aɪ/	aI	aI	AY*x*
/aʊ/	aU	aU	AW*x*
/eɪ/	eI	eI	EY*x*
/ɔɪ/	OI	OI	OY*x*
/ɪə/	I@	Ir	IH*x*R
/eə/	e@	Er	EH*x*R
/ʊə/	U@	Ur	UH*x*R
/ð/	D	D	DH
/θ/	T	T	TH
/ŋ/	N	N	NG
/ʃ/	S	S	SH
/tʃ/	tS	tS	CH
/ʒ/	Z	Z	ZH
/dʒ/	dZ	dZ	JH
/j/	j	j	Y

interval to the next, `PRAAT03` must identify whether the current phoneme is a coda, in order to create a boundary on the TextGrid. This process is made more complicated by two factors: *(i)* the differences in transcription systems, shown in table 1.4. These differences require checks for matches between transcription systems to be made, since English syllabification is indicated by the LPD transcription, but parsed phonemes are transcribed in either SAMPA or ARPAbet[8]. *(ii)* the variations in numbers of syllables, often caused by the diverging degrees of rhoticity between the British (LPD) and American (SPPAS & P2FA) dictionaries. This

[8] This issue is compounded by the fact that stress and syllabification in the LPD transcription are indicated by numbers ("1", "2" or "3" for primary, secondary and tertiary stress) or forward slashes ("/") respectively.

issue is discussed in greater detail in section 1.6.2. Code-wise, the key to align syllables correctly is to identify both syllabic nuclei and codas, regardless of how they are transcribed. This identification was harder with vowels than with consonants, but variations happened with the latter too: intervocalic /t/ and /d/ are flapped in the CMUPD used by SPPAS (*e.g.* "water" is transcribed /w O: 4 3:r/, /ˈwɔːɾɚ˞/, with "4" indicating a flap, but the /t/ is not transcribed as flapped in either the LPD or the CMUPD used by P2FA). With such hurdles, errors are likely - and happened. Section D.4 lists all the words whose *vowels*[9] feature different numbers of syllables between the LPD and the aligner's CMUPD, listed in the `LPDSC` and the `SC` columns (*c.f.* section 1.2.3): 1,450[10] monophthongs have syllable count mismatches in the SPPAS-generated dataset, with 1,318 monophthongs in the P2FA-generated dataset. These two figures account for 2.18% and 2.05% of the total number of monophthongs as aligned by SPPAS and P2FA respectively. More details about these mismatches can be found in section 1.6.

The next two sections present the resulting documents (*c.f.* figure 1.4) generated by `PRAAT03`: first, the multitier TextGrid, then the `.csv` spreadsheets based on the SPPAS and P2FA alignments.

1.2.3 The multitier TextGrid

This section lists the 17 tiers of the final TextGrids generated by `PRAAT03`. These final TextGrids are aligned with the original recordings of the students in each session. Figure 1.8 is a screenshot of a short section of one of the 102 final TextGrids after running SPPAS, P2FA and `PRAAT03`.

The 17 tiers respectively correspond to: *(i)* Transcription of the learner's speech / the short interval extracted with `PRAAT02`. *(ii)* Transcription of the native speaker's speech.

The string retrieved by `PRAAT03` from the dictionary therefore contained metaphonemic information which needed to be both stored (for syllable placement) and dispensed with (for phoneme parsing and matching).

[9] More precisely, monophthongs, since the focus of this work is on monophthongs.

[10] This number applies to monophthongs with durations longer than 0.03s., see section 2.1 for details).

(iii) Empty tier. Tiers 4 to 9 are P2FA-aligned mirrors of SPPAS-aligned tiers 10 to 15: *(iv)* P2FA-aligned phonemic tier. *(v)* P2FA-aligned word tier. *(vi)* P2FA-aligned French syllabic tier with manners of articulation (MOA) – based on the SPPAS algorithm for French syllables (V=vowel, O=occlusives, F=fricatives, N=Nasals, G=Glides). classes (C=consonants, V=vowel). *(vii)* P2FA-aligned English syllabic tier, using the same MOA-based transcription. *(viii)* LPD-based phonetic transcription of the current P2FA-aligned syllable. *(ix)* Stress of the current P2FA-aligned syllable (Primary/secondary/tertiary stress, unstressed or monosyllabic). *(x)* SPPAS-aligned phonemic tier. *(xi)* SPPAS-aligned word tier. *(xii)* SPPAS-aligned French syllabic tier with manners of articulation. *(xiii)* SPPAS-aligned English syllabic tier. *(xiv)* LPD-based phonetic transcription of the current SPPAS-aligned syllable. *(xv)* Stress of the current SPPAS-aligned syllable. *(xvi)* SPPAS-aligned consonantal and vocalic intervals (regardless of word and syllable boundaries); this tier is to calculate pairwise variability indices. *(xvii)* P2FA-aligned consonantal and vocalic intervals. The first, second and third tiers are the exact same as those from the main original TextGrid. The first tier is therefore the one that was adjusted manually, and the one from which the shorter sound files and TextGrids to be used by the two aligners were extracted. The numbers of intervals of this tier correspond to header `REFINT` in the final dataframes (*c.f.* section 1.2.4): with possible differences in transcriptions and syllable counts from one of the three dictionaries to another, these intervals are the only truly firm basis on which cross-comparisons between the two aligners can be made: they are a necessary (but more often than not, not sufficient) condition to the accurate retrieval of a given vowel in a given recording. Word and phonemic tiers (*i.e.* tiers 4 & 5 for P2FA, and 10 & 11 for SPPAS) are imported straight from the TextGrids created by the aligners, and merged into the main TextGrid. All other tiers are created by `PRAAT03`, although at different moments in the script (*c.f.* section 1.2.2). If we exclude tiers 2 & 3, which are neither affected by `PRAAT03` nor useful for the current purpose of our analyses, 1 tier out of 14 preexisted `PRAAT03` (tier 1), 3 were imported from SPPAS

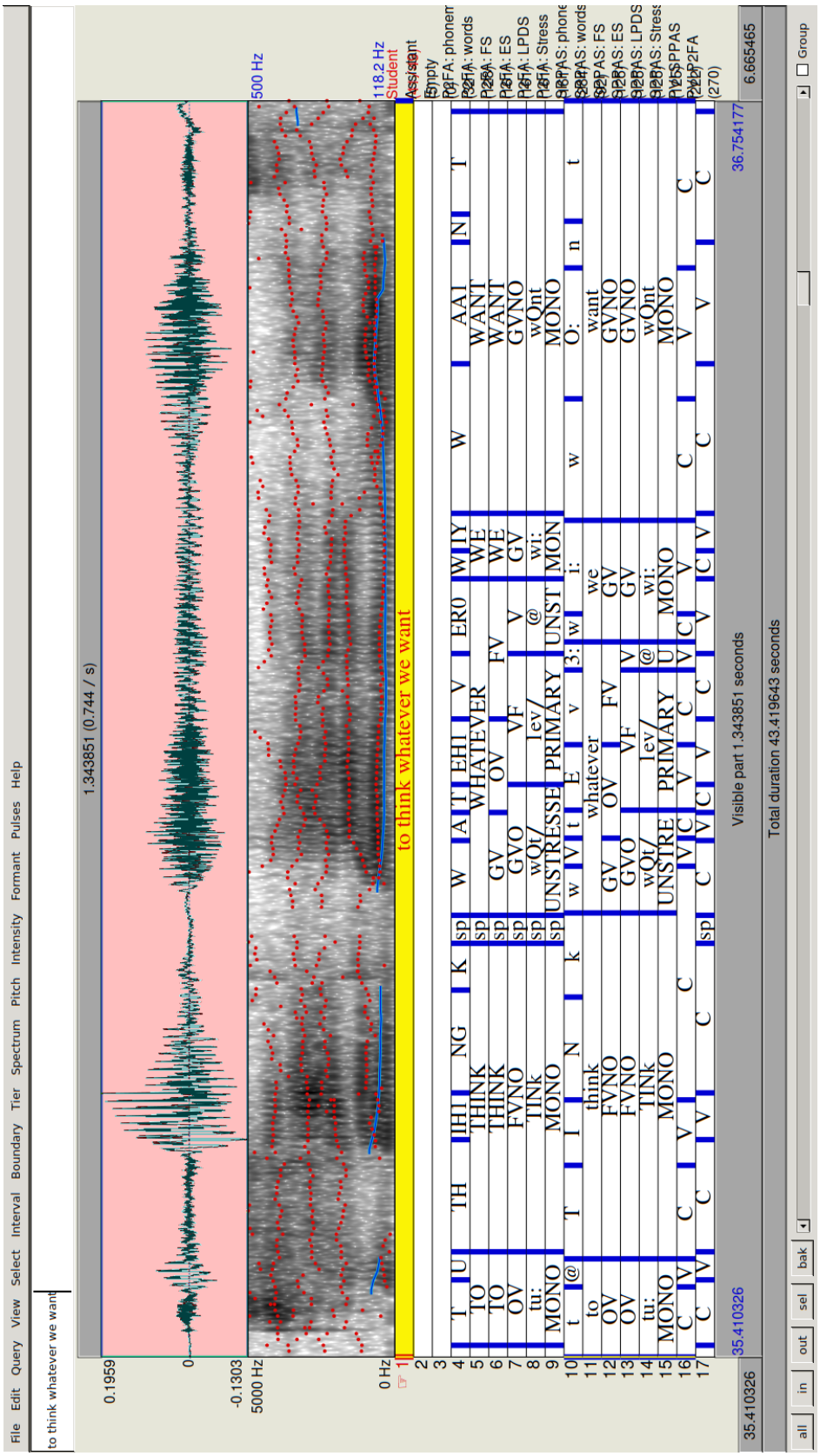


Fig. 1.8: Section of the final multitier TextGrid after running SPPAS, P2FA and PRAAT03.

(the phoneme tier, the word tier, the French syllable tier), 2 from P2FA (the phoneme and word tiers), and the rest were created by PRAAT03 based on the aligners' boundaries.

The reasons why these tiers were created are given in section 1.4, and the error-prone obstacles that were encountered are described in section 1.6.

1.2.4 The generated dataframes

Section D.1 gives the names of the 542 columns of the two final dataframes. There is one dataframe for each aligner, but the names of the columns are common to the two files, in order to make comparisons, and script-writing, easier. This section explains the content of each column.

The first dataframe is based on SPPAS-aligned data, which was extracted from tiers 10 to 15 of the multitier TextGrid (*c.f.* section 1.2.3 and figure 1.8). Likewise, the second dataframe is based on P2FA-aligned data, extracted from tiers 4 to 9 of the multitier TextGrid. Each dataframe is the end result of a dedicated loop in PRAAT03: the second loop in the case of SPPAS-aligned data, the third loop for P2FA-aligned data (*c.f.* section 1.2.2). The two dataframes are interchangeable: their only difference is the aligner used to extract the data, with all the changes this entails, especially with respect to the aligner-dependent transcription method. Because transcriptions vary (*c.f.* section 1.4 for details), the safest way to cross-reference data between the two dataframes (*i.e.* to ensure the correct retrieval of a given vowel in a given word in a given recording) is by using the very last column of both dataframes, REFINT, which indexes the interval number of the first tier in the multitier TextGrid: recall from section 1.2 that this number corresponds to the short TextGrids from which alignment was accomplished, and which the two aligners therefore have in common.

The first four columns (SPEAKER, SEX, SESSION, and WORD) are self-explanatory – WORD corresponds to the label of the interval of the SPPAS-aligned word tier at the current

TRP[11].The fifth column, CLXFREQ, gives the frequency of use of the word as mentioned in CELEX2 (Baayen et al. (1995)). Column LPDPRON gives the transcription of the LPD pronunciation of the word; the next column, PRON, the pronunciation of the word in the aligner's transcription method. LPDSC and SC indicate the word's number of syllables according to the transcription of the LPD or of the aligner respectively. Likewise, columns 10 & 11, LPDPHONEME and PHONEME, list the phoneme (*i.e.* the vowel) being scanned. The next six columns provide data on the syllable structure the phoneme is the nucleus of: columns 13 (ESYLLSTRUC), 15 (ECVSTRUC) & 17 (ESKELS) deal with English syllables, whereas columns 14 (FSYLLSTRUC), 16 (FCVSTRUC) & 18 (ESKELS) deal with French syllables. *x*SYLLSTRUC columns encode the manner of articulation of the syllable's phoneme in the following fashion: "O" for "occlusives", "F" for "fricatives", "N" for "nasals", "G" for "glides", "L" for "liquids" and "V" for vowels (*e.g.* "strings" would be encoded as "FOLVNF"). *x*CVSTRUC columns subsume all consonantal MoAs under a "C" category (*e.g.* "strings" would be encoded as "CCCVCC"). Finally, *x*SKELS[12] columns compounds all adjacents "Cs" into a single "C" (*e.g.* "strings" would be encoded as "CVC", *i.e.* a closed syllable). Whether the vowel's syllable is stressed is shown in STRESS; STRESS can be of values "MONO" for monosyllabic words, "PRIMARY" if the syllable carries the primary stress, "SECONDARY" and "TERTIARY" for secondary and tertiary stresses, or "UNSTRESSED". Column 20, PHONDUR, measures the vowel's duration. The next one, LOCINFILE, locates the beginning of the SPPAS vowel interval, while INTNB corresponds to the interval number of the phoneme's interval on the phoneme tier. INTENSITY gives the mean intensity of the SPPAS-aligned vowel from 10% of the vowel's duration to 90%. In the next eight columns, six give the preceding (PHONBEFORE) and succeeding (PHONAFTER) phonemes, along with their MoAs(BEF/AFTMOA), PoAs(BEF/AFTPOA) and voice features. Columns 25 & 30, PRECOART and POSTCOART address coarticulatory effects: if the vowel is preceded or succeeded: *(i)* by

[11] Recall that the TRP is the beginning of the interval on the phonemic tier which is currently scanned by PRAAT03 (*c.f.* section 1.2.2 for more details).

[12] "SKEL" stands for "skeletal".

a silence, these columns will return NONE; *(ii)* by a phoneme belonging to the same word, they will return INTERNAL; *(iii)* by a phoneme not belonging to the same word, EXTERNAL. The next column, EPENTHETIC, is an experimental set-up aimed at capturing learners' grunts and nasal fillers. If an interval on the SPPAS-aligned phonemic tier is empty and lasts more than 0.1 second, measurements are taken at $t_1 = t + 0.05$, $t_2 = t + 0.1$, $t_3 = t + 0.15$ and $t_4 = t + 0.2$. Pitch is then measured at these four time locations, along with intensity, which is averaged over from t_1 to t_4. EPENTHETIC, which returns a Boolean-like value ("YES" or "NO"), will return "YES" if mean intensity is superior to 40, and F0 readings exist at all four time points. The specific coding to obtain the value can be found in section C.2. TOTALDUR gives the total duration of the recording. The next 500 columns can be read in the following manner: *(i)* the aligner, SPPAS or P2FA, is specified; *(ii)* the formant (F0, F1, F2, F3 or F4) is specified; *(iii)* the number that follows gives the relative time location in the vowel where the formant was extracted. For example, SPPASF167 gives the F1 value 67% into the SPPAS-aligned vowel; P2FF34 gives the F3 value 4% into the P2FA-aligned vowel. BIRTHYEAR gives the learner's birthyear, and ESCDAYS indicates the number of days the learner reported spending in an English-speaking country. Finally, the last three columns give the duration of the word the vowel appears in (WD); the cumulative number of phonemes per parsed syllable, mostly a debugging feature, with NPW; and the interval number of the first tier in the multitier TextGrid (REFINT). After the execution of PRAAT03, an important change to the files is made: the value "–undefined–" is changed to "-1". "–undefined–" is the value assigned by PRAAT when the programme cannot perform a task. In this case, all "–undefined–" values. The reason why the "–undefined–" is changed to "-1" is because a numeric value is preferable to a string of character when importing the dataset to *R*: "-1" as a numeric value is consistent with the values of the columns in which the "–undefined–" values were (*c.f.* figure 1.9 and below. If these replacements are theoretically consistent, they also have valuable pragmatic consequences: with "-1" values, loading the datasets into an *R*

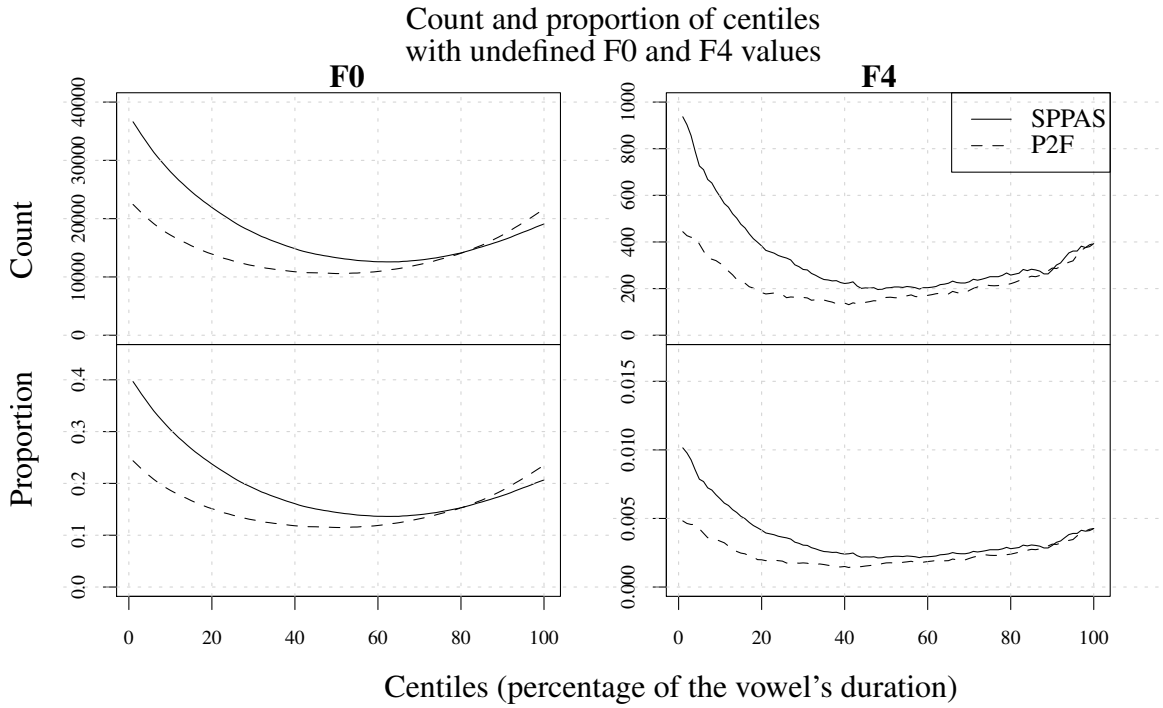


Fig. 1.9: Per-centile undefined F_0 and F_4 values. Top row: count; bottom row: proportion. Left panel: F_0; right panel: F_4.

environment using `fread` (Dowle & Srinivasan (2017)) considerably reduces computation times[13]. The total number of undefined values for SPPAS was 1,813,210; for P2F, 1,424,200. These numbers being so high in appearance, further study was required.

For SPPAS, 98.4% of those values (1,780,146) can be found in F_0 columns. Likewise, 98.2% (*i.e.* 1,401,993) of all undefined values in the P2FA dataset are F_0 values. The rest of the undefined values can be found predominantly among F_4 values (32,998 and 22,165 for SPPAS and P2FA respectively; undefined F_1,F_2 and F_3 values are insignificant: 24 and 14 each.). Figure 1.9 shows the distribution of undefined F_0 and F_4 values across the centiles of vocalic durations using both aligners. The bulk of the undefined values take place at the onset of the vowel, where the proportion of undefined values almost reaches 40% for SPPAS (24.2% for P2FA). Half-way through the vowel, at the 50$^{\text{th}}$ centile, the proportion drops down to 14.3% (11.5% for P2FA). At the end of the vowel, it rises back to 20.7% (24.3% for P2FA).

[13] On our computer, loading one dataset with "–undefined–" values takes more than two minutes – less than nine seconds with "-1" values....

Overall, 19.3% of all SPPAS F_0 values are undefined, against 15.2% for P2FA. It is not the purpose of this study, which will focus on F_1, F_2 and, to a lesser extent, F_3, to investigate the reasons why `PRAAT` returned so many undefined F_0 values. A cursory look at the data did not reveal any consistent patterns: no specific vowels or consonantal environments seem to be more likely to trigger undefined values[14].

Along the main `.csv` dataframe, `PRAAT03` created three other `.csv` files:

1. a duration file, in order to calculate the duration of the short, manually aligned TextGrid (*cf.* section 1.2.1 and figure 1.5);
2. a file containing the series of words pronounced by all speakers in all sessions, along with their location in each file (column `LOCINFILE`) for easy retrieval;
3. a file with the duration of the SPPAS- and P2FA-aligned consontal and vocalic intervals, in order to calculate the pairwise variability indices; each datapoint also features the number of phonemes pronounced in each interval, in order to calculate the Control/Compensation Index;

The headers of all these files can be found in section D.3. The same files were generated for the reading task in English (*The Selfish Giant*) and in French (with no P2FA-aligned data in this case, since P2FA does not support French). The same procedures of alignment and data extraction was applied to three subcorpora, detailed in the next section.

1.3 Sub-corpora

The workflow described in section 1.2 was applied to three subcorpora:

1. a list of words to be read, which can be found, along with the given instructions, in section A.2.1;

[14] An immediate assumption to explain the phenomenon, especially at the onsets of vowels, would be that missing values are due to the misalignments of phonemic boundaries after initial, potentially aspired plosives: the ensuing late Voice Onset Time (VOT) could have been interpreted as part of the vowel already, rather than part of the plosive itself. Unfortunately, this does not seem to be supported by the data. But as said above - this was only looked at in a cursory way.

2. a text in French, also to be read: the translation of Oscar Wilde's *the Selfish Giant* (1888), which can be found in section A.2.2;
3. a set of recordings from native speakers, taken from various podcasts.

The rationale behind these subcorpora is twofold: first, using native speakers' recordings will provide a basis to assess the quality of the workflow. Should implausible formant values be found for vowels pronounced by native speakers, then they cannot be attributed to the French learners' potential errors, but rather to the procedure itself. Second, these subcorpora provide bases for useful comparisons: the list of English words theoretically tests the students' phonological knowledge of the vocalic system of English (in Chomskyan terms, their competence); the text in French provides the students' base formant values of their native system. This is particularly useful to investigate the differences between their native /i/ and /u/ on the one hand, and their realizations of /iː/-/ɪ/ and /uː/-/ʊ/ on the other; finally, the English speakers' recordings will provide the basis for comparing the values of natives and learners in spontaneous connected speech. These values will be prefered over those found in various studies (*c.f. e.g.* Ferragne & Pellegrino (2010), Hillenbrand et al. (1995)) with different eliciting methods (mostly recorded /hVd/ words). As the recordings of

	List of words	**French Text**	**Native Speakers**
Number of speakers	13	13	15
Number of vowels (SPPAS)	1750	2902	4273
Number of monophthongs (SPPAS)	1310	–	3380
Number of vowels (P2FA)	1750	–	4283
Number of monophthongs (P2FA)	1303	–	3314

Table 1.5: Summary of the subcorpora data

the list of words and the text in French were made during Session 4, the thirteen speakers are those who took part in all four sessions (*c.f.* table 1.1). 83,019 substitutions of "–undefined–", *i.e.* 6% of all formant cells across all centiles, were made in the SPPAS-aligned data of the native speakers' subcorpus; in the P2FA-aligned subcorpus, 63,328 substitutions, *i.e.*

4.6%, were made. For the reading subcorpus, the numbers of subsitutions are 29,823 (5.7%) and 16,740 (3.2%) in the SPPAS-aligned and P2FA-aligned data respectively. In the French reading corpus, 3,426 "–undefined–" cells, amounting to 1.36% of all cells, were replaced. Finally, the breakdown of accents in the native speakers' subcorpus, later on refered to as the Native Speakers Subcorpus (NSS), the following: four male and female British speakers, one male Irish speaker, one female Scottish speaker, one female American speaker and four male American speakers. These accents were encoded in the `SPEAKER` column under the following labels respectively: `NATBRIT`, `NATIREL`, `NATSCOT` and `NATUSMI`. In terms of collected data, the NSS and the reading lists in English feature the same number of datasets (one SPPAS-aligned and one P2FA-aligned) and columns as the main learner corpus. Only the French reading corpus is different, since it has no corresponding P2FA-aligned tiers in the `TextGrids`, and no P2FA-based dataset, since P2FA does not handle French. Also, the formant values were only retrieved at every decile of the vowels' durations. The columns of the dataset specific to the French reading corpus can be found in section D.2.

1.4 Theoretical justifications

This section provides the technical, then theoretical and linguistic, justifications for the type of data that were collected.

One first remark must be made about the format of the data structure. There are two main reasons for the choice of a univariate (long) format, as opposed to a multivariate (wide) format. The first reason lies in the technical impossibility of obtaining a univariate format, because the number of occurrences of each phoneme in each session is highly variable. This state of affairs is fortunate: coding-wise, the way `PRAAT03` parses the aligned phonemes in the `TextGrids` harmoniously corresponds to the requirements of a univariate format. This issue is also compounded by the extraction of (among other variables) five different formant values at every centile of a vowel's duration. Such a number of variables would make a

wide format unreadable. The second reason is less a matter of happenstance: Linear Mixed Effects Regression (LMER) analyses, which will be carried out in section 3.4, require the data structure to be in long format (Long (2012)).

Several columns in the data frame presented in Appendix D serve as control entries to ensure that the data collected for each aligner is correct. This is the case for instance of `P2FWORD`, `P2FPHONATTRP`, or `P2FMATCH`: since `PRAAT03` parses the SPPAS-aligned phonemes, controls were necessary to ensure that P2FA-aligned phonemes were also correctly retrieved. `LPDPHON` and `LPDSYLLPRON` are also control entries: `LPDPHON` is the output of a complex `PRAAT03` procedure that accesses a `.txt` file version of the LPD. It is from `UKPRON` that the theoretical syllable structure (`ESYLLSTRUC`) is derived. Likewise, in order to narrow down the source of potential errors during the (lengthy) debugging phase, `LPDSYLLPRON` made it possible to ensure that the syllable-aligning procedure in `PRAAT03`, which led to the creation of Tiers 12 & 13 in the final TextGrid, was correct. On a side-note, the designing of this procedure in the development stage of `PRAAT03` was particularly excruciating, and required that control entries such as `LPDSYLLPRON` be devised. The main reason why aligning syllables based on the LPD pronunciation was so difficult, and therefore error-prone, is that, as described above, SPPAS uses SAMPA as a transcription alphabet, and American as the base language, whereas the LPD uses the International Phonetic Alphabet (IPA), and Southern British English as the base language. This difference entailed that correspondences between phonemes, on which correct syllable alignment crucially relies, were far from obvious: rhoticity (as seen in figure 1.8 with "favourite"), conventions for /iː/, /ɪ/ or /i/, /t/- or /d/-flaps are some of the obstacles that the procedure had to overcome.

Why, then, take the trouble to align syllables? One of the main reasons is that phonotactics in both French and English exerts influence on segmental realizations. One way to explain the role of syllables in French is to adopt generativist terminology. It can be said that French features three archiphonemes, /E/, /EU/ and /O/, which are underspecified for height. Height

specification depends on whether the syllable has a coda or not (*i.e.* is closed or open): thus /E/ will *tend* to be pronounced /e/ in open syllables, as in "cocher" (/koʃe/), /ɛ/ in closed syllables ("cochère", /koʃɛʀ/); likewise with /EU/ and /O/, as in "ceux" (/sø/) and "seul" (/sœl/), and "saut" (/so/) and "sol" (/sɔl/) respectively. Of course, numerous exceptions exist (*e.g.* "fée" /fe/ *vs.* "fait" /fɛ/, "jeûne" /ʒøn/ *vs.* "jeune" /ʒœn/ or "saute" /sot/ *vs.* "sotte" /sɔt/, to name but a few), but the principle holds in other numerous instances as well. Likewise, although in an arguably stricter fashion, in English, lax vowels only appear in closed syllables. One question therefore arises, which justified collecting the theoretical English syllabic structure (`ESYLLSTRUC`) and the algorithm-based, SPPAS-generated, French syllabic structure(`FSYLLSTRUC`): are instances of English /e/, or /ɔː/, possibly even /ɒ/, influenced by the syllabic structure in which they appear? This non-trivial issue will be addressed in future research. The difference between columns `FSYLLSTRUC`, `FCVSTRUC` and `FSKELS`[15] lies in the granularity of consonantal labelling: MoA-based labels from `FSYLL` such as "O" for "Occlusive", "F" for "Fricative", "N" for "Nasal", etc., are all subsumed to "C" (for "Consonant") in `FCVSTRUC`. Consonantal clusters disappear in `FSKELS`: a CCCVCC syllabic structure (as in "straps") in `FCVSTRUC` will be labelled "CVC" in `FSKELS`. The reason why such simplifications were made lies in the following theoretical questions: do all consonants affect vocalic realizations in the same manner? Likewise, to what extent do consonant clusters affect these realizations? The answer to the first question is arguably trivially negative (*cf. e.g.* Hillenbrand et al. (2001) for the influence of plosives and /h/ in initial and final syllable position on English vowels), and therefore entails that the manners and places of articulation, along with the voicing feature, of preceding and suceeding consonants had to be listed in order to control their influences on vocalic realizations. This is the *raison d'être* of columns `PHONBEFORE` (*i.e.* the preceding phoneme), `BEFMOA` (the MoA of the phoneme before the vowel), `BEFPOA` (the PoA of the phoneme before), `BEFVOICE` (the

[15] The same applies to `ESYLL`, `ECVSTRUC` and `ESKELS`.

voicing feature of the phoneme before), PHONAFTER (the succeeding phoneme)[16], AFTMOA, (the MoA of the phoneme after), AFTPOA (the POA of the phoneme after) and AFTVOICE (the voicing feature of the phoneme after). To answer the second question, bases for comparisons needed to be provided: if different statistical correlations can be found, for a given type of vowel, between that vowel's formant values and the three types of consonant labelling, then the labelling method whose model displays the lowest deviance should be kept. The influence of consonant clusters on vocalic formant signals will be addressed indirectly in chapter 3 where vowel-inherent spectral change (section 3.2) and discrete cosine transforms (section 3.5) will be investigated.

This section having detailed the theoretical justifications underlying the design of the extracting scripts and the structure of the generated dataset, a crucial question now arises: what treatment were context-dependent labels given? The answer is provided in the next section.

1.5 Vowel reductions and weak forms

This section explains the achievements, compromises and shortcomings of labelling phonemes likely to undergo vowel reduction. The main issue that was encountered with labelling weak forms is that weak forms are caused by two different sorts of context. The first type of context is phonological: this is the case of instances of "the" pronounced /ði/ when preceding a vowel, and /ðə/ when preceding a consonant. By the same logic "to" is usually pronounced /tu/ before a vowel, /tə/ before a consonant. These weak forms are easy to hard-code into a PRAAT script that can retrieve the content of the succeeding interval, as described in figure 1.7. It is technically impossible, however, to infer weak forms when these are induced by syntactic or semantic conditions. Cases such as the reduction of "had"

[16] The plan is also potentially to investigate the equation of locus for the vowels of French non-native speakers. Do we observe the same coarticulation as for natives? The results in datasets could be used to answer this question in the future.

from /hæd/ to /həd/ in sentences where "had" serves as an auxiliary, or cases of stranded prepositions, cannot be predicted: the workflow as it stands is blind to syntax and semantics. No attempt at overcoming this hurdle was made. An ambitious venue of research would be to merge syntactic tree-taggers into `PRAAT`-generated `TextGrids` in order to predict the ideal pronunciation of vowels likely to adopt weak forms. The following paragraphs study in detail the transcriptions of the vocalic nuclei of two very frequent words in the corpus, "the" and "to".

For now, when not coded in `PRAAT`, phonologically induced weak forms can also be deducted in *R* by replacing the strings in `LPDPHONEME` when conditions are met in `PHONAFTER`. This solution is much handier than hard-coding in `PRAAT`, which would require a lengthy generation of all the grids and datasets. Besides, it enables interested researchers to amend the mistakes of programmers: if weak forms of "the" have indeed been properly coded in `PRAAT03`, this is not the case of "to", whose weak forms must therefore be determined in *R* scripts[17].

Now it looks as if SPPAS attempted to implement weak forms in its computations. The reason why that may be the case is that both "the" and "to" are ascribed different transcriptions (these transcriptions can be found in the `PHONEME` column in the SPPAS dataset). "The" is transcribed as either /tə/, /tʌ/ (*sic*) or /tiː/ (*sic*); "to" is at times transcribed /θə/, at times /θɪ/ (*sic*), at times /θu/ (*sic*). What logic did the algorithm follow? It is very unlikely that the succeeding phonemes were taken into account. Table 1.6 gives the number of times "the" (in the left table) and "to" (in the right table) were given one of their three respective transcriptions for each phonemes succeeding them[18]. The pronunciation of the vocalic nuclei of the 2,874 occurrences of "the" can be broken down as follows: 1,995 /ə/, 325 /ʌ/ and

[17] This option is not only more convenient, but also much simpler than in `PRAAT`. Two lines of code suffice.

[18] As the first row of the two tables shows, it would be more accurate to speak of succeeding *intervals*: these phonemes are technically the retrieved strings of the intervals following the parsed vowels. That interval may occasionally be empty. In the P2FA-aligned data, empty intervals are also sometimes labelled "sp". P2FA-aligned being somewhat shorter than SPPAS-aligned ones, P2FA will detect "sp" intervals where SPPAS will see a continuing vowel. This explains the discrepancies in numbers for the empty or "sp"-labelled rows between table 1.6 and table 1.7.

Table 1.6: Distribution of the pronunciations of the SPPAS-aligned nuclei in "the" (left table) and "to" (right table) against the succeeding phoneme (transcriptions are in SAMPA).

	@	I	u
	8	327	216
3:r	1	1	0
@	3	15	14
@U	0	2	0
A	1	0	0
D	5	121	53
E	0	6	4
I	1	20	5
O:	2	4	3
S	0	9	5
T	2	10	8
V	3	4	9
aI	0	7	2
b	42	40	46
d	75	14	37
dZ	3	4	3
eI	2	4	2
f	14	7	46
g	108	17	81
h	7	25	40
i	1	22	13
i:	0	7	2
j	0	14	4
k	17	16	12
l	53	7	24
m	40	42	24
n	11	3	6
p	15	28	26
r	7	27	39
s	126	26	27
t	44	49	57
tS	1	6	8
v	2	4	6
w	23	4	41
{	0	3	6

	@	V	i:
	257	57	80
3:r	3	0	4
@	9	0	21
@U	15	0	18
A	6	1	10
D	55	29	29
E	12	0	28
I	7	3	8
O:	3	0	7
S	10	1	5
T	35	5	7
V	11	0	27
aI	7	0	7
aU	0	0	1
b	118	9	8
d	27	5	5
dZ	4	1	3
eI	1	0	2
f	176	19	17
g	31	2	11
h	32	8	9
i	9	3	11
i:	3	0	4
j	25	1	53
k	149	8	39
l	89	53	19
m	115	8	5
n	24	8	9
p	149	11	9
r	106	33	5
s	288	23	38
t	58	19	38
tS	15	0	1
u	1	0	0
v	4	2	3
w	127	15	6
z	1	0	0
{	13	0	7

554 /iː/; the 2,382 nuclei of "to" are transcribed 617 times as /ə/, 895 times as /ɪ/ and 870 times as /u/. However, the table shows no clear pattern indicating that the transcriptions

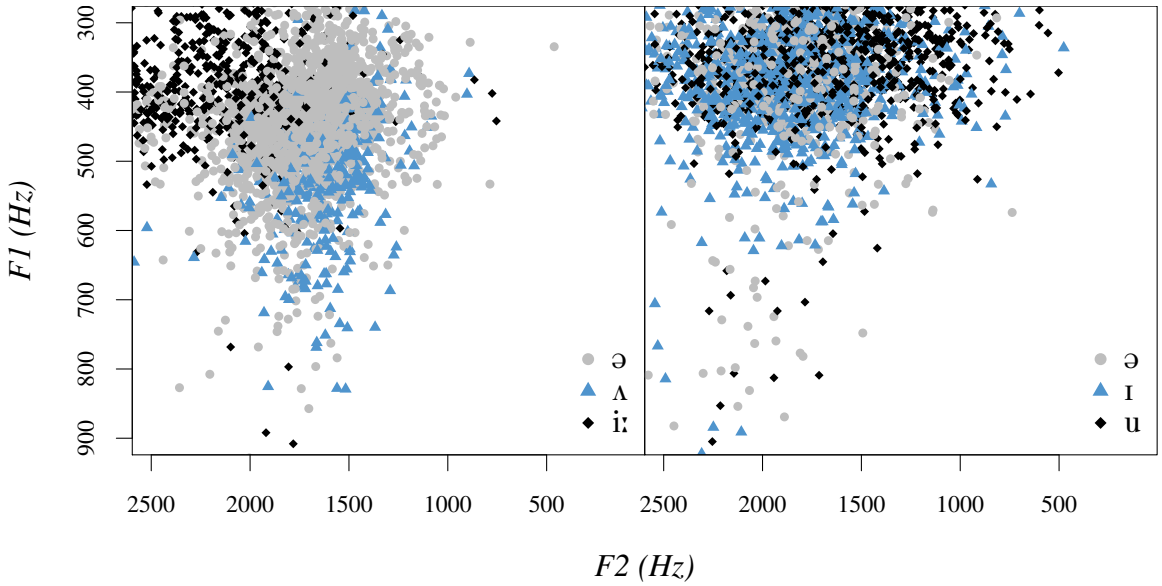


Fig. 1.10: F_1 and F_2 values of the SPPAS-aligned vocalic nuclei of "the" (left plot) and "to" (right plot).

for either words are selected on the basis of the succeeding phonemes, thereby emulating

the phonological rules of vowel reduction. Considering that at run-time, SPPAS has access to the acoustic data from PRAAT, another solution to explain the choices of transcriptions is to have a look at the distribution of those transcriptions as a function of acoustic data. The simplest parameters are mid-temporal F_1 and F_2 values. Figure 1.10 plots the F_1 values in Hertz of the nuclei of "the" (in the left panel) and of "to" (in the right panel) against their F_2 values, both taken in the middle of the vowels' durations. Graphic examination reveals a pattern for "the": the locations in the vocalic trapezoid of the nuclei is consistent with their transcripions. Instances of /iː/ exoectedly feature high F_2 values and low F_1 values; /ə/-transcribed nuclei have higher F_1 values and lower F_2 values than the previous ones; and, also expectedly, tokens with /ʌ/ have roughly the same F_2 values, with however a greater aperture of the mouth than their /ə/ counterparts, and therefore higher F_1 values. The picture is completely different for the nuclei of "to", whose transcriptions look more chaotic. No consistent pattern is apparent. In order to confirm these observations, a linear discriminant analysis (henceforth, LDA) using *R* package MASS (Venables & Ripley (2002)) was carried out. Although the dependent variable, *i.e.* the transcription of the nuclei, is categorical, a simpler logistic regression to classify the instances was not possible because the variable has more than two categories. A leave-one-out cross-validation was implemented simply by setting the CV argument to true in the lda command. The results bear out the insights provided by the observation of the plots: the LDA for "the" returns a classification rate of 75.6%, against a mere 47.8% for "to". These findings are arguably all the more robust as the distribution of the three categories are more imbalanced in the case of "the", with a clear bias in favour of instances of /ə/ (69.4% of all occurrences for /ə/, against 11.3% and 19.3% for /ʌ/ and /iː/ respectively); the proportions are more even with "to": 27.9%, 37.6% and 36.5% for /ə/, /ɪ/ and /u/ respectively. It goes without saying that this does not mean that the transcriptions of "to" as listed in the PHONEME column of the SPPAS-based dataset are random: all that can be concluded is that these transcriptions are not selected on the basis of

mid-temporal F_1 and F_2 values. Neither does it mean that the apparent logic of transcriptions of "the" is one that takes into account the values of F_1 and F_2.

Do weak forms or alternative transcriptions exist in the P2FA-aligned data? The situation is slightly different in that for instance, there is no variation in the transcriptions of "to", whose nucleus is transcribed as `UW1`, the ARPAbet equivalent of a stressed /ʊ/. Transcriptions of "the" however follow the same patterns as the SPPAS-aligned cases. The possible transcriptions are `AH0` (/ə/), `AH1` (/ʌ/) and `IY0` (unstressed /ɪ/). The 2,778 occurrences of "the" are respectively distributed across those three categories as follows: 66.4%, 14.6% and 19%. Just as in the cases of SPPAS-aligned instances of "the", no apparent pattern emerges

Table 1.7: Distribution of the pronunciations of the P2FA-aligned nuclei in "the" (left table) and "to" (right table) against the succeeding phoneme (transcriptions are in ARPAbet).

	AH0	AH1	IY0
	79	104	86
AA1	2	0	11
AA2	2	0	1
AE1	3	0	9
AE2	0	0	3
AH0	3	0	21
AH1	2	1	31
AO1	1	1	6
AY0	1	0	7
AY1	0	0	3
B	99	3	8
CH	12	0	3
D	24	3	2
DH	12	9	9
EH1	2	0	34
ER0	2	1	2
ER1	0	0	2
EY0	1	0	0
EY1	0	0	1
F	152	9	2
G	24	1	12
HH	21	3	9
IH0	4	0	6
IH1	2	3	9
IH2	2	0	4
IY1	0	1	3
JH	4	0	2
K	115	7	27
L	104	37	11
M	107	8	3
N	36	2	2
OW1	3	1	28
P	123	7	3
R	88	31	6
S	260	12	6
SH	11	0	0
T	72	5	11
TH	31	2	1
V	5	1	1
W	119	11	3
Y	22	2	49
Z	1	0	0
sp	293	140	92

when inspecting the phonemes succeeding P2FA-aligned nuclei of "the". These succeeding phonemes are listed in table 1.7. When looking at the corresponding mid-temporal F_1 and F_2 values, however, the same correlation appears between the label and the formant values,

as figure 1.11 shows. The same LDA run on the P2FA-aligned instances of “the”, with

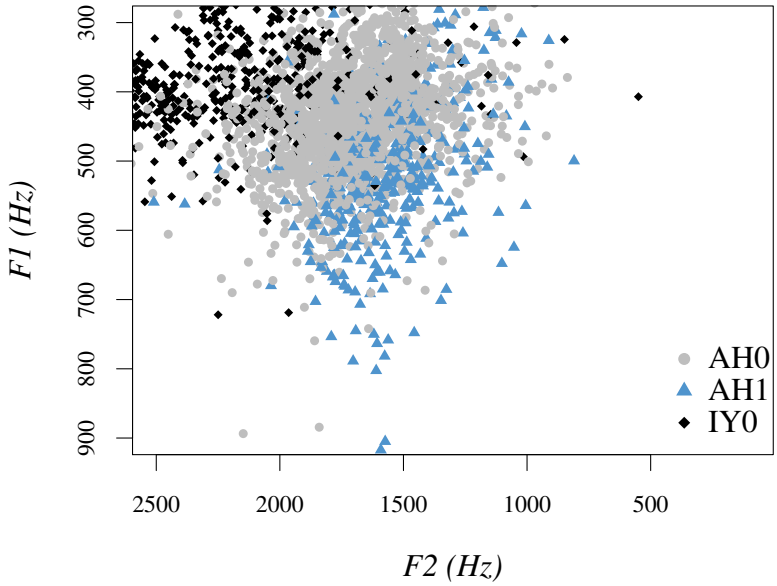


Fig. 1.11: F_1 and F_2 values of the P2FA-aligned vocalic nuclei of “the”.

the transcription given in the `PHONEME` column as the categorical dependent variable, and raw mid-temporal F_1 and F_2 values as the continuous predictors, returns a rate of correct classification of 77.4% – slightly higher than in the cases of SPPAS-aligned “the”.

To conclude this brief section on vowel reductions and weak forms, let it be reminded that no syntax- or semantics- induced vowel reductions have been implemented in the corpus. Changes caused by the nature of the following phones, whether consonants or vowels, were hard-coded by SPPAS and P2FA in the case of “the”, but not in the case of “to”. It was here found that the chosen transcriptions for “the” are consistent with formantic values; for this word, the strings listed in the `PHONEME` column, which was generated by the aligners themselves, reflect an acoustic reality, but not a normative, linguistically desirable target. These desirable targets are defined as functions of the nature of the following sounds, a feature that was implemented in `PRAAT03` for “the”, and “the” only. These desirable targets are listed in the `PRAAT03`-generated `LPDPHONEME` column. Expanding the same logic of taking the succeeding sounds into account to words such as “to” is nonetheless easily feasible in *R* scripts. Future research should endeavour to explain the variations in transcriptions by SPPAS for other words such as “was”, for instance, but especially to make syntactic and

semantic information available either at alignment run-time, or at extraction run-time. Only then will vowel reductions and weak forms be properly implemented.

1.6 Syllabification: technical details

This section summarizes the issues and challenges raised by syllabification while building the corpus. The previous section and figure 1.8 have already briefly presented the issues at hand from the point of view of the output. Many obstacles had to be overcome in order to create accurate syllabic tiers: the differences in encoding alphabets and base accents between the transcription systems and the variations within those transcription systems turned out to be quite formidable coding challenges. How these hurdles were overcome, along with an assessment of the overall performance of the solutions adopted, is presented in section 1.6.1. Section 1.6.2 details the causes of the remaining errors.

1.6.1 Coding principles and challenges for syllabification

Screenshot 1.8 encapsulates all the issues of the project. One first issue is that of the transcription convention each aligner chooses. Although both aligners use the Carnegie Mellon University Pronouncing Dictionary (CMUPD, Weide (1994)), P2FA uses an ARPAbet symbol set, while SPPAS uses SAMPA[19]. For the reader's convenience, the first three lines of table 1.4 have been reproduced below to illustrate the differences in the sets coding the vowels. Machine-readable symbol sets have been invented to represent IPA-like transcriptions with characters. Two types of problems arose: the existence of symbols that corresponded to supra-segmental features (*i.e.* stress marks and syllable divisions); the choice of character strings may not be the same from one conventional system to the other. Similar issues are described in the documentation of the CELEX (Baayen et al. (1995)). In the CELEX, they

[19] More precisely, SPPAS's developper Brigitte Bigi converted the CMUPD to X-SAMPA.

Table 1.8: Correspondences between the different transcription systems (shortened version)

IPA	LPD	SAMPA (SPPAS)	ARPAbet (P2FA)
/æ/	&	{	AE*x*
/e/	e	E	EH*x*
/ɪ/	I	I	IH*x*

created their own system to make sure they only had one ASCII character per representation of phoneme. `PRAAT03` deals with a much more complex situation where: *(i)* suprasegmental marks may correspond to the initial ASCII character before the representation of a phoneme; *(ii)* the representation of a phoneme corresponds to a variable number of ASCII characters. *(i)* shows that this is not a bijective phoneme-to-ASCII character mapping. The script was programmed to capture the segmental transcriptions between the suprasegmental marks to align the LPD transcriptions with the various word tokens[20].

(ii) is far from being a trivial issue: although both aligners are based on the CMUPD, there is no bijective relationship between the ARPAbet and SAMPA transcriptions because reduced vowels may vary. For instance, the word "civilization" is transcribed "sIv@lizeIS@n" in SAMPA (*i.e.* /ˌsɪvəliˈzeɪʃən/, and "S IH2 V AH0 L AH0 Z EY1 SH AH0 N" in ARPAbet (*i.e.* /ˌsɪvələˈzeɪʃənˈ/). When the TRP is at the /ə/ of the second syllable on the SPPAS-phoneme tier, the only way to make sure that the corresponding P2FA-aligned phoneme is the one in the second syllable, as opposed to that of the third syllable, is by keeping track of syllable counts and syllable structures (as given by the LPD). Besides, it is worth noting that in this example (as in many other instances), SPPAS-aligned /i/ is matched in the third syllable by P2FA-aligned /ə/: correspondences between the two aligners based on phonemes only are therefore insufficient.

In order to overcome those obstacles, multiple checks had to be used, loosely inferred from linguistic principles. One first check was to determine at the beginning of every word

[20] This roughly corresponds to procedure `GetWellsPron` from line 2922 of `PRAAT03` onwards.

how many syllables that word features in the three systems. This was carried out by retrieving the three transcriptions, and counting the number of vocalic nuclei in the strings. Because once again the transcription systems are different and one same phoneme can be transcribed with different numbers of ASCII characters, three procedures were designed – one for each transcription system. The procedures store all the different types of possible vowels that may occur within their respective transcription method. Once the syllable counts were calculated, the exact location of syllable boundaries had to be determined. Once again, the key issue was to come up with a way to guarantee that the phoneme at the TRP would be safely identified across the three systems. Multiple checks had to be implemented because of the differences in transcriptions of rhotic vowels – the LPD giving non-rhotic transcriptions, SPPAS and P2FA giving rhotic ones but with variations within their own dictionaries – but also because of variously encoded phenomena such as flapping. Flapping was problematic because it undermines the intuition that correspondences for consonants are more easily determined than for their vocalic counterpart. This intuition quickly turned out to be wrong: not only did the case of the transcription matter (for instance, "t" in SAMPA is "T" in ARPAbet, while "T" in SAMPA is "TH" in ARPAbet), but /ɾ/, transcribed "4" in SAMPA, can either correspond to a "d" or "t" in the LPD – *i.e.* a "T" or "D" in ARPAbet. Several constructs were therefore designed in `PRAAT03`: metaphonemes, metaonsets and metacodas. These constructs served the purpose of unifying all possible transcriptions under a single label which could then be cross-identified regardless of the system adopted. A metaphoneme in the script was fundamentally a list of attributes attached to transcription-dependent phonemes based on their MoAs, PoAs and voice features. At the beginning of every word, when syllable boundaries are established, the metaphonemic features of both codas and onsets are stored so that when the phoneme at the TRP, and the phoneme after it in the next interval, have metaphonemic properties that match both those of the pre-stored metacoda and metaonset

respectively, a syllable boundary was added[21]. Because of the multiple issues caused by the variations in transcriptions, another check was added: the number of phonemes of every word was also stored as soon as one began. A counter kept track of the number of phonemes parsed in every transcription system, and the number of phonemes before a syllable boundary was also calculated beforehand. The combination of all these checks made for a relatively error-proof system as shown in table 1.9. The table gives the number of *phonemes* featuring

Table 1.9: Table of skeletal syllabic structures for SPPAS (top table) and P2FA (bottom row)

	C	CV	CVC	CVCV	CVCVC	CVVC	V	VC	VVC	
	1	1	19802	35027	3	2	11	13236	19158	29	

	CV	CVC	CVCV	CVCVC	CVCVVC	CVV	CVVC	V	VC	VVC	VVCVCVC
1	19117	32106	4	6	1	1	11	13670	20153	20	11

one of the listed skeletal syllabic structures, as retrieved from the ESKELS column. This means that the number of syllables and words featuring a particular syllabic structure is lower. Of particular interest here are the structures containing two syllabic nuclei. In the SPPAS-aligned data, the affected words are the following: “bodies” and “ladies” (CVCV); “schedules” and “there’ll” (CVCVC); “ideas”, “las” and “Korean” (CVVC); and “different”, “several”, “favorite” (US spelling), “favourite” (UK spelling) and “Orpheus” (VVC). These last cases, except for “Orpheus”, are all instances of issues caused by rhoticity and the status of /r/ (coda or onset or part of the nucleus) in the transcriptions. For the P2FA-aligned data, the words containing syllables with two nuclei are the following: “bodies”, “Glasgow” and “ladies” (CVCV); “question”, “learn” and “there’ll” (CVCVC); “consensual” (CVCVVC); “jewelry” (CVV); “several”, “different”, “favorite”, “difference” and “Orpheus” (VVC); and “education” (VVCVCVCVC). These anomalies amount to 45 datapoints in the SPPAS-aligned data, out of 87,269 datapoints (vowels with durations shorter than 0.03 were excluded from that grand total), *i.e.* 0.05% of all datapoints; and 54 out of 85,100 datapoints in the

[21] Adding an interval on any given tier is always a delicate operation in PRAAT03, since if an interval boundary already exists where one is added, the script crashes.

P2F-aligned data, *i.e.* 0.06% of the entire dataset. Bearing in mind that these numbers count instances of vowels, these error rates were deemed acceptable and no further improvement of the algorithms were undertaken after reaching those targets[22].

Having attempted to explain the principles and challenges underlying syllabification, the following subsection deals with the errors that remain and attempts to explain their causes.

1.6.2 Errors in syllabification

In order to check the consistency of the syllable divisions in the different methods, two variables associated to each token were generated. These two variables indicate the number of syllables of the LPD transcription and of the aligner transcription respectively. In each dataframe, they are listed under columns n°8 `LPDSC` and n°9 `SC`. An *R* script[23] retrieved the number of discrepancies in syllable division. Somehow that discrepancies exist is not exclusively attributable to the transcription systems used by the aligners, and should come as no surprise. Syllabification theories themselves are complex, and at times contradictory. For instance, the two main reference pronouncing dictionaries used in Europe, the LPD and the English Pronouncing Dictionary (EPD, Jones et al. (2011)), have conflicting conceptions and transcriptions for syllable divisions. The LPD thus favours the MaxCoda rule for stressed syllables (Wells (1990)), *i.e.* as many consonants as possible are attached to codas; whereas the EPD follows the Maximum Onset Principle (MOP) (*c.f.* Ballier (2014)) – the exact other way round. A word like "*country*" will be syllabified /kʌntr.i/ in the LPD, /kʌn.tri/ in the EPD. French speakers are more likely to adopt an EPD-like MOP syllable division (*c.f.* Dell (1973)). However, regardless of these theoretical tenets followed by the two dictionaries, the fact remains that English has more CVC patterns (Cutler et al. 1995, Levelt et al. (1999)).

[22] Of course, there is room for improvement. The cases of "bodies" and "ladies" for instance are errors most likely caused by improper plural suffixation. Why the suffixation procedure failed in those cases is itself most likely due to the flapping of <d> before the suffix. Assuming this explanation is correct, amending the suffixation procedure for the two occurrences of "bodies" and the single occurrence of "ladies" was not cost-effective – if this work was to ever see the light of day.

[23] This script can be found in Appendix C.5.

The list of words with mismatches in syllable counts can be found in section D.4. These mismatches are those caused by a discrepancy between the phoneme counts as calculated from the LPD transcription on the one hand (listed in the `LPDSC` column) , and as calculated from the aligner's transcription on the other (listed in the `SC` column). In total, there 2.13% of all datapoints feature a syllabic mismatch in the SPPAS-aligned data (*i.e.* 1,857 datapoints), and 2.34% of all P2FA-aligned datapoints (1,988 datapoints). Once again, these numbers refer to datapoints, that is to say to vowels, not to words or syllables. The numbers of words is much lower, 597 words for the SPPAS-aligned data, and 549 for the P2FA-aligned data, out of 68,212 and 66,621 respectively (*i.e.* 0.88% and 0.82% of the total number of words). The ten most frequent words featuring syllabic mismatches are plotted in figure 1.12. As an

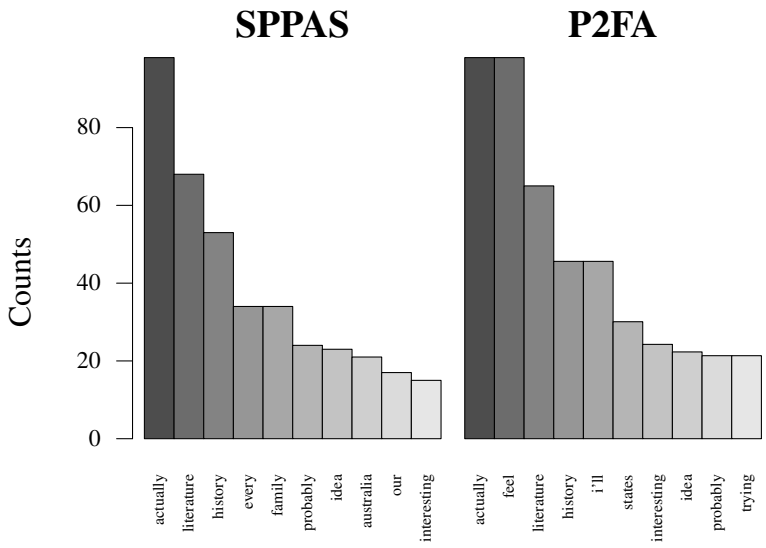


Fig. 1.12: Number of occurrences of words featuring syllabing mismatches. *Left panel:* SPPAS-aligned data; *right panel:* P2FA-aligned data.

indication, "actually" in the SPPAS-aligned data accounts for 13.1% of all the 748 words featuring syllabic mismatches; in the P2FA-aligned data, "actually" and "feel" each account for 9.9% of the 1,016 mismatching words. The question now is then the following: for those words, is the mismatch systematic or does it only occasionally occur within each word category? These proportions are plotted in figure 1.13. In the SPPAS-aligned data, "australia", "idea" and "literature" have 100% of their occurrences featuring a mismatch

between the number of syllables determined by the LPD in the LPDSC column, and that in the SC column, determined from the transcribed pronunciation provided by the aligner. In the P2FA-aligned data, "feel" "I'll", "idea", "literature" and "states" also feature syllabic mismatches in all their occurrences. These cases are of lesser interest because they are based on localized misinterpretations or hard-coded discrepancies in transcriptions due to differences in the rhoticity of the varieties of English taken into account in the dictionaries. An example of the former case in SPPAS is "idea", transcribed as /aɪˈdɪə/ in the LPD and correctly labelled as disyllabic. SPPAS however transcribed it as /aɪˈdiːə/, so that PRAAT03 failed to analyze /iːə/ as a single diphthong, and therefore ascribed the word three syllables. The same logic applied, unfortunately, to "I'll" in the P2FA-aligned data: this contraction was misinterpreted by PRAAT03 as a disyllabic word because of the superscripted /ə/[24]. An example of differences in rhoticity is that of "literature". SPPAS gives the following

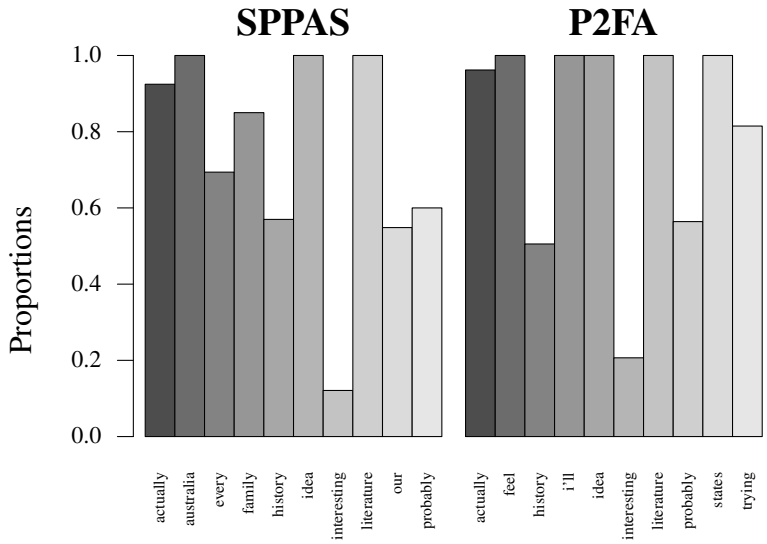


Fig. 1.13: Per-word proportions of occurrences of mismatches.

transcription: /l I 4 3:r @ tS 3:r/, whose IPA equivalent is /ˈlɪɹɜˑɹətʃɜˑː/. PRAAT03 correctly

[24] This is also the of "feel", transcribed as /fiːəl/ in the LPDPRON column of the two datasets... But for some reason, its number of syllables in the SPPAS-aligned dataset is correctly listed as 1 in the LPDSC column... But as 2 in the same column of the P2FA-aligned dataset... Besides being an appeal to the reader's indulgence, this example lifts the veil on the complexity of the syllabifying process. Another instance of such an unexpected discrepancy between two values which the two datasets are supposed to share is "states" – here again, it is correctly listed as a monosyllabic in the SPPAS-aligned dataset. But according the P2FA-aligned dataset, the LPD says this word has 0 syllable.

interprets the transcription as one of a four-syllable word. Likewise, P2FA gives an ARPAbet transcription of /L IH1 T ER0 AH0 CH ER0/ for the word, which is also interpreted as a four-syllable word. However, the LPD transcription, being British, is the following: /ˈlɪtr.ətʃ.ə/[25] – that of a three-syllable word. All words whose occurrences are systematic mismatch can also be easily corrected, in one fell swoop. Words featuring variations, on the other hand, cannot be, because the two aligners somehow change the transcriptions of the same words from one occurrence to the other.

How then to account for those mismatches? The study of one particular case, that of the transcriptions of "interesting" is particularly revealing. Such a case "mid-stream"

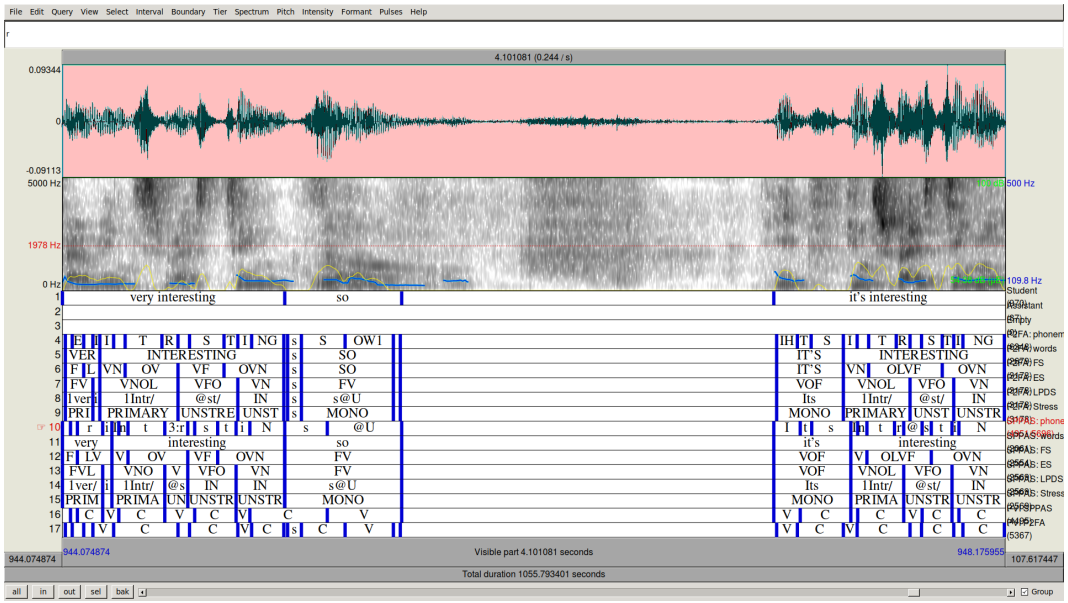


Fig. 1.14: Example of a varying transcription: on row 10, SPPAS chooses two different transcriptions for "interesting".

transcription change is illustrated in figure 1.14. Note that when listening to the file, neither occurrence sounds like a four-syllable word. The first occurrence displays four syllables – /Int3:r@stiN/, whereas the second occurrence only displays two – /Intr@stiN/. It is unclear why such variability in the transcriptions is observed, even when inspecting the spectrograms. The LPD version used by PRAAT03 only uses one entry for "interesting", namely /ˈɪntr.əst.ɪŋ/,

[25] This transcription abides by the principles posited in Wells (1990). The affricate is not split (principle n°3). MaxCoda is posited for the stressed syllable (principle n°1). In cases where the two adjacent syllables have the same stress level, the consonant goes in coda position of the preceding syllable (principle n°2).

and it always contains three syllables. The SPPAS version of the CMUPD lists three entries for "*interesting*": .

1. Int3:r@stiN (/ˈɪntɜˑːrəstɪŋ/)
2. Int3:ristiN (/ˈɪntɜˑːrɪstɪŋ/)
3. Intr@stiN (/ˈɪntr.əst.iŋ/)

The first two options feature four syllables in total, while the last one, just like the LPD transcription, has three. All three entries[26] are present in the dataset. SPPAS can read customized orthographic transcriptions that point to a specific realization (for instance, we could have <intresting> to force a trisyllabic realization). In our case, "*interesting*" is spelt identically, so that the variation probably has to be attributed to the acoustic models implemented in SPPAS[27]. Likewise, P2FA uses two transcriptions:

1. IH1 N T R AH0 S T IH0 NG (/ˈɪntr.əst.ɪŋ/)
2. IH1 N T ER0 AH0 S T IH0 NG (/ˈɪntɜˑːrəstɪŋ/)

Is there a way to determine the criteria used by the two aligners to select one entry rather than another[28]? There are 15 and 25 instances of "interesting" transcribed with four syllables in the SPPAS- and P2FA- aligned datasets respectively. Seven of these instances are common to the two corpora. Conversely, for the three-syllable versions of "interesting", there are 88 common instances out of the 109 SPPAS-aligned ones and the 96 P2FA-aligned ones. Another word worth investigating because it appears in the list of the 10 most frequent words with syllabic mismatches for the two aligners, and also features instances with and without mismatches, is "history". It is transcribed as a disyllabic in the LPD, /ˈhɪs.tri/. In the SPPAS-aligned data, it is at times transcribed as a disyllabic (/hIstri:/) or a trisyllabic (/hIst3:ri:/). In the P2FA-aligned dataset, it comes up alternatively as two-syllable /HH IH1 S T R IY0/ or three-syllable /HH IH1 S T ER0 IY0/. There are 40 SPPAS-aligned disyllabic instances,

[26] The choice of vowels for these transcriptions will not be discussed here.

[27] Brigitte Bigi in SPPAS 1.7 uses Julius by default.

[28] The focus here is on the variations in syllable numbers, so the study of the vocalic qualities of the third syllable of SPPAS-aligned four-syllable versions of "interesting" will be left to further research.

against 46 P2FA-aligned one. 30 of these occurrences are common to the two datasets. In the case of the trisyllabic versions, 37 of the 53 and 47 SPPAS- and P2FA- aligned occurrences are shared across the two datasets. Further research on other words would be required to find out whether the aligners select given transcriptions according to detectable criteria. The common instances are too few to assume any relationship between the processes underlying the selection of the transcriptions. However, the logic underlying the variations in transcriptions is common to both words "interesting" and "history": in both cases, the statuses of /ɜː/ and /r/ are at the heart of the decision process[29].

This section has attempted to show how syllabification was made in the `TextGrids` and how the syllable counts were determined. It is hoped that the obstacles encountered, the problems that remain, and possibly some solutions to solve them, have been described here in a clear enough way. It is now time to deal with preliminary analyses regarding the durations of the phones and the leareners' speech rates.

1.7 Preliminary analyses

Section 1.7.1 provides an analysis of the formant tracks extracted from the spectrogram visible in figure 1.8, as well as a presentation of the distribution of the vowels' durations (regardless of their quality); section 1.7.2 assesses the learners' speech rates.

1.7.1 Formant tracks and vowel durations

This section serves as a preliminary investigation of the accuracy of the automatic extraction procedure. The formant tracks of the vowels shown in figure 1.8 are compared to the data gathered by `PRAAT03`. The top row of figure 1.15 shows the formant tracks of

[29] Incompetence is hereby declared regarding issues in rhoticity, and the syllabic status of /r/. In its four-syllable LPD transcription of "history", /r/ is attached to the coda: /ˈhɪs.tər.i/. For "interesting", /r/ is an onset: /ˈɪnt.ə.rest.ɪŋ/. Why this is the case, whether it matters on the decision process – are questions best left to experts in rhoticity.

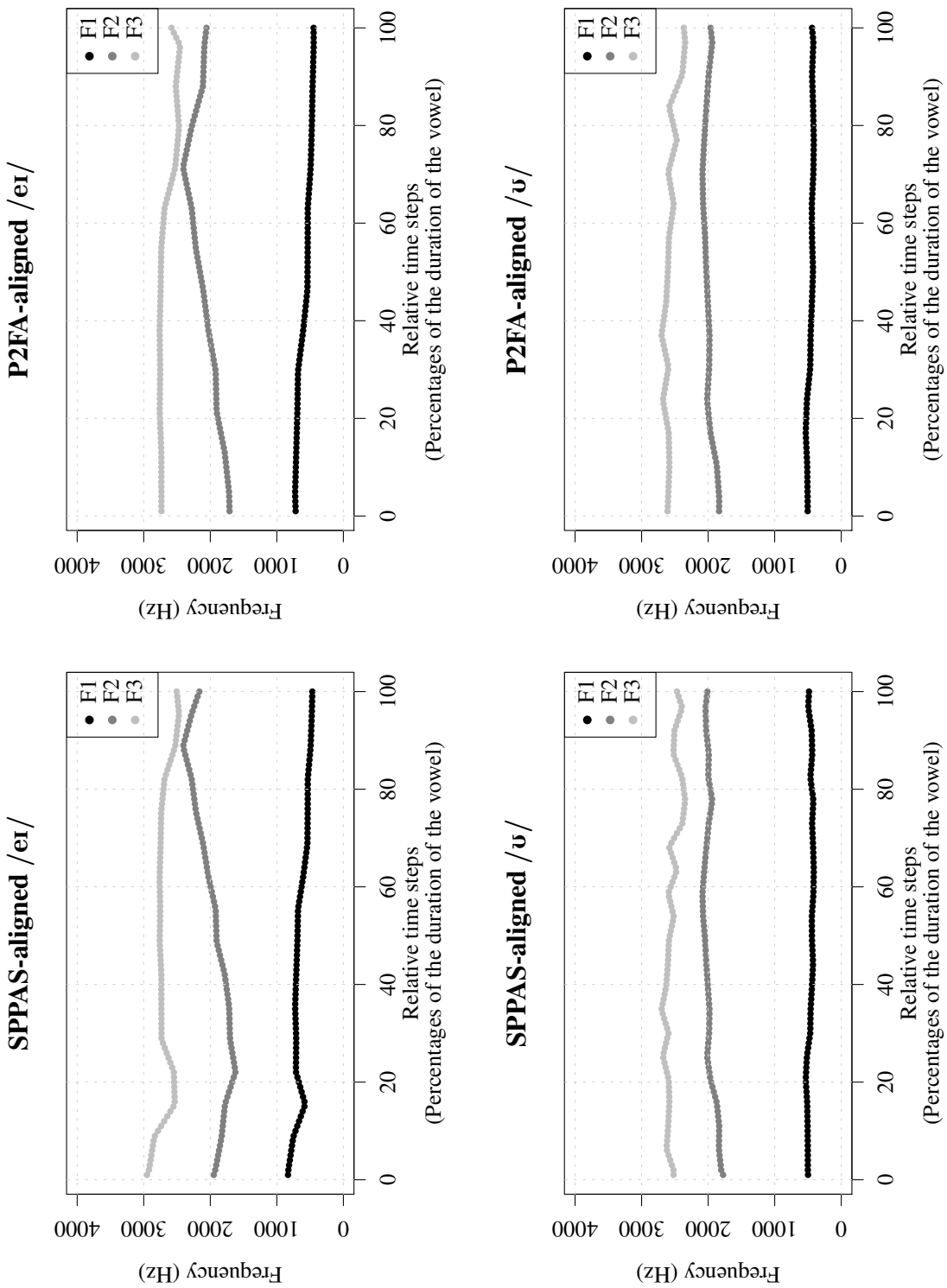


Fig. 1.15: Formant tracks for the first syllable of the word "favorite" (top row) and for the word "book" (bottom row), as pronounced by speaker DID0108 in Session 2.

vowel /eɪ/ in "favorite" for F1, F2 and F3, while the bottom row shows /ʊ/ in "book". The values are obtained from the SPPAS-aligned and P2FA-aligned phonemic tiers for speaker DID0108 in session 2. Visual inspection reveals that the formants' curves in the spectrogram of figure 1.8 and those of figure 1.15 for /eɪ/ (in the top row) seem to match to a certain extent. The case looks different for /ʊ/ in "book". The formant tracks obtained *via* `PRAAT03` seem to be more precise than what can be observed in the spectrogram. Assuming this impression is correct, it may be explained by the differences in duration between the two examples. In practice, the procedure of extracting formant values at 100 relative time locations, regardless of the vowel's duration, normalizes these durations. As far as automatic extraction is concerned, however, no other solution could be adopted. Taking measurements at absolute steps of 5 milliseconds, for instance, would have resulted in discrepancies in the number of columns in the `.csv` file from one vowel to another. As shown in table 1.10 and figure 2.3, in the case of absolute five-millisecond steps, 20 columns would have been needed on average $(0.1/0.005)$ – a procedure less precise than 100 measurements at relative time locations.

The durations of vowels feature almost no scatter: only 778 SPPAS-aligned vowels are longer than 0.5 second, out of 92,458 in total; this number drops to 480 for P2FA-aligned vowels. 26 and 19 vowels are longer than a second with SPPAS and P2FA respectively. This relative absence of spread can be seen in the upper panel of figure 2.3, which shows the boxplots of SPPAS- and P2FA- aligned vowels' durations, along with their kernel densities. The minimal value for SPPAS-aligned vowels is 9.74×10^{-6} ("it's", DID0108, session 3). One other vowel has a similar value: 4.17×10^{-6} for "I" (speakers DID0068 in session 3). These values are clearly the sign of a misalignment[30]. Although nothing is mentioned in their respective documentations, the two aligners seem to have a cut-off duration threshold under which no vowel is recognized (the anomaly mentioned above set aside). The cut-off

[30] Fine-tuning the interval alignment and repeating the procedure several times on these files did not iron out the anomaly. Its cause remains unknown for the time being.

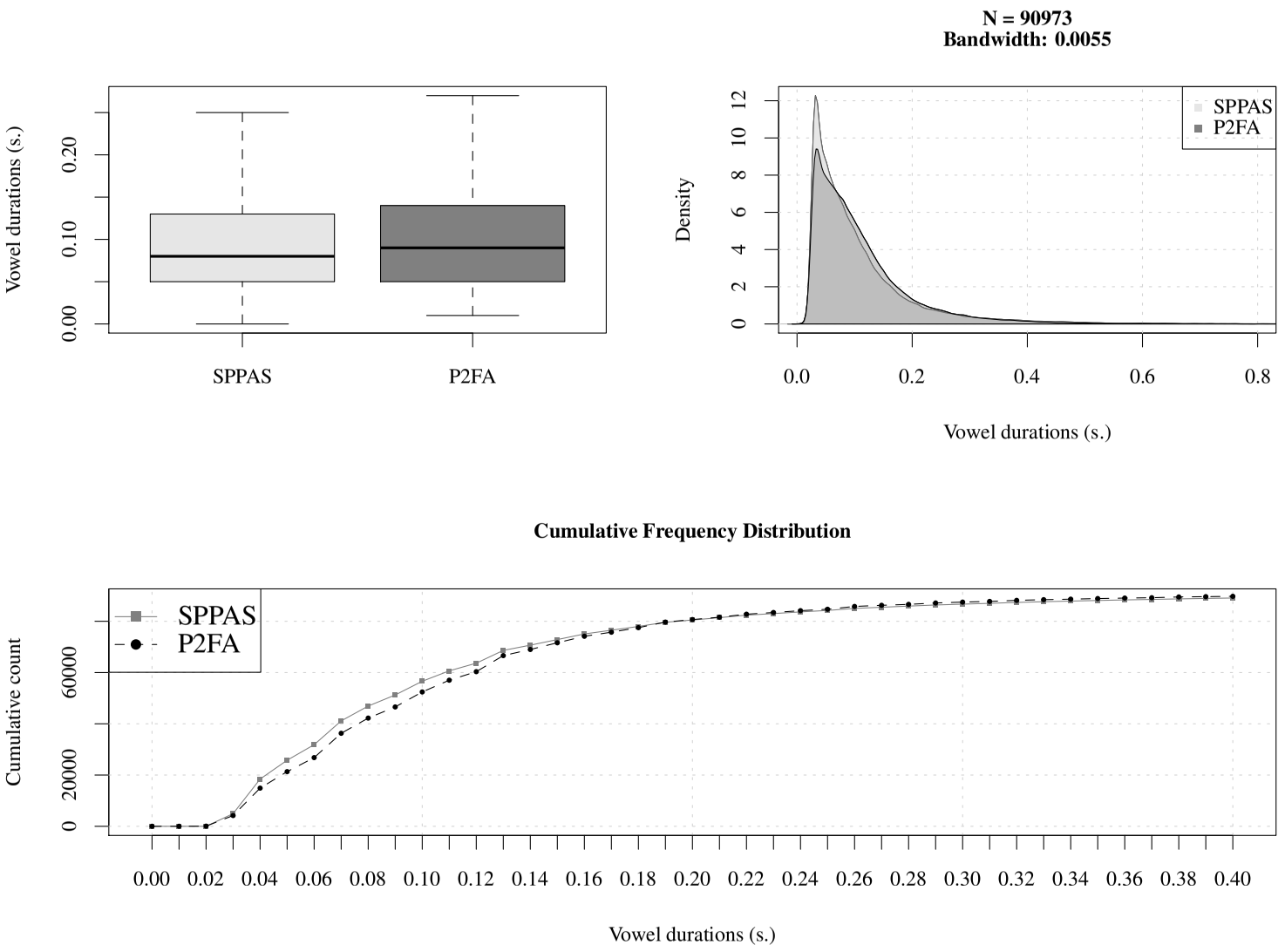


Fig. 1.16: Per-aligner distribution of vowel durations in spontaneous speech; medians and quartiles (left); kernel density plot (right); cumulative frequency (bottom).

Table 1.10: Minimum and maximum vowel durations (in seconds)

		SPPAS	P2FA	French
SPONTANEOUS	Minimum:	9.74×10^{-6}	0.01	NA
	Maximum:	1.62	1.55	NA
	μ:	0.107	0110	NA
	σ:	0.095	0.087	NA
	m:	0.08	0.09	NA
READING	Minimum:	0.03	0.03	1.53×10^{-6}
	Maximum:	1.09	0.78	0.41
	μ:	0.20	0.18	0.10
	σ:	0.13	0.08	0.06
	m:	0.17	0.17	0.08

thresholds are 0.03 second for SPPAS, and 0.01 second for P2FA. Evidence for this is the following: using *R* v3.2.0 (R Core Team (2015)), no entries are returned for SPPAS durations under 0.029 second; at 0.030, 4,909 entries are listed. The same phenomenon applies to P2FA: at 0.009, no vowels are returned; 11 when the threshold is set at 0.01. In the case of P2FA, this low threshold at 0.01 seems mostly theoretical, however. The real cut-off point seems to be 0.030 – the same as SPPAS. At 0.029, 94 P2FA-aligned vowels are returned. At 0.030, the number rises to 4,192. This phenonemon is clearly visible in the bottom panel of figure 2.3.

Boxplots and density plots for the vowel durations of reading tasks can be seen in figure E.1.

1.7.2 Speech rate

Speech rate is often used as a measure of proficiency (*e.g.* Towell et al. (1996), O'Brien et al. (2007)). It is usually calculated by counting the number of syllable nuclei which is then by the duration of the word (*cf.* Towell (2002), de Jong & Wempe (2009) and the references therein) or by the number of syllables pronounced by second (*cf.* Dellwo & Wagner (2003)). There are several ways to define speech rate. One is to calculate the average number of words

per seconds; this method being dependent on the lexicon used, whose distribution in terms of length is itself can vary considerably, it is not adopted here; another way is to count the number of syllables; one last way is to count the number of phones. Those last two ways are investigated here.

Syllables per second

Figure 1.17 plots the number of syllables against the aggregated durations of all five sessions (session 1, session 2, session 3, session 4 and the reading task in French), using both the SPPAS- and P2FA-aligned intervals. As is clear from visual inspection, the two variables seem to be strongly linearly correlated. This is confirmed by the Pearson correlation coefficient: $r_{SPPAS} = 0.9986165; r_{P2FA} = 0.9951977$. A simple linear regression

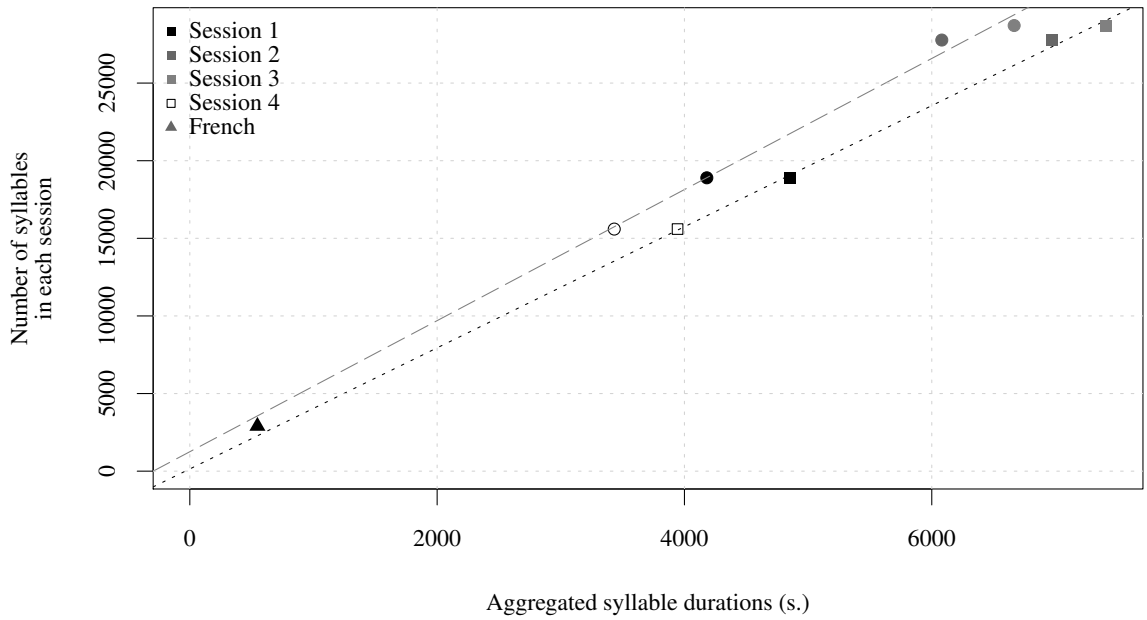


Fig. 1.17: Scatter-plot of the number of syllables against the aggregated syllables durations; squares: SPPAS-aligned intervals; circles: P2FA intervals; grey dashed line: regression line for the P2FA-based model; black dotted line: regression line for the SPPAS-based model.

was calculated to predict the number of syllables based on their durations for both aligners.

Table 1.11: Statistics of the SPPAS and P2FA linear models for the syllable-based and phoneme-based calculations of per-session speech rates.

		F	df	p	R^2	Intercept	Slope
SYLLABLE	**SPPAS**	721.3	1	0.001384	0.9972	143.45	3.9007
	P2FA	206.7	1	0.004802	0.9904	1252. 52	4.2222
PHONEME	**SPPAS**	209.9	1	0.00473	0.9906	4612.48	8.9173
	P2FA	96	1	0.01026	0.9796	7380.78	9.5294

A significant regression equation was found in the case of both aligners, as displayed in table 1.11 (in the first two top rows). A one-second increase in the duration of a given session therefore corresponds to a 3.9007 increase in the number of syllables pronounced in that session for SPPAS-aligned intervals, and to a 4.2222 increase for P2FA-aligned intervals. These very close findings for the two aligners can be accounted for by the fact that differences in alignment result in differences in phonemic intervals, but not in the number of syllables. From this it can be infered that the time gaps between SPPAS-aligned intervals and P2FA-aligned intervals are only marginal. No effect from sessions is visible: speech rate did not evolve from one session to another. The two simple linear models that were calculated did not include, of course, the values for the reading task in French. That the values for the reading task in French fit the regression lines may serve as an indication of a certain syllable-based isochrony. If we apply the equation for calculating the residuals: $e = y - \hat{y}$ to French values, we obtain: $e_{SPPAS} = 2901 - (3.9007 \times 542.93 + 143.45)$ and $e_{P2FA} = 2901 - (4.222 \times 542.93 + 1252.52)$, *i.e.* $e_{SPPAS} = 639.74$ and $e_{P2FA} = -643.77$. In the case of the SPPAS linear model, the maximum absolute residual value is 445.87 (in session 2); e_{SPPAS} is roughly 50% as high as the maximum absolute residual value. The P2FA linear model is more clear-cut: the maximum absolute residual value is 845.564 (session 2), and the second highest residual value is 691.005 (session 3); in other words, e_{P2FA} is lower than two residuals out of four. These two observations are arguably strong arguments in

favour of some sort of syllabic isochrony in the learners' spontaneous speech. This hypothesis will have to be further investigated in future research.

Phones per second

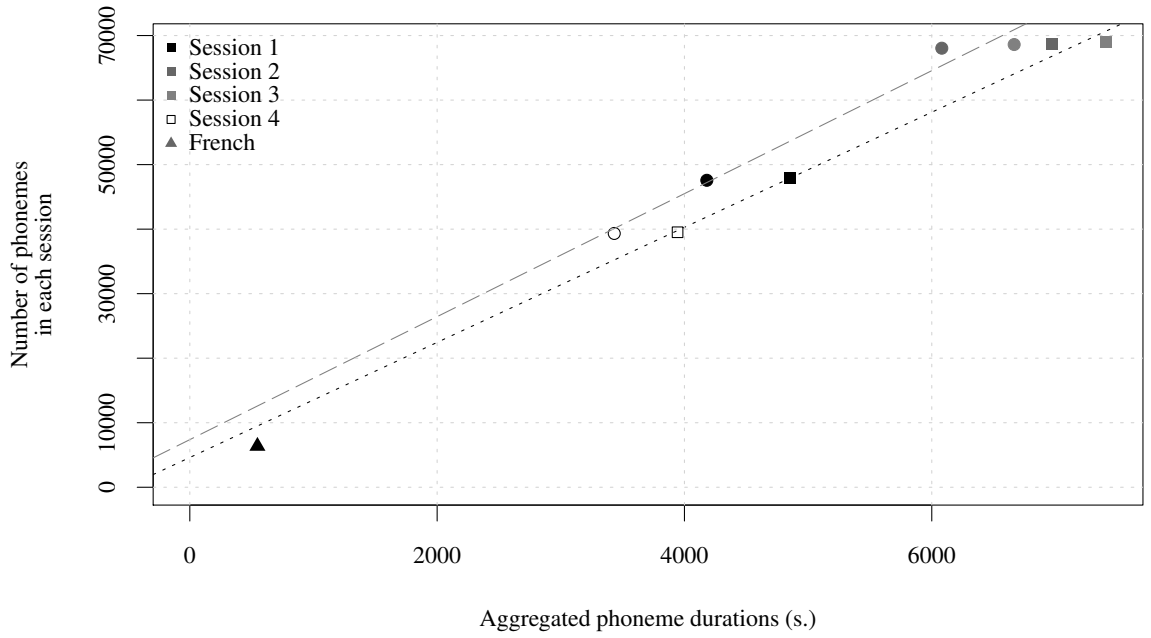


Fig. 1.18: Scatter-plot of the number of phonemes against the aggregated phonemes durations; squares: SPPAS-aligned intervals; circles: P2FA intervals; grey dashed line: regression line for the P2FA-based model; black dotted line: regression line for the SPPAS-based model.

At first sight, the findings from the previous section do not seem to be borne out if the speech rate is calculated from the number of pronounced phonemes, rather than on the number of pronounced syllables. Along the same lines as the previous figure, figure 1.18 plots the number of phonemes against the aggregated durations of all five sessions (session 1, session 2, session 3, session 4 and the reading task in French). The Pearson correlation coefficients for the SPPAS and P2FA values are $r_{SPPAS} = 0.9953$ and $r_{P2FA} = 0.9897$. Two linear models were also calculated, and their statistics can be found in the bottom row of figure 1.10. In this instance: $e_{SPPAS} = 6371 - (8.9173 \times 542.93 + 4612.48)$ and $e_{P2FA} =$

$6371-(9.5294\times542.93+7380.78$, *i.e.* $e_{SPPAS}=-3082.95$ and $e_{P2FA}=-6183.577$[31]. The maximum absolute values of the residuals in the SPPAS and the P2FA models are 1874.53 (in session 2) and 2709.9 (also in session 2). It would therefore not be reasonable to assume some sort of continuity between the durations of phonemes in the reading task in French and those in the recordings of spontaneous speech. At least the French values for the calculation of speech rate based on the number of phonemes per second are not predictable by the linear models to the same extent as they can be from the syllable-based linear models. In terms of acquisition, this could be interpreted as a paradigmatic shift away from the native language, and therefore as evidence of acquisition[32]. This seeemingly further reinforces the assumption that there may exist syllable-based isochrony: the overall picture is however more complex. Table 1.12 shows that the speech rate in French is higher than the speech rate in English, with more syllables and more phonemes pronounced per second in French than in English . However, the table synthesises session speech rates to the detriment of individual variation, which is hidden behind the aggregated data. Still, these rates roughly correspond to the slopes of the model lines plotted in figure 1.17 and 1.18. The greater number of syllables per second in French (5.31 *vs.* 3.93 and 4.49 on average for SPPAS and P2FA respectively) can be accounted for by the smaller number of phonemes per syllables in French than in English (2.20 against 2.5). Unsurprisingly, more phonemes are articulated per second in the learners' native language: on average, 9.77 phonemes per second for SPPAS, 11.08 for P2FA, and 11.65 in the French data. The question therefore arises of whether the syllabic speech rate is a

[31]The reader will have noticed that 6371 is not a multiple of 13, while 13 participants completed the reading task in French, and the number of phonemes can be reasonably be assumed to be the same from one participant to the other. This is however not the case because of the title, "*le géant égoïste*", which some participants chose to read, and others did not. The same remark of course applies to the total number of syllables, 2901. The aggregated durations on the *x*-axis of both figures 1.17 and 1.18 are also the same: the aggregated durations of syllables are the same as those of the phonemes that make up the syllables.

[32] This of course raises the crucial issue of temporal cues in the recognition of phonemes in English. Depending on the dialects (*cf.* Morrison (2008) and the references therein), the difference between /iː/ and /ɪ/ is one of quality exclusively, not one of quality *and* length, making the "ː" symbol redundant in the transcription. This problem is compounded by teaching practices in France, where /iː/ is often referred to as "*le i long*" – the long "i". Regardless of whether this is a correct way of teaching the pronunciation of this vowel, in this state of affairs, our observed gaps between the durations of French and English phonemes *are* evidence of some sort of acquisition, namely the taking into account – right or not – of temporal cues.

Table 1.12: Speech rates in syllables per second (top row) and phonemes per second (bottom row) for SPPAS- and P2FA-aligned spontaneous speech and the reading task in French; third row: number of phonemes per syllable (*i.e.* the ratio of the second and first row); fourth row: average phoneme durations; last row: standard deviations of phoneme durations.

	SPPAS				**P2FA**				**French**
	Session 1	Session 2	Session 3	Session 4	Session 1	Session 2	Session 3	Session 4	
SYLLABLE/S.	3.90	3.99	3.87	3.95	4.52	4.57	4.31	4.54	5.31
PHONEME/S.	9.88	9.85	9.32	10.02	11.38	11.19	10.29	11.46	11.65
PH. PER SYLL.	2.54	2.47	2.41	2.53	2.52	2.45	2.39	2.52	2.20
AVG. PH. DUR.	0.101	0.102	0.107	0.100	0.088	0.089	0.097	0.087	0.086
PH. DUR. σ	0.086	0.102	0.097	0.084	0.064	0.066	0.082	0.066	0.056

simple translation of the characteristics of the phonemes – whether the whole is other than the sum of its components. Attempting to answer this question by looking at speech rates only is far from trivial. One first reason is that the only adjustable variable is of course phoneme duration. The last row of table 1.12 shows greater standard deviations in English phoneme durations than in French. Such greater variability can however have several explanations: *(i)* it could simply be an effect of the acquisition process, and of articulatory difficulties; *(ii)* the available data are of fundamentally different natures: the greater stability in French could be an effect of reading, as opposed to producing spontaneous sentences; *(iii)* temporal cues matter more in English than in French (*cf. e.g.* Hillenbrand et al. (2000)): the higher standard deviations may therefore simply be a consequence of *improved* mastery of English; *(iv)* the apparent predictability of the French syllabic speech rate from the English data, and the similarly apparent impredictability of the French phonemic speech rate, could be an artefact of the corpus size: 6,371 phonemes were collected in total for the French reading task, when the smallest number of phonemes, in Session 1, reaches 39,518. An analysis of speech rates without more data, access to spectral specificities or speakers' idiosyncrasies therefore seems unlikely to provide a satisfying answer to the question of whether French speakers somehow resyllabify their productions in English.

This cursory, homemade approach mostly aimed at checking the consistencty of the two aligners . More detailed analyses of rhythm should factor out tasks, probably distinguishing

the cut off point from initial monologal situations of the interviews to final dialogues for the LINDSEI-inspired tasks. It may well be the case that the speech rate is not consistent over time for some speakers as accommodation seems to have played a role for some of the files investigated in (Burin & Ballier (2017)).

1.8 Conclusion

In this chapter, the processes by which the data was generated have been detailed. A grand total of 81 `TextGrids`, one for each of the 23 learners across the three or four sessions they took part in, have been created, with tiers for the alignments carried out by the two aligners SPPAS anf P2FA, and for the pronunciations listed by the LPD. French and English syllable boundaries have also been emulated, making it possible for future research to investigate possible acoustic cues, and the influence of syllabic templates on vocalic realizations. Parallel to this process, two main datasets, one for each aligner, centralized extra-linguistic, linguistic and acoustic information for all the vowels aligned in the `TextGrids`. The two spreadsheets contain 92,330 (for SPPAS) and 92,091 (for P2FA) datarows, each row corresponding to one extracted vowel. Each vowel has 542 attached cells for data on the speaker, the session, the word, the duration, the formant values taken at each centile of the vowel's duration, etc. The same workflow, of generating `TextGrids` and centralized spreadsheets, was also applied to three subcorpora: a homemade subcorpus of native speakers with 4,542 and 4,586 vowels extracted for SPPAS and P2FA respectively; a subcorpus of a list of English words recorded by the learners as part of the LONGDALE project, whose spreadsheets contain information for 1,750 for each aligner; and finally a subcorpus for the vowels extracted from a French text read by the participants, also as part of the LONGDALE project. Since the recordings are in French, only SPPAS could align it, and information for 2,901 vowels was collected. Unlike the three other corpora, however, formant values were extracted not at each centile of the vowels' durations, but at each decile.

Some observations were also made about the quality and reliability of the information that was collected: F_0 and F_4 for instance were shown to feature more undefined values than the three other formants. The processes underlying the labelling decisions of the aligners have also been tentatively shed light on, although "black box" effects undeniably remain. "The" for example seems to see its vowel labelled by both aligners according to formant-based decisions. This potent feature however is not applied across the board, and a word almost equally as frequent and subject to vowel reductions as "to" does not benefit from it. An assessment of the quality of the syllabification alignments, along with the complex processes required to carry them out, has been undertaken. A lot of mismatches were found to have been caused by discrepancies between the American-based dictionaries used by the aligners on the one hand, and the British-based dictionary used by `PRAAT03` for syllabification on the other. However, in cases not so rare, alternative pronunciations, mostly involving variable syllabic statuses for /r/, either as a coda or an onset[33], were selected by the aligners. Future research will have to determine along what guidelines the choices for one transcription over another are made. Finally, preliminary analyses were made of formant tracks, vowel durations and speech rates, in an attempt to assess the viability of the extraction procedures regardless of vowel qualities. These vowel qualities, and especially those of the monophthongs, are the main focus of the next chapter.

[33] *c.f.* footnote 29.

Chapter 2

Speaker-independent Analyses

This chapter analyses the vowels collected following the procedures described in chapter 1, without taking into account the cross-speaker differences. This means that acoustic analyses will compound formant values regardless of the learners who pronounced the vowels. The first section details technical preliminaries such as the parts of *R* codes which are recurring across various scripts, or the colours chosen to represent vowels. The second section aims at assessing the accuracy of the automatic alignment and extraction. The third section investigates the disparities, such as formant standard deviations or lexical distribution, between the monophthongs. The fourth section deals with issues in normalization: it descrives the obstacles and theoretical contradictions underlying normalization methods applied to the sort of corpus under study (*i.e.* spontaneous learner speech), while experimenting with a procedure to study the effects of normalization on corpus analysis. The final section investigates the connection between the Euclidean distances of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts in the vowel space, and the surface of that vowel space.

2.1 Technical Preliminaries

In the following pages, the same R code will have been used to extract and analyze data. Because it was established in section 1.7.1 that both SPPAS and P2FA had minimal thresholds for vowel durations, under which no vowels were recognized, only the phonemes lasting longer than the minimal duration (0.03s.) will be taken into account.

Calculations were always made on both datasets, *i.e.* the SPPAS-aligned and the P2FA-aligned datasets. However, comparing the differences is not always justified – especially when they are small or even non-existent. In those cases, results using the SPPAS-aligned dataset are presented. Choosing SPPAS as the default dataset makes sense, since part of the P2FA data, especially the data related to syllabic structure, is inferred from SPPAS-generated alignments (*c.f.* section 1.2.2).

When selecting vowels for study in either dataset, the transcription system from the LPD was used. The reason why the LPD transcription was chosen is that SPPAS and P2FA use SAMPA and ARPAbet respectively (*c.f.* table 1.4). Cross-comparisons between the two datasets are therefore much easier to make by resorting to their common transcription system. Another reason is that the LPD provides a British-based pronunciation, whereas the CMUPD versions of the aligners are American-based. With learners' interlanguage being more likely to be unstable, it makes sense to use the more complex vocalic system (*i.e.* the British one) as reference: should learners try to contrast "dog" and "door"[1], for instance, their endeavour (or success!) will be taken into account. The ARPAbet and SAMPA versions of the CMUPD are also coarser grained in their transcriptions than the LPD. One example is the absence of "happy"-tensing /i/ in the SPPAS transcriptions: "easy" is thus transcribed /iːziː/[2]. Perhaps more crucially, there is more consistency in the labeling of vowels by the LPD than by either of the two aligners. The next paragraph explains why.

[1] "Dog" is transcribed /d O: g/ and /D AO1 G/ in SPPAS and P2FA respectively; "door", /d O: r/ and /D AO1 R/.

[2] No stress marks exist either. The exact transcription in the dictionary is the following: /i: z i:/.

The main source of inconsistency is the observed variety of labels for the same vowels in the same words. There are 770 monosyllabic words in the SPPAS dataset (taking the SC column as the base reference, not LPDSC). 39 of these words have their nucleus labeled in more than one manner. The issue is that a lot of these variations cannot be ascribed to reduced values: "and" shows 1,456 entries as strong form (/æ/), 576 as weak form (/ə/). The vowel in "was" is labeled /ə/ 381 times; /ʌ/ 87 times; but also /ɑ/ 166 times, and /ɔː/ 14 times. The nucleus of "will" may well be labeled /ə/ 65 times, and /ɪ/ 78 times, but there is no clear syntactic or semantic reason in the occurrences that may explain the choice of one pronunciation over the other. The vowel in "the" is transcribed as /ə/ 1,995 times, 555 times as /iː/, and 325 times as /ʌ/. Sometimes, the vowels chosen in weak forms[3] are not consensual: in "been", /ɪ/ appears 79 times, and /ə/, 19 times... "Just" appears as /dʒʌst/ 67 times, and 256 times as /dʒɪst/. "Your" has 32 occurrences under /ʊ/, 12 under /ɔː/; but "you're" is listed under /ʊ/ 38 times, and under /u/ 22 times... Such variations are not limited to potentially reducible function words: "want" is transcribed /wɑnt/ 46 times, /wɔːnt/, 77 times; "walk" shows as both /wɑk/ and /wɔːk/ 14 times each... When it comes to polysyllabic words, 41 words out of a total of 1,271 words feature vowels with alternative pronunciations. "Accent" is transcribed as /ˈæksɛnt/ 25 times, /ˈæksənt/ 2 times; "upset" is listed as /ʔəpˈsɛt/ one time; <-ed> can be transcribed /ɪd/ or /əd/: 9 times in "started" for the former, 13 times for the latter. "Because" features an alternation of /ɔː/ (491 times) and /ʌ/ (193 times) on the second syllable. The P2FA alignment is overall as subject to variations as the SPPAS alignment: for monosyllabic words, 42 words out of a total of 769 feature at least two different pronunciations of their nucleus. This is the case of "your" (/jɔːr/ 12 times and /jʊr/ 29 times), "you're" (/jʊr/ 49 times, /juːr/ 14 times), "them" (/ðɛm/ 114 times, /ðə/ 15 times), "the" (/ðə/ 1,844 times, /ðʌ/ 405 times, /ði/ 529 times), "if" (/ɪf/ 184 times, /əf/ 77 times)... If 98 out of 1,195 disyllabic words contain a vowel with more than two pronunciations, the explanation seems to be different, however: a lot

[3] A choice, once again, which is not clearly determined by the context in which the words appear.

of these mismatches can be attributed to either vowel reductions, *e.g.* "wasn't", or changes in stress patterns due to the grammatical nature of the words (*e.g.* "subject", "contrast", "conflict"...) "Began" however is transcribed as /biˈgæn/ 3 times, and /bɪˈgæn/ three times too; "because" is listed as /biˈkʌz/ 635 times, and as /biˈkɔːz/ 12 times. From the perspective of this work, alternations in either aligner between /ɪ/ and /i(ː)/ in words like "because" are problematic, because they sever the link between the lax/tense feature of vowels and the syllabic structure they appear in: it is traditional and consensual to consider that lax vowels only appear in closed syllables. If that view were to be challenged in a data-driven approach like this study, the starting point would be to adopt a tagging system that preserves the link between phonemes and syllabic structure. This is one more argument in favour of using the `LPDPHONEME` column in both datasets. Finally, it is worth noting that it was possible to exert control, through PRAAT03, over the `LPDPHONEME` labels, in a way that was not possible with the aligner-dependent `PHONEME` column. For instance, occurrences of "the" were labeled as /ði/ when preceding a vowel sound, /ðə/ otherwise. This introduces a normative aspect which is not necessarily undesirable when dealing with specialized learners of English who are likely to end up teaching the language themselves. For all these reasons, *i.e.* cross-comparisons between the two datasets, consistencies in the labelling of vowels, and manual control through `PRAAT03`, it was deemed reasonable to base vocalic analyses on the `LPDPHONEME` column, rather than on the aligners' `PHONEME` column.

As this study focuses on monophthongs, diphthongs and triphthongs were excluded from the data. Section C.6 provides the piece of *R* code common to all the scripts used to obtain the results detailed below. This common piece of code loads the datasets using `fread` (Dowle & Srinivasan (2017)); excludes the vowels whose duration is shorter than 0.03s.; and only selects monophthongs in the remaining datapoints. Figure 2.1 lists the per-session, per-aligner number of monophthongs thus collected, along with their respective proportions across all sessions in the bottom panel. The labels of these vowels are from the `LPDPHONEME`

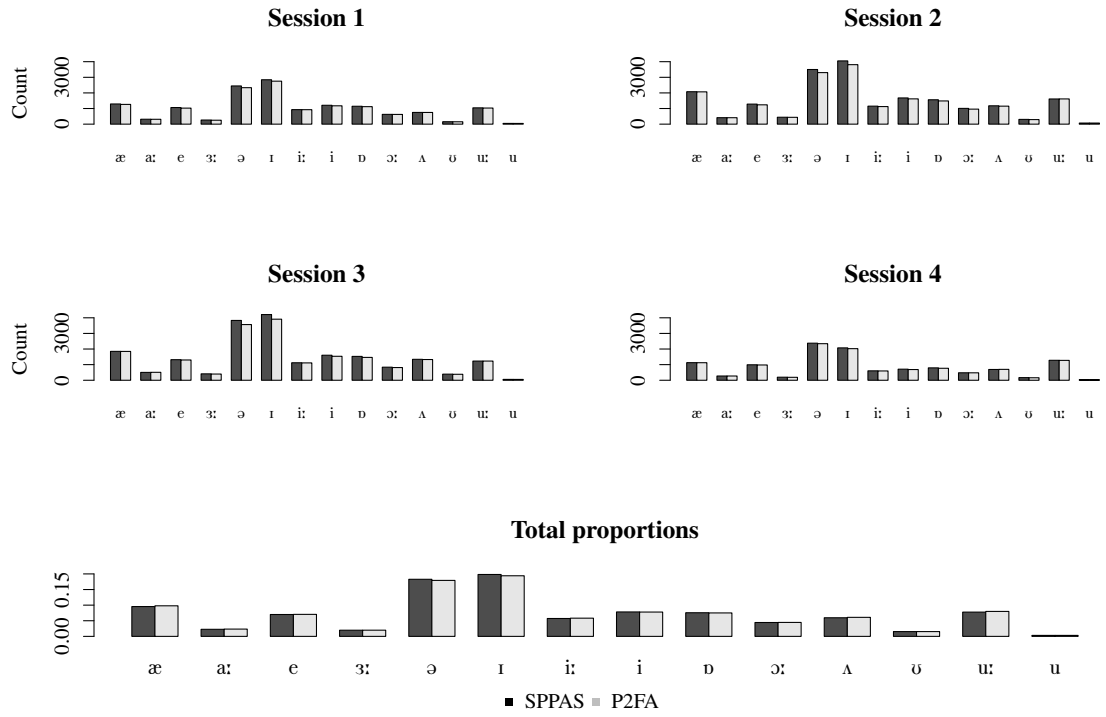


Fig. 2.1: *First four panels:* per-session, per-aligner count of LPDPHONEME monophthongs; *bottom panel:* per-aligner proportion of each LPDPHONEME monophthong across all sessions.

columns, for all the reasons mentioned above. The distribution of vowels across the four sessions is roughly the same. The total numbers of monophthongs for each aligner are the following: 66,470 SPPAS-aligned monophthongs, and 64,407 P2FA-aligned monophthongs. The disproportions in numbers from one vowel to another are worthy of attention: /ɪ/ and /ə/ account for 38% of all the SPPAS-generated data, and 33.4% of all the P2FA-generated data. This skewness in phonemic distribution is an issue that will be discussed in further detail when dealing with normalization (*c.f.* section 2.4). One final note needs to be made

Fig. 2.2: Colour codes of phonemes used in graphs

about the colour codes used in graphs representing the monophthongs. The same colours are applied to the same monophthongs, and are displayed in figure 2.2. These colour codes try to

respect some sort of logic (low front vowels in blue, high front vowels in green, rhotic vowels with a slightly darker shade, central vowels in dark colours, etc.), and make the contrasts which will be focused on, *i.e.* /iː/-/ɪ/ and /uː/-/ʊ/ more visible.

2.2 Assessing alignment and extraction quality

One of the first questions that needs to be dealt with before proceeding forward is that of the accuracy of the alignment and formant extraction[4]. The extraction procedure described in Chapter 1 tentatively tackled this issue through the study of speech rate (*cf.* Section 1.7.2). However, the specifics of spectral analyses have yet to be addressed. Two structural obstacles lie in our way: *(i)* the very nature of the corpus, *i.e.* connected speech; phonation, speech rates and their related coarticulatory effects are likely to affect formants in a way that might compromise their extraction. Besides, although laughter, hesitation markers, coughs and overlaps have been carefully excluded from the segments under study, clear-cut boundaries between words or between silent and noisy moments were at times difficult to establish. *(ii)* That the corpus is also a *learner* corpus compounds these difficulties, as formant instability within a vowel category cannot but be a feature of learner speech. With high dispersion a defining and expected characteristic of the learners' vowels' formants, sorting out which outlying formant values pertain to genuine idiosyncratic pronunciations, and which pertain to errors in automatic extraction, is no easy task. The next section, section 2.2.1, aims at assessing the latter, *i.e.* automatic extraction, by exploring the number of formants with plausible values.

[4] In general, no technical distinction will be drawn in the following paragraphs between "extraction" and "alignment". While we are well aware that "alignment" refers to the process of creating an interval boundary at a given location in the sound signal, and "extraction" to the process of retrieving acoustic data from a particular location in the signal, the fact that these two processes are intrinsically linked (the acoustic data will depend on the chosen point in the signal) means, for our purposes here, that assessing the quality of one is assessing the quality of the other. The two terms will therefore be used interchangeably in this section.

2.2.1 Assessment with predefined formant ranges

One way to look at the problem of the accuracy of automatic extraction is to find out the number of centiles, for each vowel, whose F1, F2 and F3 values fall within a predefined range which would include all potentially realistic values, and exclude straightforwardly abnormal ones. The arbitrary cut-off values adopted here were the following: $F_1 \in [250, 850]; F_2 \in [500, 2500]; F_3 \in [1500, 3500]$. These values were chosen after cross-referencing data in phonetic research. Sundberg (1977, p. 109) states that *"[The range] in adult males averages approximately from 250 to 700 hertz for the first formant and from 700 to 2,500 hertz for the second"* [5]. In their study of journalistic broadcast speech by French and German native speakers, Gendrot & Adda-Decker (2005) chose gender-, language- and vowel- specific ranges for the automatic extraction of formants (e.g. F_1 values for male speakers pronouncing French /i/ were to be superior to 300 Hz and inferior to 2050 Hz – 350 Hz and 2,400 Hz for female speakers). Extreme F_1 and F_2 values for French, independently of gender, were (in Hz) $F_{1min} = 300; F_{1max} = 750$ and $F_{2min} = 850; F_{2max} = 2400$. F_3 was not taken into account in their study. In their chapters on vowels, Ladefoged & Maddieson (1996) do not specifically mention ranges of formants for vowels, but standard axes in graphs range from 200 Hz to 800 Hz for F_1, and from 200 Hz to 2,500 Hz for F_2 [6]. F_3 is not mentioned either. The rather low value selected as a minimum for F_3 (1,500 Hz) can be explained by certain constrictive gestures in French and rhotic English. Lip-rounding to produce /y/ causes F_3 to drop towards F_2 values, from 3,000 to 2,000 Hz approximately (*cf.* Vaissière (2006, p. 75)). Rhotic vowels in rhotic varieties of English, such as /ɜ˞ː/ in "bird", are also produced by constricting anterior and posterior cavities in the vocal tract, which results in F_3 values "well

[5] Although the book deals with the acoustics of the singing voice, in context the remark above does not restrictively apply to male singing voices.

[6] Ladefoged & Maddieson (1996) give values outside the range chosen here (*i.e.* $F_1 \notin [250, 850]; F_2 \notin [500, 2500]$), which may be worth mentioning: in Eastern Arrernte, /ə/ features F_1 values superior to 850 Hz when preceded by /p/ and /t̪/, and F_2 values superior to 2,500 Hz when preceded by /k/ and /t̠/;

below 2,000 Hz" (*cf.* Vaissière (2006, p. 76)). Finally, Benesty et al. (2007, p. 219) note that "F_3 goes as low as 1,900 Hz".

Figure 2.3 shows the count (top row) and proportion (bottom row) of SPPAS-aligned (left column) and P2FA-aligned (right column) phonemes whose F_1, F_2 and F_3 values fall within the ranges defined above. The x-axis increments each centile into the duration of the phonemes. The phonemes analyzed were selected from the two aligners' datasets using the common *R* code described in section 2.1 and given in section C.6. Cursory graphic analysis shows very few differences between the two alignment procedures. These similarities can partially be explained by the inner structure of `PRAAT03`: SPPAS-aligned phonemes are looped through in priority, and P2FA-aligned phonemes are then retrieved from the time locations of the SPPAS-aligned phonemes. P2FA-aligned phonemes are therefore anchored around SPPAS-aligned phonemes. Differences between the two transcriptions (ARPAbet and SAMPA) are neutralized by the selection of `LPDPHONEME` (from the LPD, then) as the tagging method in figure 2.3. With all this in mind, it may come as a surprise that looking at the figure more into details does indeed reveal differences. Before approximately the 20$^{\text{th}}$ centile, SPPAS-aligned phonemes seem to feature fewer within-range formant values. Past that 20$^{\text{th}}$ centile, the global shapes of the proportion curves look similar from one aligner to the other, with the notable exception /u/. However, the specificity of the /u/ curve may be ascribed to the rarity of its occurrences: with only 210 and 201 occurences in the entire SPPAS- and P2FA-aligned corpora respectively (*c.f.* section 2.1 and figure 2.1 for raw phonemic distributions and section 2.3 for more details on the lexical distribution underlying phonemic count), /u/ is roughly 5 times as rare as the second rarest monophthong, /ʊ/, which occurs 1,018 and 983 times in the SPPAS and P2FA datasets. When it comes to comparing the shapes of the proportion curves within each aligner, two curves stand out: rather than the truncated, logarithmic shape all other proportion curves have, the /iː/ and /i/ curves feature a parabolic shape, with formant values more likely to be out-of-range around

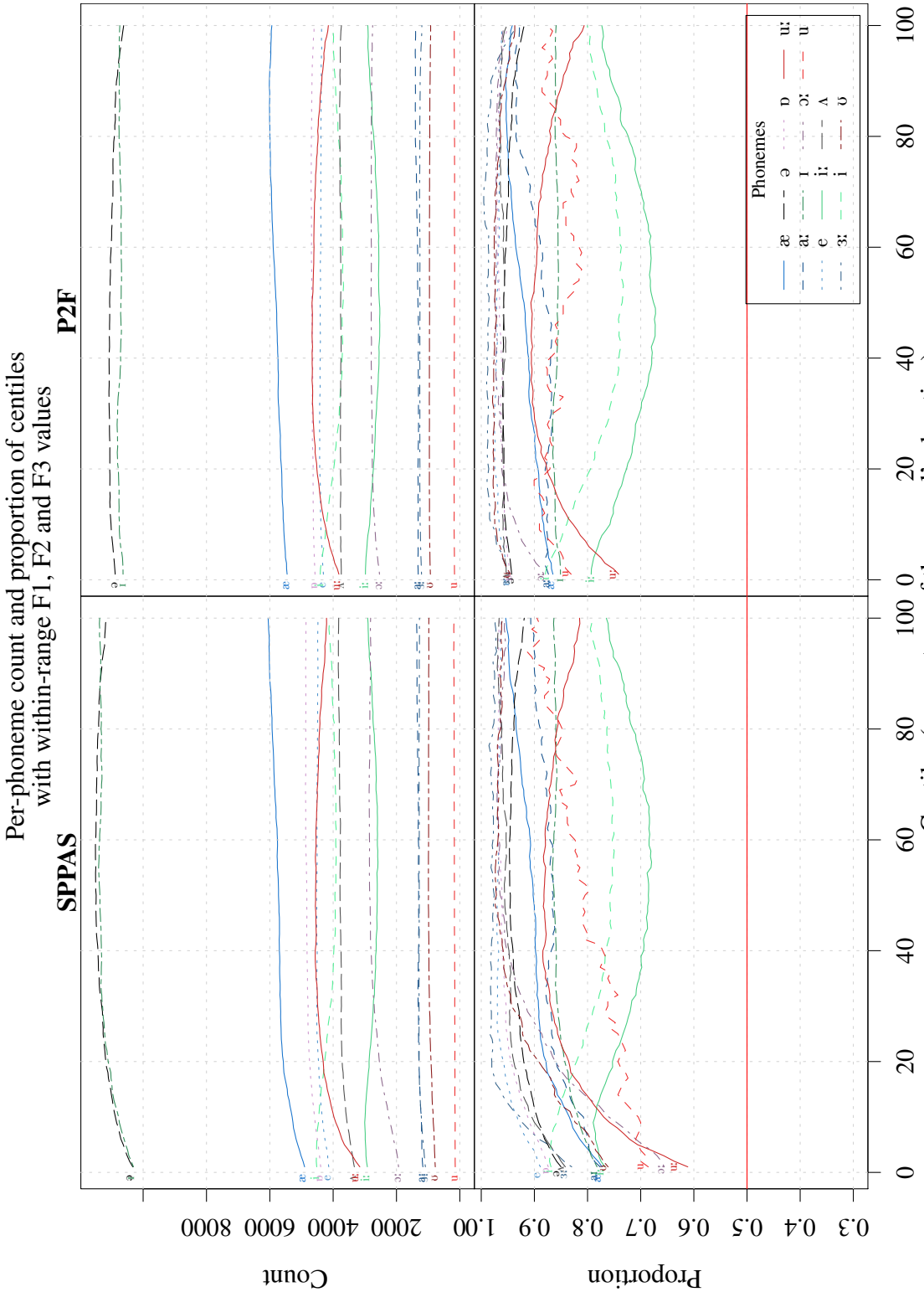


Fig. 2.3: Per-phoneme number (top) and proportion (bottom) of centiles whose F1, F2 and F3 formant values were within-range at a given centile (the $x-$axis). Left panels feature SPPAS-aligned phonemes, right panels P2FA-aligned phonemes. Within-range values are the following: $F_1 \in [150, 950]; F_2 \in [500, 2500]; F_3 \in [2500, 4800]$. Only phonemes with duration > 30 ms. were selected.

the mid-temporal point. With such distinctive differences, the question arises: what is the cause of such a decrease in within-range formant values? Isolating each of the six conditions of the pre-defined formant ranges, *i.e.* for F_1, F_2 and F_3 to be superior to 250Hz, 500Hz and 1500Hz, or inferior to 850Hz, 2,500Hz and 3,500Hz respectively and independently, shows that in both datasets, 99% of F_1, F_2 and F_3 formant values of vowels /iː/ and /i/ on all centiles respect those conditions individually[7] – except for the maximun F_2 condition, where the proportion drops to 92% in both datasets. The observed decrease of within-range formant values for /i/ and /iː/ around the mid-temporal point can therefore be ascribed predominantly to F_2 values which are superior to 2,500Hz, rather than out-of-bounds on either of the five other conditions. Still, as shown in figure 2.4, the overall mean proportion of

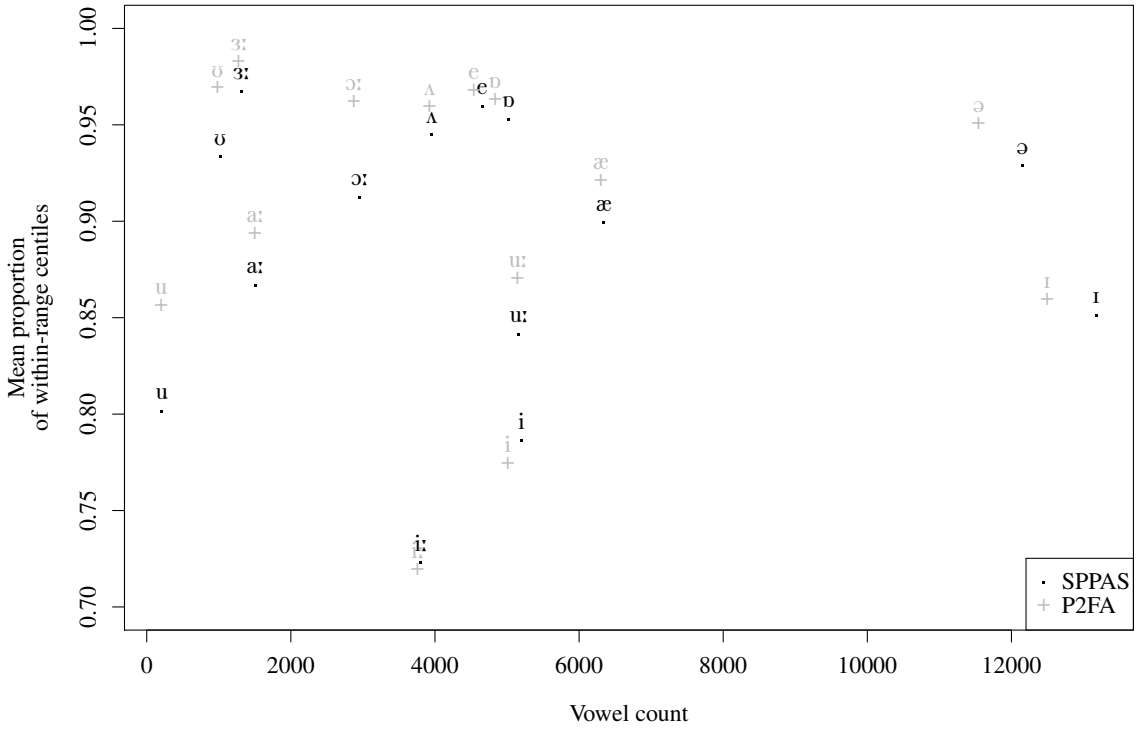


Fig. 2.4: Scatterplot of the average proportion of within-range centiles against the number of occurrences. Black: SPPAS-aligned data; grey: P2FA-aligned data.

centiles within the range of predefined of /iː/ is 72.3% in the SPPAS-aligned data, and 71.9% in the P2FA-aligned data; for /i/, it is 78.6% and 77.4% respectively. These proportions

[7] This means that, for instance, 99% of all F_1 formant values of /i/ on all centiles in a given dataset are superior to 250Hz.

are the lowest among all monophthongs, and are still arguably high. If instead of 2,500Hz, the cap for F_2 is raised to 2,600Hz, the mean proportion of within-range centiles increases substantially in high vowels, and even more in high front vowels: in the SPPAS-aligned dataset, the mean proportions for /ɪ/, /iː/, /i/, /uː/ and /u/ rise by 4.15%, 7.01%, 5.95%, 1.75% and 2.75%, to reach mean proportions of 89.2%, 79.3%, 84.6%, 85.9% and 82.9%; in the P2FA-aligned dataset, by 4.81%, 7.49%, 6.78%, 1.66% and 2.56%, reaching mean proportions of 90.7%, 79.3%, 84.2%, 88.7% and 88.2%. In both datasets, all other vowels feature increases in mean proportions under 1% – including /ʊ/, a characteristic which may at this stage be considered as incipient evidence of the (correctly) central nature of this vowel in the learners' English. Raising the maximum value of F_2 makes sense: certain studies (*e.g.* Gendrot & Adda-Decker (2005), Tubach (1989)) mention high F_2 values in French (2,365Hz in the former, with no standard deviation reported; 2,456Hz for the latter, with a standard deviation of 111Hz, *i.e.* potentially superior to 2,500Hz)[8]. Figure 2.4 also indicates that no correlation exists between the number of centiles within the range of pre-defined values and the number of occurrences of a given monophthong. This absence of correlation entails that the quality of the vowels substantially contributes to the accuracy of the automatic extraction. Looking at figure 2.4 again, the differences between the two aligners are more vertical (*i.e.* due to differences in means) than horizontal (*i.e.* due to differences in numbers of occurrences), with the exception of /ɪ/ and /ə/ (for reasons already hinted at in section 2.1). Interestingly, when ordering these vertical distances, phonological distinctions appear: the five greatest differences in means in increasing order are /ɑː/, /uː/, /ʊ/, /ɔː/ and /u/ – back vowels, with /ʊ/ the exception. Conversely, the only back vowel with a small difference

[8] Raising the F_2 cap by another 100Hz to 2,700Hz returns the same gains in within-range centiles: mean proportions of within-range centiles for /ɪ/, /iː/, /i/, /uː/ and /u/ increase by 6.6%, 12.1% (!), 10.2% (!), 2.74% and 5.1% in the SPPAS-aligned dataset; the increases in the P2FA-aligned datasets are even more substantial: 7.53%, 12.89% (!), 11.52% (!), 2.52% and 5.34%. Clearly the lower number of within-range centiles in high vowels is due to abnormally high F_2 values, rather than abnormal F_1 or F_3 values. It could make sense to raise the F_2 cap even beyond 2,700Hz: Hillenbrand et al. (1995)'s average F_2 value for /i/ among American female speakers is 2,761Hz. All in all, our original F_2 cap of 2,500Hz can be argued to be somewhat conservative.

in means between the two aligners is /ɒ/, which does not exist in American English, and is therefore potentially neutralized by the learners. These relative disparities in the means of proportions of within-range centiles between monophthongs arguably constitute further evidence of the accuracy of the automatic extraction process: the emergence of phonological distinctions in these disparities indicate that the distinctive formant profiles which establish the unique quality of vowels have been preserved.

In this section, 8,726,900 F_1, F_2 and F_3 values for the SPPAS-aligned dataset, and 8,510,000 for the P2FA-aligned dataset, came under study[9]. Regardless of vowel quality, in total 88.6% of all formant values of the SPPAS-aligned monophthongs as measured on each centile were within-range, *i.e.* were superior to 250Hz and inferior to 850Hz for F_1, *and* superior to 500Hz and inferior to 2,500Hz for F_2, *and* superior to 1,500Hz and inferior to 3,500Hz for F_3. In the P2FA-aligned dataset, the proportion is 90%. Although the arbitrary nature of the cut-off values lends itself to discussion (considering the considerable varieties and variations of languages and speakers), and although changing these values may yield very different results (as was shown with raising the F_2 cap), it is still contended here that the formant values obtained through automatic extraction on each centile are very robust and plausible, and may serve as the basis for further phonetic and phonological analysis. The next section aims at consolidating this contention, and explores in further detail the accuracy of the automatic process by looking at vowel trapezoids.

2.2.2 Vowel trapezoids

Another way of assessing the accuracy of the automatic extraction is by visually examining the distribution of monophthongs on the vowel trapezoid. Section 2.2.1 has demonstrated the plausibility of the extracted F_1, F_2 & F_3 values for each centile. But do these plausible values correspond to values in keeping with the phonemes that they are attached to? Figure 2.5

[9] As per section 2.1, these figures exclude monophthongs longer than 0.03 second, and the diphthongs and triphthongs of the datasets.

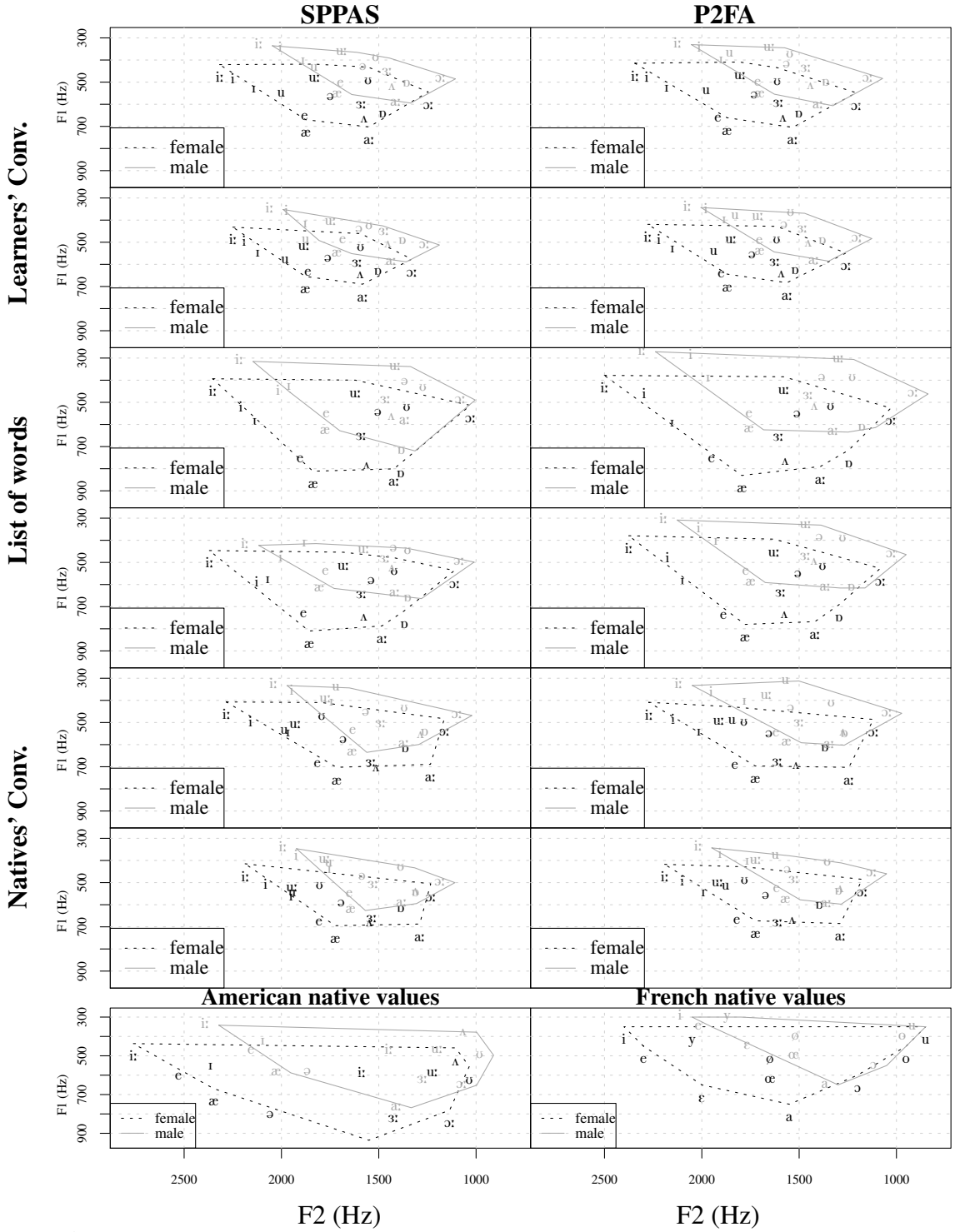


Fig. 2.5: Vowel trapezoids from mean raw F_1 and F_2 values – Odd-numbered rows: vowel trapezoids for male and female speakers from SPPAS-aligned (left) and P2FA-aligned (right) F_1 and F_2 mid-temporal formant values – even-numbered rows: same, but from means over all centiles. Rows 1 – 2: learners' conversations data; rows 3 – 4: learners' list of words; rows 5 – 6: natives' conversations; Row 7: Hillenbrand et al. (1995)'s data for American speakers (left); Gendrot & Adda-Decker (2005)'s data for French speakers (right).

represents various vowel trapezoids across all monophthongs obtained from the means of raw F_1 & F_2 values in Hertz. The conditions described in section 2.1 were implemented and only the vowels matching those conditions were retained. Values were conflated across speakers and sessions, but not gender. The dotted black line, and the continuous grey line trace the convex polygon linking outermost points for female and male speakers respectively. Two different methods were used to compute the F_1 and F_2 means: *(i)* in the first method, the means were calculated from the mid-temporal values of F_1 and F_2 of each datapoint; the results are shown in the odd-numbered rows of the first six rows in figure 2.5. *(ii)* in the second method, the F_1 and F_2 means were calculated first by averaging over each F_1 and F_2 centile value on each datapoint, then by averaging all these means for each monophthong; the obtained values are displayed in the even-numbered rows of the first six rows in figure 2.5. The advantage of computing means this way is that this second method includes all F_1 and F_2 values from all centiles, thereby making it possible to assess the accuracy of the formant extraction process: implausible means would point to inaccurate extraction. Excluding row 7 for the moment, the left column features SPPAS-aligned data while the right column features P2FA-aligned data. The main corpus of recorded conversations between learners and native assistants is used in the first four panels. The next eight panels display the trapezoids obtained from the monophthongs extracted in two sub-corpora (*c.f.* section 1.3): *(i)* the learners' recorded lists of words[10] (*c.f.* section A.2.1); *(ii)* natives' spontaneous speech, with no distinctions made between the varieties of accents (only differences in sex were taken into account). Row number 7, the last row, plots the trapezoids from the values reported in Hillenbrand et al. (1995) for American speakers in the left column, and from the values reported in Gendrot & Adda-Decker (2005) for French speakers in the right column. These two panels serve as references to compare the other twelve panels, which are all based on data generated from `PRAAT03`.

[10] The list of words being mostly a list of monosyllabic words, /u/ is not pronounced. Occurrences of /ə/ come from "cancel", "possible", "quality", "people", "serious" and "oral".

From the perspective adopted in this section, *i.e.* not one where acquisition is considered[11], but one where the accuracy, or at least plausibility, of the obtained formant values is assessed, the resulting trapezoids in figure 2.5 constitute solid evidence that the process that generated the datasets from the two alignments made by SPPAS and P2FA works: all monophthongs are correctly located, at least relatively to one another, in the vowel space; even from an absolute point of view, the areas where they are plotted are in keeping with common representations[12], and seem to reflect their places of articulation in a plausible way. It is hoped that these findings, along with those presented in section 2.2.1 justify the use of the generated dataframes as basis for actual phonemic study.

2.3 Disparities in phonemic distributions

In this section, the distribution of each monophthong is investigated. Subsection 2.3.1 investigates the per-phoneme standard deviations of F_1, F_2 and, to a lesser extent, F_3 values. Subsection 2.3.2 likewise studies the disparities in Type/Token Ratios (henceforth, TTRs).

2.3.1 Standard deviations

This subsection investigates the standard deviations along the formant tracks of each monophthong.

How distributed around a (hypothetical) centre are the formant values of each monophthong? How much variation do the monophthongs feature? These questions need to be answered: if averages of formant values over several occurrences are used in order to give a representation of how a vowel was pronounced, some assessment of the accuracy of such a method must be provided. But then another question arises: are mid-temporal measure-

[11] These vowel trapezoids will be used again in section 2.5, which explores methods to assess acquisition based on their surfaces.

[12] Only raw data has here been used. Whether, and how, to normalize the data will be tackled in section 2.4.

ments optimal? Considering that formant tracks are available in our corpora, shouldn't it be attempted to maximize the information taken into account in the representation of the vowels' pronunciations? Figure 2.6 shows the standard deviations of F_1, F_2 and F_3 values taken at each centile for each monophthong and each aligner. The top row gives the per-centile, per-phoneme F_1 standard deviations, the bottom row, the F_2 standard deviations (in Hertz in both cases). The left column plots the SPPAS-aligned data,the right column the P2FA-aligned data. The mean curves in each panel, *i.e.* the mean of standard deviations on each centile regardless of the monophthongs, are shown in a thicker, continuous black line. The shapes of the F_1 curves, regardless of the aligner, clearly indicates greater variations at the onset of the vowels. Their offset also shows a slight rise. These higher standard deviations can most likely be ascribed to coarticulation, *i.e.* the influence of the consonantal environment embedding the vowels. However, coarticulation seems to have a greater effect on F_1 values than on F_2 and F_3, which show mostly regular standard deviations after approximately the 20^{th} centile onwards. The most unstable curve in all six panels of figure 2.6 is that of /u/. This is in keeping with its count, the lowest of the corpus: its so few occurrences, combined with its phonological status, *i.e.* of a vowel only present in unstressed (likely to be clipped) syllables, the equally low number of words it appears in ((*c.f.* section 2.3.2 prevented the formation of a cluster of stable formant values, and most likely explain the high and impredictable standard deviations across all centiles. F_3 standard deviations show, if not a reversal, at least a substantial change, in the order of stability of monophthongs: the most dramatic example of this change is /ɜː/, which features the lowest F_2 standard deviations, one of the lowest F_1 standard deviations, but the highest F_3 standard deviations. It may not be that surprising that the three vowels with the highest F_3 dispersions, *i.e.* /ɜː/, /ɑː/ and /ɔː/ are all potentially rhotic: rhoticization usually entails a lowering of the third formant (*c.f.* for instance Davenport & Hannahs (2013)). This lowering might in turn lead the F_3 values to be confused with F_2 values, thereby explaining the rise in F_3 standard deviations. Coarticulatory effects are the

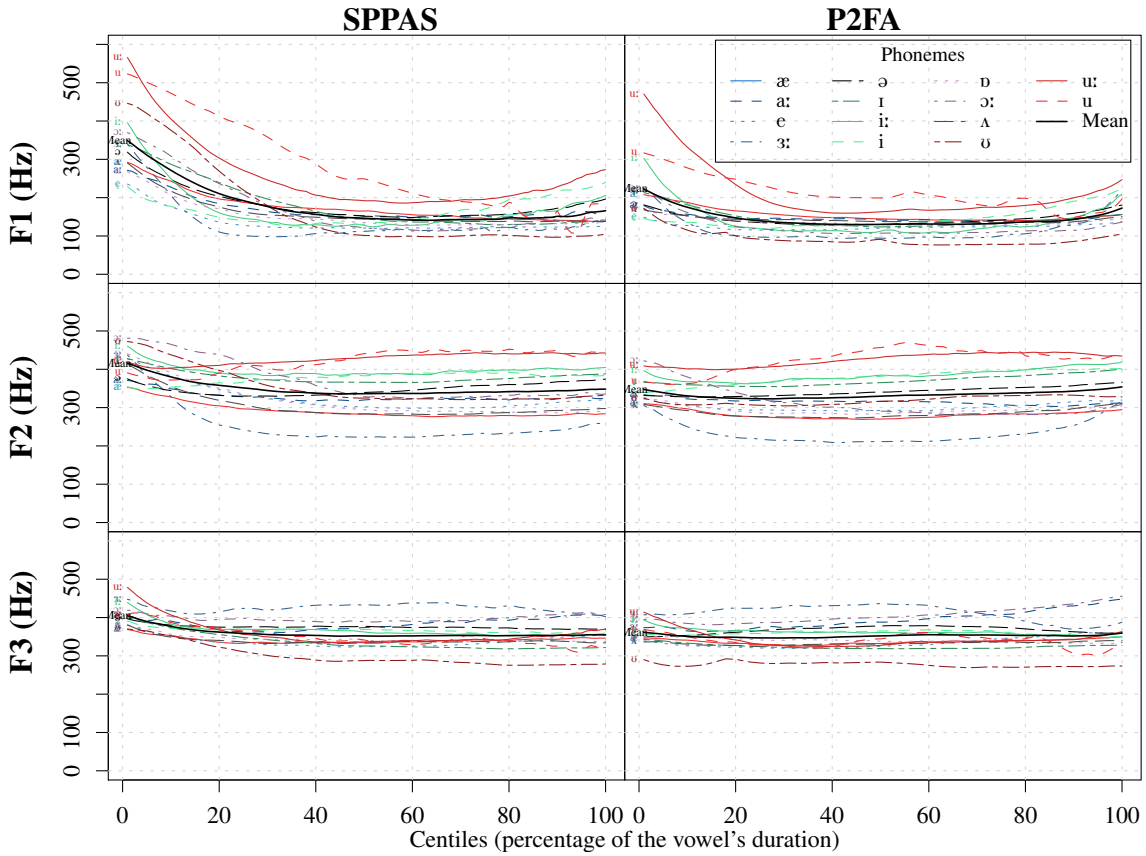


Fig. 2.6: Per-centile mean F_1 and F_2 standard deviations. Top row: F_1 standard deviations (in Hz); middle row: F_2 standard deviations (in Hz); bottom row: F_3 standard deviations (in Hz); left column: SPPAS-aligned data; right colum: P2FA-aligned data.

main reason why formant values are generally measured at the mid-temporal point of the duration of the vowel. The F_2 curves feature fewer differences between onset, middle part and offset of the vowel. The mean F_2 curve for the P2FA-aligned data is even almost flat. The mean values over all standard deviations, across all centiles and not including phonemic differences, are the following: 178.3Hz and 349.5Hz for F_1 and F_2 in the SPPAS-aligned data; 144.8Hz and 333.3Hz in the P2FA-aligned data.

With standard deviations varying across centiles, one interesting question arises: what are the centiles with the minimal variations? In other words, are there centiles where the formant values of a given monophthong are optimized, *i.e.* feature minimal dispersion? Table 2.1 gives the minimal F_1 and F_2 standard deviations for each monophthong and each aligner, along with the centiles where these minimal values are reached. Leaving aside the disparities

	SPPAS				P2FA			
	F_1		F_2		F_1		F_2	
Phoneme	Minimal	Centile	Minimal	Centile	Minimal	Centile	Minimal	Centile
/æ/	137.66	100	276.13	64	140.41	79	269.31	46
/ɑː/	147.60	89	317.79	38	137.81	82	302.85	87
/e/	119.77	53	298.15	59	114.52	28	291.20	57
/ɜː/	98.06	32	222.73	59	93.63	52	208.55	41
/ə/	148.10	60	329.73	32	128.55	39	327.57	13
/ɪ/	129.12	77	365.68	60	123.51	50	354.86	29
/iː/	123.82	52	383.45	31	107.50	58	362.17	22
/i/	126.15	36	348.21	4	119.61	26	339.15	1
/ɒ/	128.27	65	286.56	90	127.65	70	282.68	43
/ɔː/	112.39	63	323.15	60	106.38	42	281.27	64
/ʌ/	137.02	85	281.90	64	134.53	74	273.06	44
/ʊ/	96.67	89	320.76	71	76.60	62	304.62	31
/uː/	185.89	59	402.40	8	159.55	43	398.90	16
/u/	107.39	93	371.19	8	145.47	88	360.63	6

Table 2.1: Per-phoneme minimal formant SDs and centile location

between sessions and speakers for now, table 2.1 shows that no clear picture emerges across monophthongs or even across aligners: for instance, /i/ features low standard deviations rather early in the pronunciation of the vowel, when /æ/ features them rather late. However, having formant tracks for F_1, F_2, F_3 means that a potentially "optimal" centile must exist: if at each centile, the per-phoneme standard deviations for the three formants are multiplied, then the centile with the lowest product can be called "optimal", in the sense that dispersion will be minimal at that centile. We define the Optimal Centile (henceforth, OC) as the centile with the lowest product of the per-centile F_1, F_2 and F_3 standard deviations for a given monophthong. The *R* code to calculate the per-monopthong OCs can be found in section C.7 (it is presented there as a function). As an example, figure 2.7 shows the per-phoneme, per-aligner OCs. No clear trend, such as a phonological distinction or a range within which most OCs would lie, seems to stand out either. The absence of general tendency may be due to the compounding of words, speakers and sessions. The method will be used when investigating the per-speaker, per-session data: in the case of a corpus with skewed distributions (*c.f.* section 2.4 and figure 2.1), OCs seem to be a potentially effective work-around to overcome the disparities in consonantal environments while preserving idiosyncracies, and results obtained with more

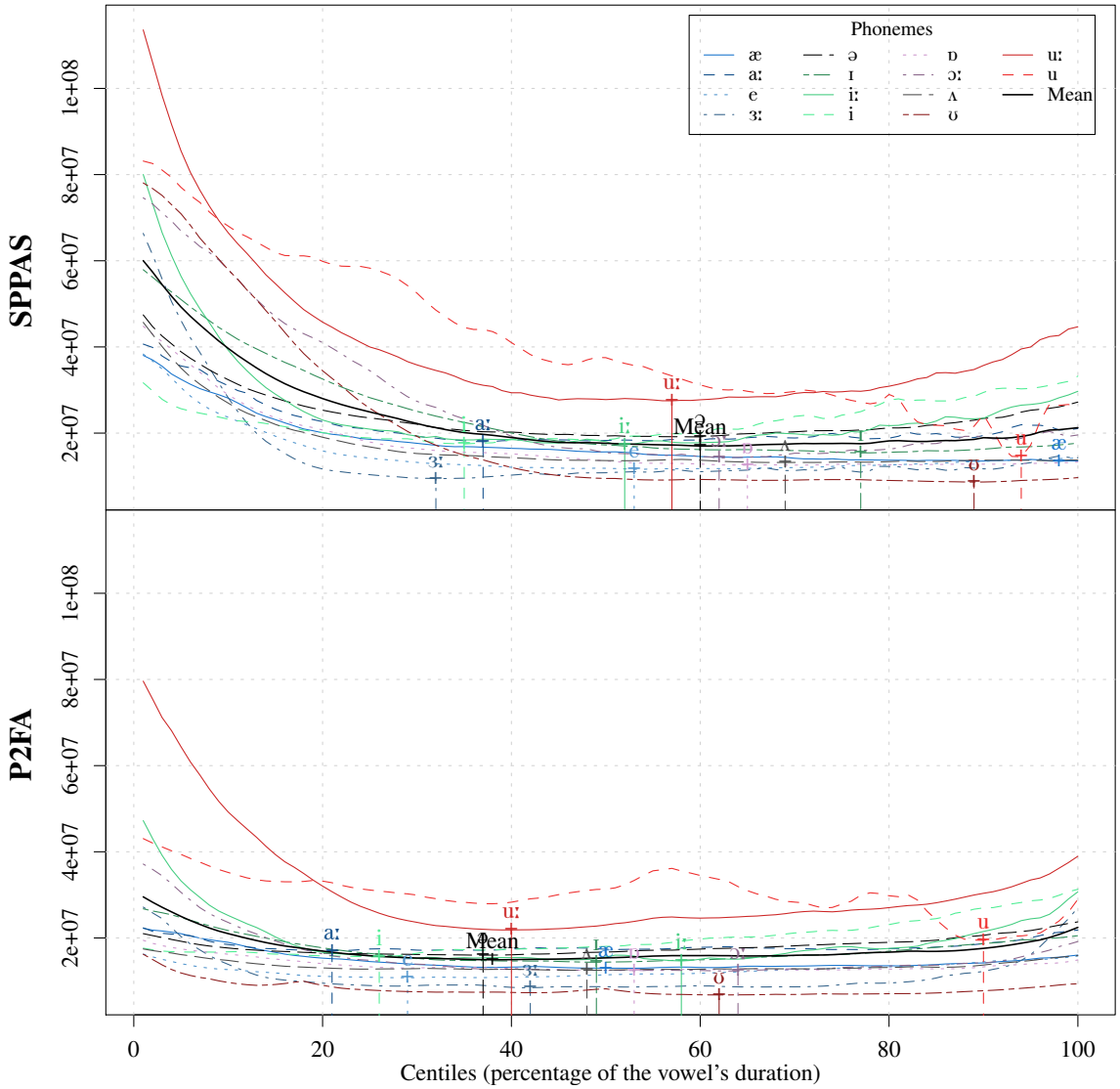


Fig. 2.7: Optimal centiles for SPPAS-aligned data (top panel) and P2FA-aligned data (bottom panel).

classic methods, such as using mid-temporal formant values, will be compared with findings based on OCs.

This analysis of standard deviations would not be complete without a cursory comparison with the two other English subcorpora, *i.e.* the list of words and the native speakers' recordings (*c.f.* section 1.3). figure 2.8 plots the F_1, F_2 and F_3 standard deviations of all monophthongs for the three corpora. It could be reasonably expected that the ascending order in standard deviations would be the following: the native conversations, the learners'

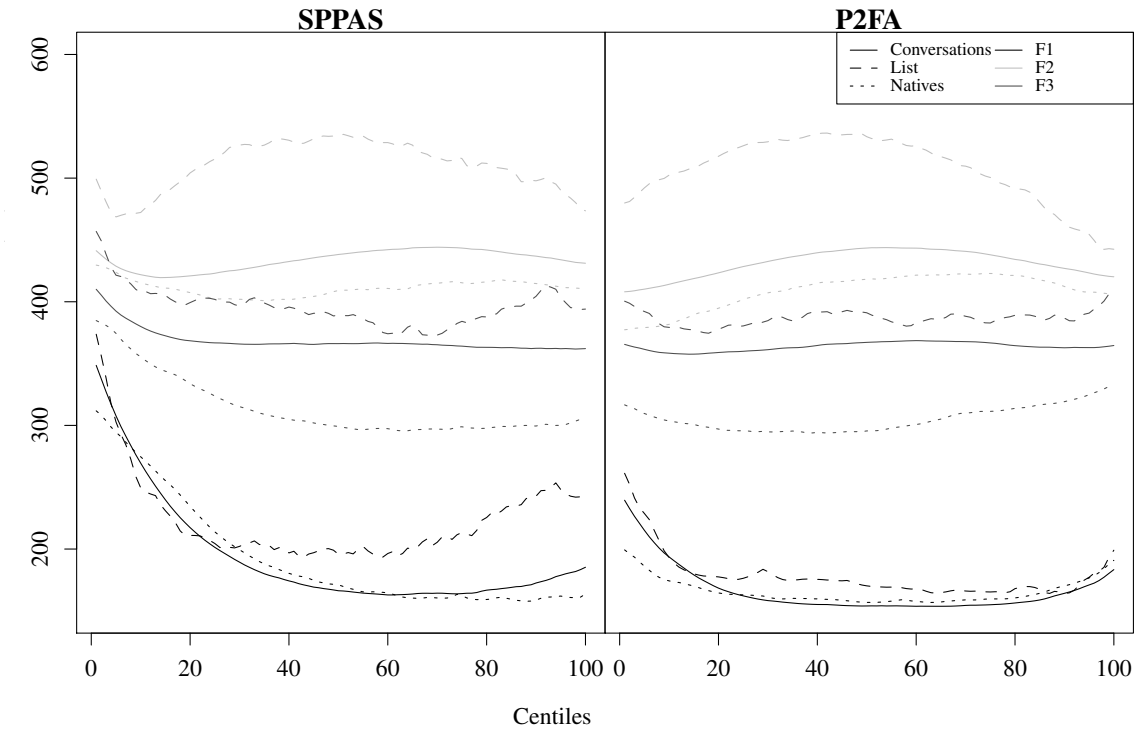


Fig. 2.8: Per-centile F_1, F_2 and F_3 Standard Deviations for all three English corpora. Plain line: conversation (main corpus); dashed line: list of words; dotted line: native conversations. Black: F_1 SDs; dark grey: F_2 SDs; light Grey: F_3 SDs.

list of words and finally the conversations. Dispersion should be the lowest among natives, and indications of how to pronounce words in the list of words should have enticed learners to produce consistent realizations of each monophthong. This, however, turned out not to be the case, as can be seen on figure 2.8: if the native subcorpus does feature the lowest standard deviations across all three formants, the main conversation corpus presents lower SDs than the reading subcorpus, in spite of its much greater number of monophthongs. Explaining these results is challenging: the underlying objective of reading lists is to tap into the learners' competence, *i.e.* their phonological knoweldge, but the objective may have been compromised by the experimental design. Noise may have been created by the phonographematic relations (*c.f. e.g.* "women", "wolf", "who"). Another assumption could be that in conversations, learners tend to use a restricted number of words, and words whose pronunciation is known. Before exploring the validity of this assumption in the next section,

let it be remembered for now that SDs in the main corpus are remarkably limited (within 300-Hertz ranges) and consistent across centiles, with the exception of F_1 SDs for /ʊ/, /uː/ and /u/.

2.3.2 Type/Token Ratios

This section investigates the lexical disparities that underlie the distribution of phonemes in the corpus: as seen in section C.6, the number of monophthongs varies greatly from one number to another. The question therefore arises whether these disparities in the numbers of occurrences can be accounted for by the nature (lexical or functional) and frequency of the words where the phonemes appear. One way to look at phonemic and lexical distributions

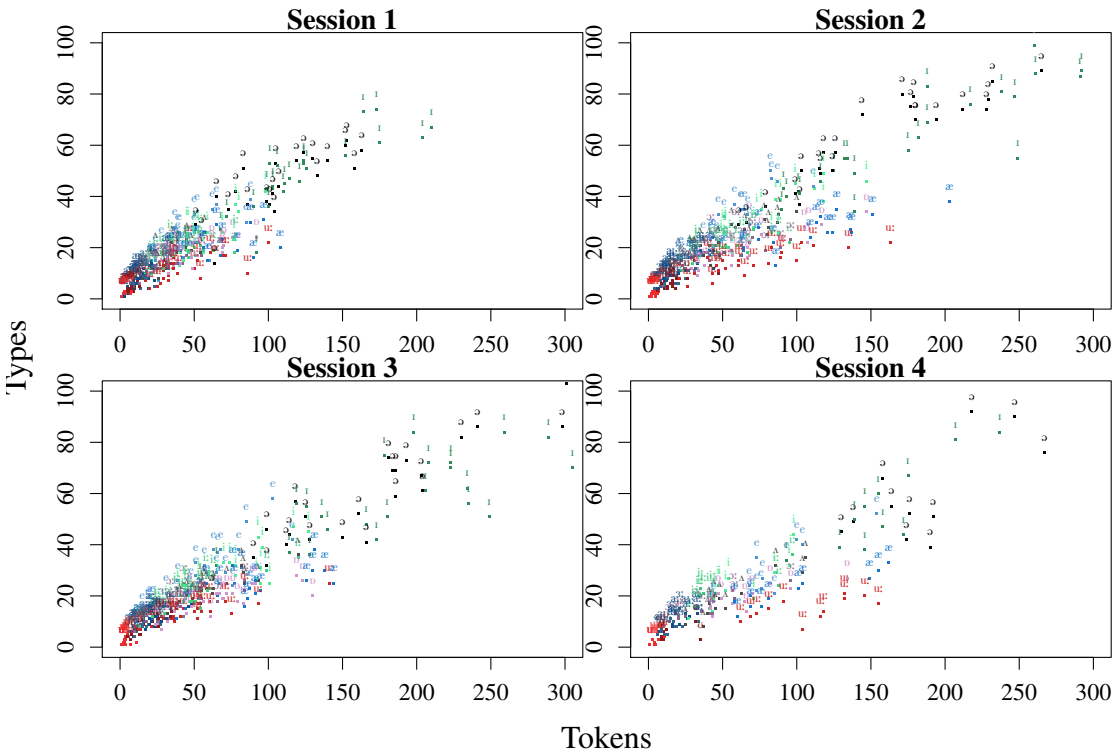


Fig. 2.9: Per-session, per-speaker types and tokens in the SPPAS-aligned data.

is to count the number of different words featuring a monophthong (types) and compare it with the number of occurrences of the monophthong (tokens). Figure 2.9 plots the number of types of monophthongs against the number of tokens for each session and each speaker in

the SPPAS-aligned data[13]. Each datapoint corresponds to the number of tokens and types of words containing a given monophthong in a given learner's session. The learners' IDs were not included not to clog the graph. One clear trend emerges: the monophthongs are distributed along the axes in the same way regardless of speakers and sessions. The lowest Type/Token Ratios (TTRs) can be found in words containing /ɪ/ and /ə/. /ʊ/, /uː/ and /u/ feature the lowest numbers of tokens, which indicates a very limited lexical distribution. /iː/ is consistently present among the monophthongs with the highest number of tokens. How do

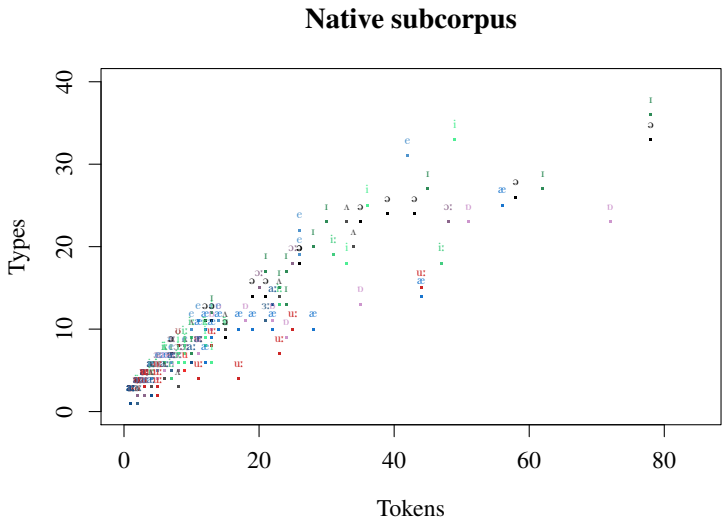


Fig. 2.10: Per-speaker types and tokens in the native speakers' subcorpus.

these results fare with native values? The same *R* code was applied to the native speakers' subcorpus, and is presented in figure 2.10. Because the subcorpus has fewer datapoints, the scale of the graph were modified, leaving 2 /ɪ/ and /ə/ items off the chart, and one/æ/ datapoint. The figure shows that the per-monophthong distributions of types and tokens are very similar, with the same vowels located in the same places of the graph. The similarities between the the two corpora in the cases of /ɪ/, /ə/,/iː/ and /uː/ in particular are especially visible. This is all the more striking as the native subcorpus is comparatively small, and made up of random recordings from conversational podcasts.

[13] Types and tokens being only marginally aligner-dependent, only the SPPAS data will be looked at in this section.

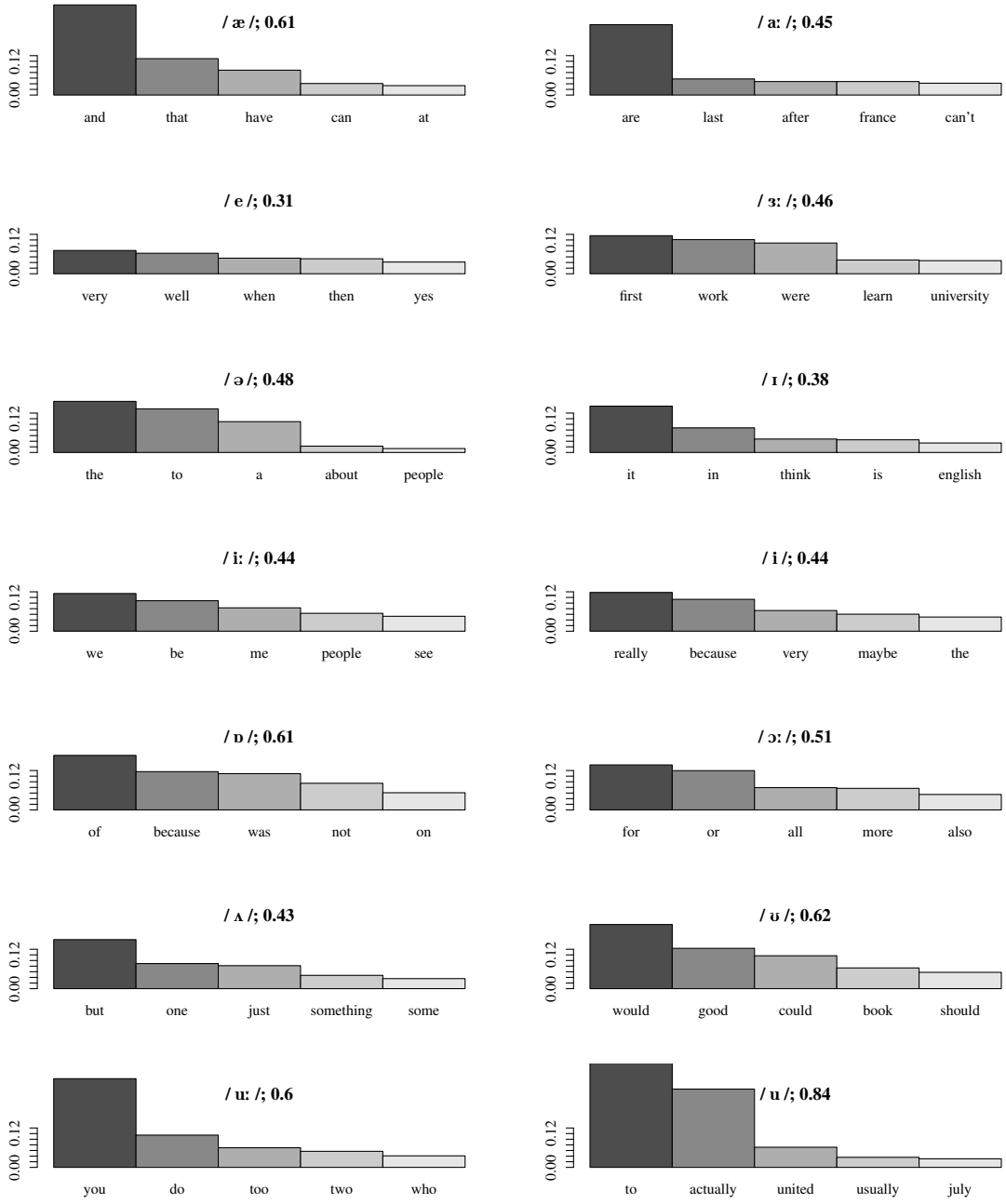


Fig. 2.11: Barplot of the proportions of the five most frequent words for each phoneme. The figures on top of each panel indicate the cumulative sums of those proportions.

How, then, are the phonemes distributed across words? Figure 2.11 gives the proportions of the five most frequent words for each phoneme. On top of each panel, the cumulative sum of those five highest proportions is indicated. It can be seen that the lowest value for these cumulative sums is that of /e/: 0.31, *i.e.* the five most frequent words featuring /e/ account

for almost a third of the entire number of occurrences of that monophthong. All other sums are higher than this. Figure 2.11 sheds excruciating light on vowel reduction issues: a case in point is that of /æ/, whose five most frequent words ("and", "that", "have", "can" and "at") are more often than not likely to undergo reduction to /ə/. As things stand, the current workflow is blind to the syntax. These crucial questions shall be set aside for now, as they are but indirectly linked to the main purpose of this work. Vowel changes due to succeeding phonemes *have* been taken into account in certain instances, however. This is the case for instance with "the", transcribed /ðə/ when followed by a consonant, and /ði/. Note that the number of occurrences are somewhat imbalanced: 2,650 occurrences of "the" are followed by a consonant, for only 224 followed by a vowel. The same logic was applied to occurrences of "to", transcribed /tu/ when followed by a vowel (153 occurrences), and /tə/ when followed by a consonant (2,229 occurrences). Connecting target pronunciation (*i.e.* to reduce the vowel or not to reduce the vowel) to syntactic and contextual information will be kept for future research. Let it be remembered, however, that all most frequent words but two ("first" for /ɜː/ and "actually" for /u/) are function words. Their dominance is well-established: out of the 70 words listed in figure 2.11, 16 are pure lexical words – the rest are function words. Looking at /ɪ/, /iː/, /i/, /ʊ/, /uː/ and /u/ in more detail, interesting differences become visible: firstly, the cumulative sums of i-sounds are much lower than those of u-sounds (0.38, 0.44 and 0.44 against 0.62, 0.6 and 0.84 respectively); secondly, u-sounds seem slightly more likely to appear in non-function words than i-sounds (6 lexical words featuring u-sounds are among the five most frequent words for the three u-sounds, against 4 for i-sounds), even though this statement must be somewhat qualified: /u/ is comparatively much rarer (*c.f.* figure 2.1[14]), and all the five most frequent words in which it appears are lexical; thirdly, the occurrences of /uː/ especially must be looked at bearing in mind that 42% of them appear in the word "you" – a function word with a potentially reductible vowel.

[14] This figure does not take into account the post-extraction changes made to the pronunciation of "to" – changes to "the" having been made at extraction run time, they do appear there.

All the differences mentioned above here and in the previous sections, of counts (*c.f.* section 2.1), of alignments (*c.f.* section 2.2), of standard deviations (*c.f.* section 2.3.1), of phonemic and lexical distributions (*c.f.* section 2.3.2 above) draw a complex picture which focusing on phonemic contrasts exclusively conceals. Whether these differences exert an influence on acquisition is a question which the rest of this work will try to answer. The next section addresses the issue of acoustic treatment, *i.e.* of normalization, in the case of a corpus featuring greatly varying numbers of occurrences for each monophthong.

2.4 Issues in normalization

This section discusses the utility of normalizing the data. After briefly introducing a few normalization methods and presenting the theoretical requirements underlying those methods, and how artificially constraining on a spontaneous speech corpus they may be, a procedure to assess the potential bias normalizing introduces when analyzing a learners' corpus.

2.4.1 Requirements of normalization

What vowel normalization method to adopt when dealing with skewed corpora? Common normalization methods such as Nearey (1978), Lobanov (1971), Wand & Fabricius (2002) are vowel-extrinsic, and require that acoustic measurements for all the vowels of a speaker's system be collected in roughly the same amount. Failure to do so will unduly skew the results, since each normalized formant value is dependent on all the other formant values either of the speaker (in the case of speaker-intrinsic methods), or of all speakers (in the case of speaker-extrinsic methods). However, such requirements hardly match the realities of language: phonemic differences are qualitative, i.e. categorical, but nothing obviates the possibility of a skewed distribution of phonemes in a language. In fact, such skewness is the norm, rather than the exception. Tambovtsev & Martindale (2007) have shown that phonemic

frequencies follow a Yule-Simon distribution in 95 languages. Besides, the distribution of phonemic frequencies in spoken or written corpora is not the same as the distributions in the lexicon. In their study on conversational American English, (Mines et al., 1978, p. 221) state that *"[t]he top ten phonemes (in order /ə, n, t, ɪ, s, ɾ, i, l, d, ɛ/) account for 47% of all the data"*. As to the English lexicon, John Higgins[15] finds that in the 1974 edition of the *Cambridge Advanced Learner's Dictionary*, the first ten most frequent phonemes are /ɪ, t, s, n, ə, l, r, k, d, z/, in order, and that they account for 60.29% of all phonemes ; another source[16] compiling data from the Carnegie Mellon University Pronouncing Dictionary along with Adam Kilgarriff's unlemmatized frequency list for the British National Corpus lists /ə, n, r, t, ɪ, s, d, l, i, k/ as the first ten most frequent phonemes, which account for 58.48% of all phonemes.

There therefore exists a structural mismatch between the requirements of vowel normalization and the realities of language. In these conditions, it seems impossible to study vocalic realizations in spontaneous speech: whatever the chosen corpus, the numbers of occurrences of each vowel are bound to be unevenly distributed, and the normalized values of vowels, in the case of the most common vowel-extrinsic methods, are bound to be skewed. However, such a mismatch raises another question in turn: could it be that by comparatively inflating the numbers of occurrences of the rarest phonemes and decreasing those of the most frequent ones, normalized values themselves provide a skewed and inaccurate representation of the vowel space? Most studies (e.g. Hillenbrand et al. (1995), Ferragne & Pellegrino (2010), Clopper et al. (2005)) resort to lists of words, with vowels usually embedded in the same consonantic template /hVd/. But could it be that normalization might lead to increasing or decreasing contrasts unduly? These methods can be argued to overlook the role of phonological neighbourhood density and frequency: in English, there are 466 minimal pairs distinguishing /ɪ/ from /iː/ – 18 for /ʊ/ and /uː/ (morphosyntactic variations included).

[15]http://myweb.tiscali.co.uk/wordscape/wordlist/phonfreq.html, retrieved on June 4, 2014.

[16]http://cmloegcmluin.wordpress.com/2012/11/10/relative-frequencies-of-english-phonemes/, retrieved on June 4, 2014.

Are speakers not aware of the necessity to enhance contrasts in high-density words, the extent of the enhancement being itself dependent on the discourse context? Words with high-density neighbourhoods have been shown to be processed differently from low-density ones in children, adults and aphasic speakers. In all these cases, facilitating effects have been observed (cf. e.g. Middleton & Schwartz (2010)). In the case of frequency, phonetic details are processed by adults in a finer-grained way in high-frequency words than in low-frequency ones (White et al. (2013)). Besides, prosodic positions have been shown to influence the realization of phonemes (Keating et al. (2004)); phonemic processing and speech-errors likewise depend upon phonological neighbourhood density and clustering coefficients (the similarities between phonological neighbours, Chan & Vitevitch (2010)). Phonemic realizations are therefore highly likely to depend upon super-phonemic parameters which normalization methods somehow force out of consideration. Focusing, as so many studies have done, on the /hVd/ template in experimentally balanced corpora, increases the likelihood of overlooking parameters which may turn out to be crucial in understanding speech production and perception. However, multiplying parameters makes cross-comparisons impossible, and decreases the likelihood to account for the stable nature of phonemes without which communication would be impossible. What will be tentatively studied here is the relevance of using normalization methods without leveling out the numbers of occurrences of each vowel. The issues mentioned above are of course compounded by the fact that the phonemes under study here are non-native, and are therefore unstable in nature: within-speaker variations for a given phoneme, possibly even for a given word, are to be expected, and this dispersion must be taken into account in order to assess phonemic acquisition.

In order to illustrate the issue further, and to summarize the findings from sections 2.1 and section 2.3.1, figure 2.12 plots the F_1 and F_2 standard deviations, taken at mid-temporal values for each monopthong and each aligner (SPPAS in black and P2FA in grey) using the main learners' corpus (top panel) and the natives' subcorpus (bottom panel). The symmetry

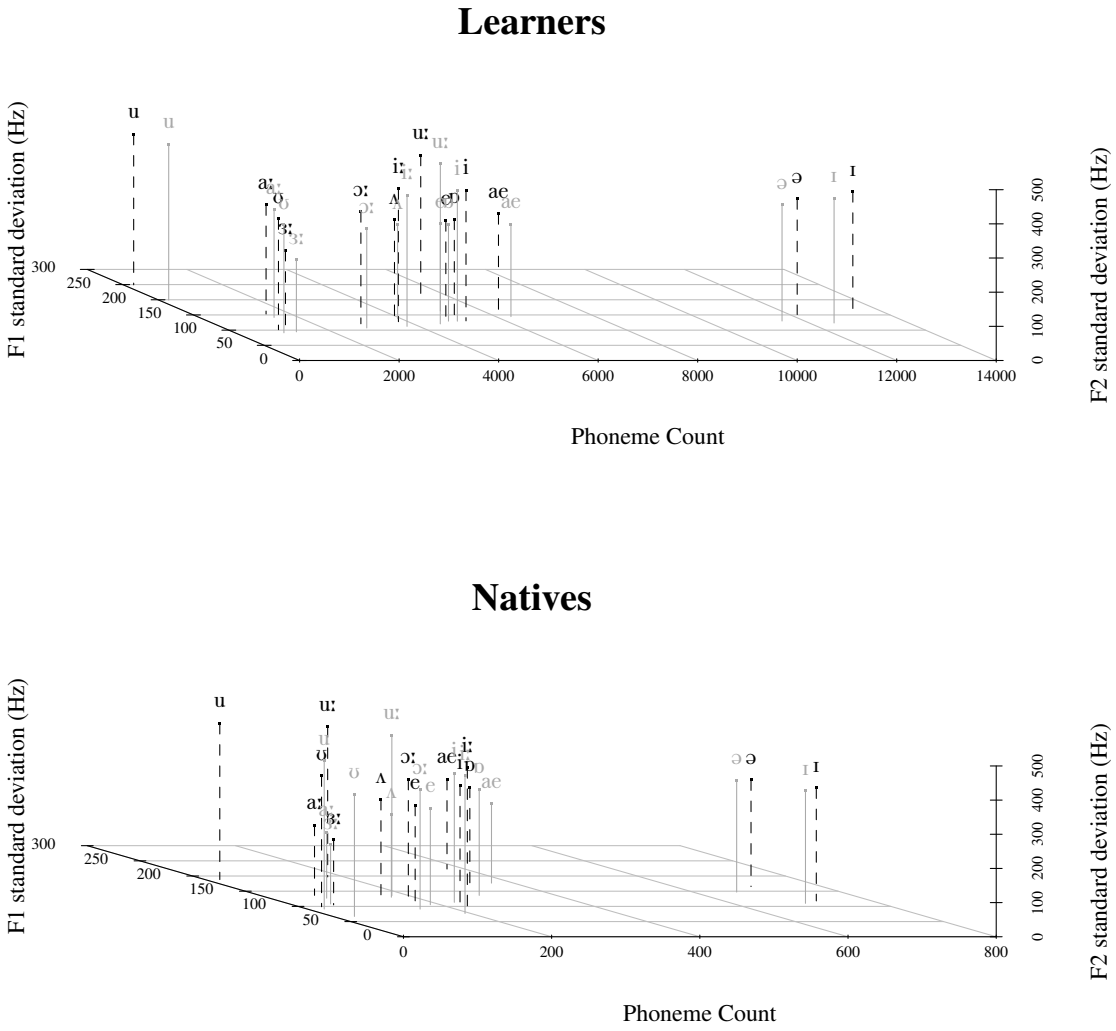


Fig. 2.12: Scatterplot of the mid-temporal F_1 & F_2 standard deviations of monophthongs against their number of occurrences. Black: SPPAS-aligned data; grey: P2FA-aligned data; *top panel:* main learners' corpus; *bottom panel:* natives' subcorpus.

between the two corpora, is once again all the more surprising as the numbers of occurrences vary greatly from one corpus to the other. Such similarities, in both the distributions of the standard deviations and in the proportions of the different monophthongs with respect to the overall count may offer a solution to the harmonisation required by normalization methods: the fact that the respective categories in the learner data come in proportions similar to the native data that no normalization requiring even numbers of tokens in each phonemic

category is needed. The view is even held here that such a normalizing procedure would introduce a counter-productive bias, in that the natural skewness of phonemic distributions found in spontaneous speech should be preserved, as it is contended it is bound to exert influence on the acquisition of phonemic contrasts.

2.4.2 Phoneme-gating

This section studies the impact of normalizing acoustic data when attempting to assess phonemic acquistion. A simple method, called "phoneme-gating", is proposed: phonemes with formant values inferior or superior to the respectively maximum and minimum values of the corresponding phonemes from a native data set (here Peterson & Barney Peterson & Barney (1952)) are then sorted according to whether they meet theseminimal and maximal requirements. The procedure is applied to the datasets with four different methods of normalization: Traunmüller's Bark method (Traunmüller, 1990); the Bark Difference Metric (Syrdal & Gopal, 1986) (henceforth, BDM); Nearey's extrinsic method (Nearey, 1978); and Lobanov's method (Lobanov, 1971). For these calculations, the mid-temporal values of each formant were adopted. The equations for each procedure of normalization are the following (where F_i^v is the i^{th} formant of a vowel v):

1. **Bark:** $Z_i^v = \frac{26.81}{(1+1960/F_i^v)}) - 0.53$
2. **Bark Difference Metric:** $Z_{1/2}^v = (\frac{26.81}{(1+1960/F_3^v)} - 0.53) - (\frac{26.81}{(1+1960/F_{1/2}^v)} - 0.53)$, where $F_{1/2}^v$ is vowel v's F_1 or F_2.
3. **Nearey Extrinsic:** $Z_i^v = log(F_i^v) - (\mu_1^L + \mu_2^L + \mu_3^L)$, where μ_1^L, μ_2^L and μ_3^L are the log-means of the F_1, F_2 and F_3 values of all vowels.
4. **Lobanov:** $Z_i^v = \frac{F_i^v - \bar{x}_i}{\sigma_i}$, where $\bar{x}_i$ is the mean of *all* the speaker's i^{th} formant values, and σ_i their standard deviation.

Other methods of normalization exist (cf. Adank et al. (2004) for a review), but these four were chosen because of their differences in their pre-requisites. Normalizing a given formant

of a given vowel may or may not, depending on the method chosen, require data outside this particular formant of a particular vowel. If no data outside of the formant under study is needed for normalizing, the formant, the method is formant-intrinsic (*i.e.* collecting F_2 data, for instance, is not necessary to normalize F_1) – formant-extrinsic otherwise. Likewise with vowels, if studying a given vowel does not require collecting acoustic data on other vowels, then the method is vowel-intrinsic – vowel-extrinsic otherwise. These differences create four different categories of normalization methods: whether they are formant- or vowel- extrinsic or intrinsic. There also exist speaker-extrinsic methods (*cf.* Morrison & Nearey (2006) or Labov et al. (2006)). In our corpus however, the number of occurrences for each monophthong varied greatly from one speaker and one session to another. All normalizing procedures described below are therefore speaker-*intrinsic*. Each method retained here is representative of one of these four categories, as shown in table 2.2. The computations were made using the statistical software R R Core Team (2015), and the PhonTools Barreda (2014) package for the last two methods, Lobanov and Nearey 2. The procedure experimented to compare various

Table 2.2: Reminder of the specificities of the normalization methods.

Method	Vowel	Formant
Bark	Intrinsic	Intrinsic
BDM	Intrinsic	Extrinsic
Lobanov	Extrinsic	Intrinsic
Nearey 2	Extrinsic	Extrinsic

methods of normalization applied to our skewed corpus was the following: the data from Peterson & Barney Peterson & Barney (1952)), which comes with the PhonTools Barreda (2014)) package, was first normalized using the four methods of normalization. The first two methods are not parts of the PhonTools package, but only require simple operations to be applied on raw values. Before proceeding any further, let it be emphasized that this method does **NOT** offer any objective insight on acquisition or pronunciation accuracy[17] *per se* –

[17] During the research phase, the same experiment was carried out on the native speakers' subcorpus: the proportions of phonemes with within-range formants were on average lower than with the learners' corpus... One possible explanation could be that the native speakers' subcorpus is made up of recordings of

rather, it aims at showing that the results one may obtain, and the conclusion one may draw, can be different whether one method of normalization or another is chosen. The procedure to compare the normalization methods was the following. Firstly, sex-dependent minima and maxima of the F_1, F_2 and F_3 formant values for each vowel and each method of normalization were calculated from the Peterson & Barney data. As their data consist of recordings of 10 vowels *i.e.* /æ/, /ɑː/, /ɛ/, /ɜː/, /ɪ/, /iː/, /ɔː/, /ʌ/, /ʊ/ and /uː/, these calculations returned 10 (speaker-independent) minimal and maximal values for each formant, for each of the two sexes and for each normalization method. Subsets of the main corpus were then created by sex and session, and the F_1, F_2 and F_3 mid-temporal values of each datapoint of these subsets was normalized in turn, following the four methods. Finally, the obtained normalized values were then checked against the corresponding (*i.e.* by formant, sex and normalization procedure) minimal and maximal values of the native Peterson & Barney data. The idea behind gating learners' normalized formant values is twofold: *(i)* assess the influence of normalization methods on acoustic analysis – if normalization methods return varying results, then conclusions are not so much data-driven as method-driven; *(ii)* explore whether a normalization-independent method existed which might shed light on phonemic acquisition.

The counts and proportions of vowels whose formant values met the minimal and maximal requirements for the corresponding normalization procedure are presented in figure 2.13. One clear trend emerges from the figure: vowel-extrinsic methods of normalization (*i.e.* Nearey extrinsic and Lobanov) return a higher number and a higher proportion of gated[18] phonemes, 65.3% and 56.3% respectively, than vowel-intrinsic procedures, with 31.7% for Bark and 42.7% for BDM in the case of the SPPAS-aligned data. For the P2FA-aligned data, the proportions are 33.4%, 45%, 66.7% and 57.4% for the Bark, BDM, Nearey extrinsic

various accents (British, Scottish, Irish and American, *c.f.* section 1.3, which in all likelihood increases the dispersion of formant values.

[18] From now on, a phoneme whose formant values fall within the range of minimal and maximal values defined by the Peterson & Barney data will be refered to as "gated".

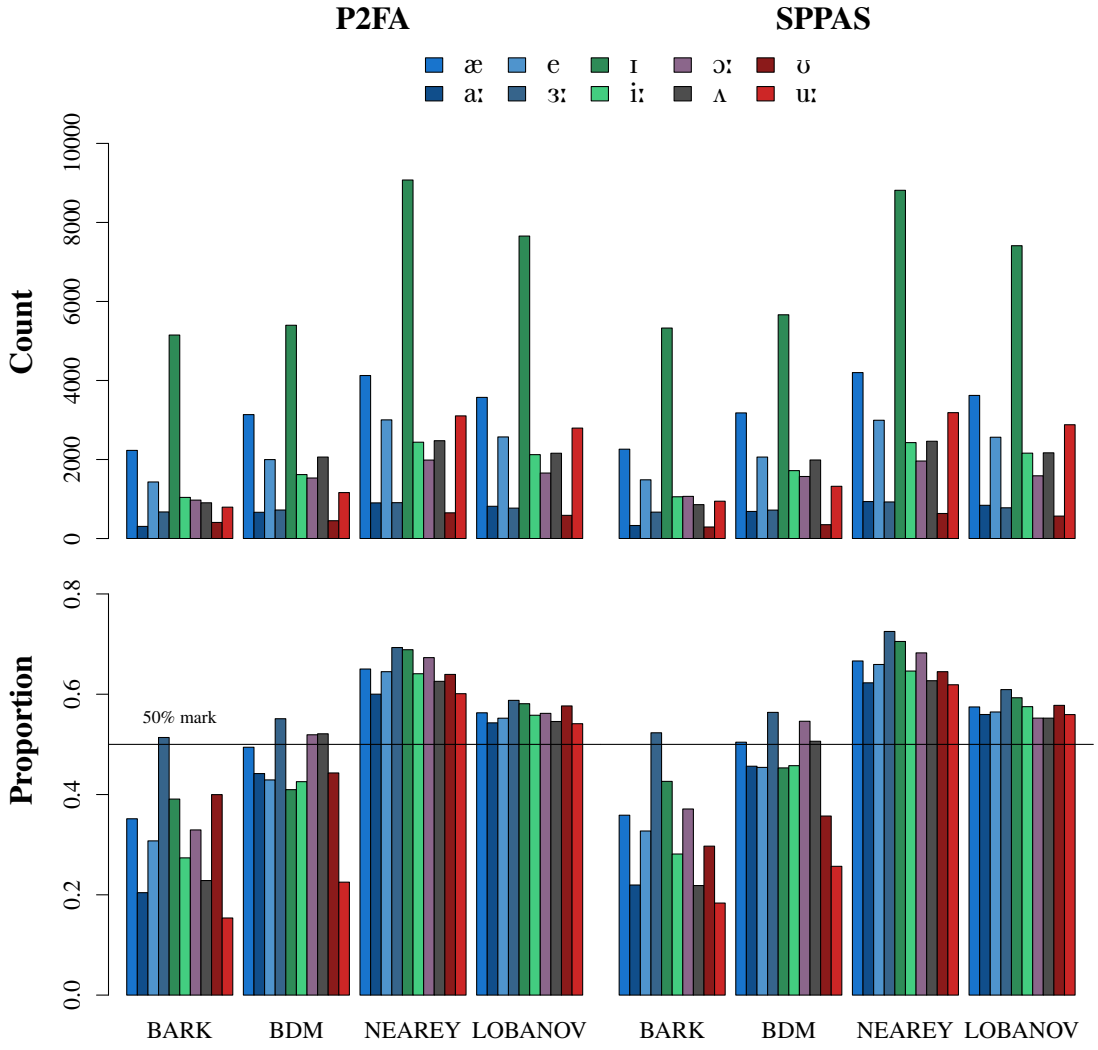


Fig. 2.13: Per-normalization method counts and proportions of SPPAS-aligned (left panel) and P2FA-aligned (right panel) phonemes meeting the minimal and maximal requirements from the native Peterson & Barney data.

and Lobanov methods respectively. Great differences from one phoneme to another can be observed too: /ɜː/ is consistently gated, with a mean proportion across all normalization methods of 58.6% of within-range values for the SPPAS-aligned data, and 60.5% for the P2FA-aligned data. All other phonemes feature proportions around 50%, between 48% and 52%, except /ɑː/ (46.4% and 46.5% for the SPPAS-aligned and the P2FA-aligned data respectively), and /uː/. /uː/ is the phoneme with the lowest proportion of gated values, with 38% and 40.5% for the two aligners respectively.

In the more global perspective of this work, which aims at establishing whether the acquisition of /ɪ/-/iː/ and /ʊ/-/uː/ follow similar patterns, these findings, and especially the gap in numbers of gated values for /uː/ and the other phonemes, are only evidence to a limited extent. The main issue is that an analysis may differ considerably depending on the normalization method that was chosen[19]. To test this statement, the same experiment was carried out using the native speakers' subcorpus as the reference values, *i.e.* replacing the Peterson & Barney data. To keep some consistency, only the native British speakers, 4 women and 4 men, were retained – for a total of 1,038 monopthongs for female speakers, and 263 for male speakers. The greatest difference, however, is that the numbers of each individual monophthong vary from one category to the other (*c.f.* figure 2.10). Once again, the minimal and maximal values across speakers of the same sex were stored for each formant. The numbers and proportions of gated monophthongs using these new reference values are presented in figure 2.14. The counts and proportions of gated phonemes are greater than when using the Peterson & Barney data. The remarks about how stricter and more exclusive than vowel extrinsic ones vowel-intrinsic methods of normalization are seem to hold. Two phonemes stand out because of their low proportions of gated datapoints, /ɜː/and/ʊ/[20]. The high proportion of gated datapoints in the case of /ɜː/ is all the more surprising as the monophthong was the most gated one with the Peterson & Barney data (*c.f.* above). /ʊ/ features comparatively lower proportions in the P2FA-aligned data than in the SPPAS-aligned data (46.9% against 51.5%).

The overall results are summarized in figure 2.15, with a per-session breakdown. The top row shows the mean proportions of gated phonemes using the Peterson & Barney (P & B) data as reference. The bottom row shows those proportions using the native speakers' corpus (NSS) as reference, restricted to its British elements[21]. In the top row, possibly the most striking feature is that the differences between the two aligners are minimal. As to

[19] Note that the effect of the aligner seems comparatively limited.
[20] Bark-normalized /æ/ could also be worth mentioning, but it is only an outlier using the Bark method.
[21] As mentioned in the caption of figure 2.15, the scales between the top and bottom rows vary.

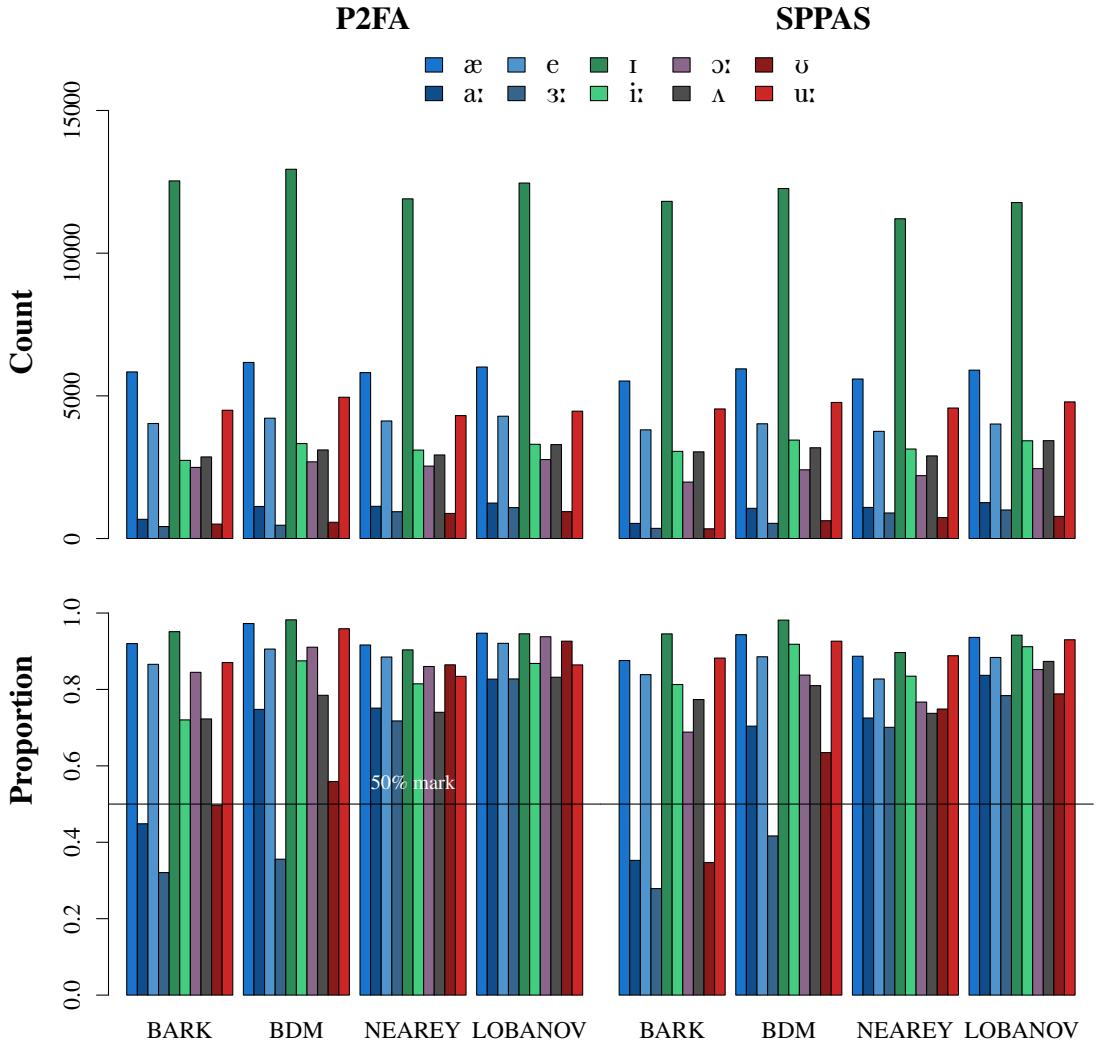


Fig. 2.14: Per-normalization method counts and proportions of SPPAS-aligned (left panel) and P2FA-aligned (right panel) phonemes meeting the minimal and maximal requirements from the native (British) speakers' subcorpus.

the differences between normalization procedures, they seem to consist of simple vectorial translations, except for the dip between session 3 and 4 in the case of BDM. In the bottom row, the differences between the two aligners are more visible, with the *caveat* that the y-axis covers a much shorter range ($y \in [0.6, 1]$ against $y \in [0.1, 1]$) and that the proportions are much higher using the native speakers' subcorpus as reference than using the Peterson & Barney data. This smaller range makes the permutations in the orders of normalization methods (Lobanov being the procedure with the highest proportions of gated phonemes,

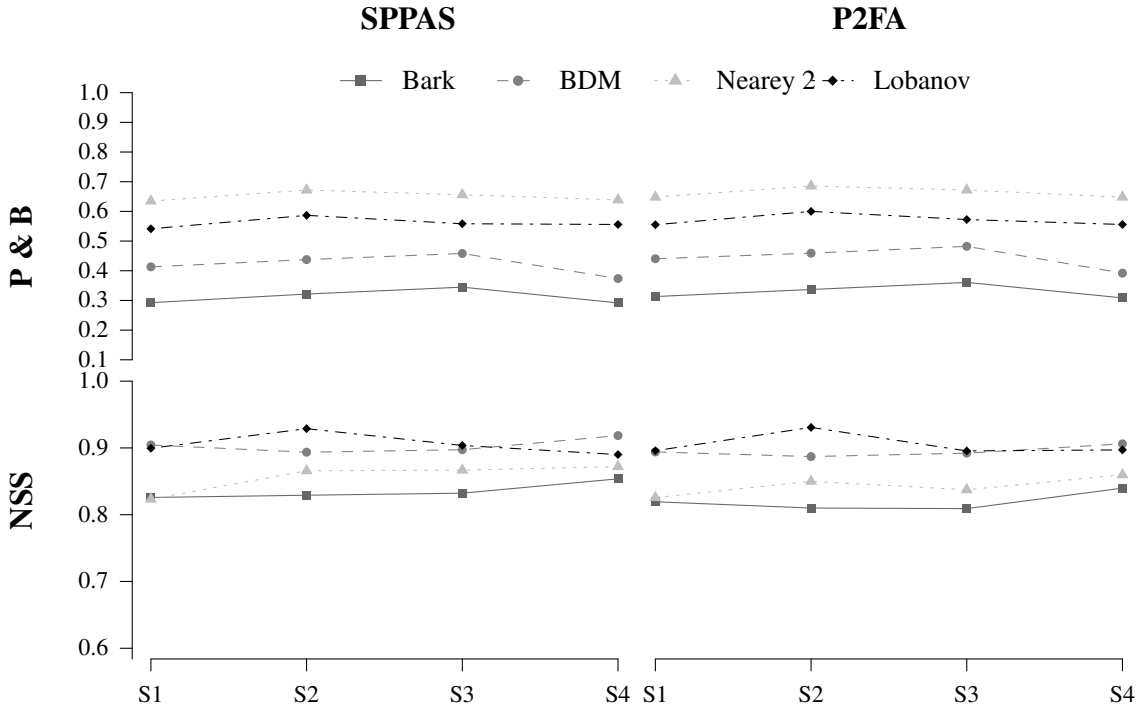


Fig. 2.15: Per-normalization method, per-session means of proportions of gated phonemes. *Top row:* using the Peterson & Barney (P & B) data as reference; *bottom row:* using the (British) native speakers' subcorpus (NSS). **Scales vary.**

and BDM being a close second) of relative importance, even though Bark does remain the procedure with the lowest proportion of within-range phonemes regardless of the aligner and of the referecne data. The results using the Peterson & Barney data are split along the 50% mark: they are on average below the mark for vowel-intrinsic methods, and above for vowel-extrinsic methods – a crucial threshold when attempting to assess acquisition. Two things remain to explain, if only tentatively: *(i)* the overall higher proportions of gated phonemes when normalizing the data with vowel-extrinsic methods; *(ii)* the higher proportion when the native speakers' subcorpus is used. The first point, and, to some extent, probably the second point as well, may be explained by the fact that a bias is introduced by the overwhelmingly numerous occurrences of /ɪ/[22]. The overall results in the case of vowel-extrinsic methods

[22] Another set-up for the experiment, which was actually tested, could have been to equalize the number of phonemes for each speaker and each phoneme, as per the requirements of the normalization procedures. The question then arises of which phonemes to select? With such a low number of /ʊ/ and such a huge number of /ɪ/ (*c.f.* figure 2.1), the total count of monophthongs per speaker was bound to be determined by the lowest number of occurrences of /ʊ/ (this is not even speaking about per-speaker, per-session numbers). This means

will heavily depend on the accuracy and dispersion of the most frequent phonemes. This of course also holds if the reference data, as is the case with the native speakers' subcorpus, itself features a bias matching the tested data. Note that this is the likeliest explanation to the second point: a tempting justification to the greater proportions of gated phonemes in the case of the native speakers' subcorpus could be that since the reference data is spontaneous speech, then dispersion is higher, and the minimal and maximal formant values of each monophthong spread across a greater range. Interestingly, this is actually wrong: the Peterson & Barney data has standard deviations at 201Hz, 637Hz and 519.5Hz for F_1, F_2 and F_3 respectively (regardless of sex). The SPPAS-aligned data has corresponding standard deviations of 172Hz, 418Hz and 304Hz – 155Hz, 427Hz and 300Hz for the P2FA-aligned data.

In conclusion to this section, it may be asserted that normalization methods, especially vowel-extrinsic ones, at least when applied to learners' data and spontaneous speech data[23], distort the data in a way that may drastically change the analysis. For this reason, and those mentioned in footnote 22 and section 2.4.1, the adopted solution here is to resort either to raw values[24], or to the BDM-normalized values – these values presenting the comparative advantage of including F_3. For these reasons, vowel-extrinsic normalization methods will not be used in the rest of this work. The procedure used to test the effect of normalization, phoneme-gating, is probably not without flaws itself, at least in the way it was implemented: speakers' idiosyncrasies, including the reference speakers', were not taken

that *all* occurrences of /ʊ/ would have been selected on the one hand, while a wealth of /ɪ/ remained on the other. Random selection of a given number of occurrences for each monophthong was a solution that was tried, but the problem was that formant values *never* converged even after thousands of random selection loops. The results obtained would have therefore been totally random. This requirement that the numbers of occurrences of each monophthong should be the same before normalizing is at the core of our contention towards normalization: it is our view that the frequencies of phonemes vary greatly, and that combined with word frequency, they form a complex and skewed system which normalization artificially distorts by equalizing and neutralizing frequencies. A bias, in favour of the rarer phonemes, is therefore introduced, especially in the case of vowel-intrinsic methods. The bias that our experimental design introduces is in our view the lesser of two evils, as at least it is a bias which can be contended (*c.f.* figure 2.9 and figure 2.10) to be present in spontaneous speech.

[23] The combination of these two factors increases skewness in occurrences and dispersion in formant values.

[24] Bark values can somehow be considered raw too, as they are simply translated from raw values. One considerable advantage of using Bark data is that emulations of Bark-based experiments can be attempted.

into account. Further research should carry out these tests speaker by speaker, in order to confirm these concluding statements, and establish the robustness of phoneme-gating. The next section offers a method to assess the acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts, by investigating the relationship between the Euclidian distance in the vowel space of each contrast's vector and the surface of the vocalic trapezoid.

2.5 Contrasts and vowel space

This section investigates whether the locations and distributions of all the monophthongs in the vowel space may yield useful information about the acquisition of a contrast. More specifically, is there an observable, consistent relationship between the specific locations of the contrastive vowels and the rest of the vowel space? This question is theoretically motivated by the Theory of Adaptive Dispersion (henceforth, TAD; Liljencrants & Lindblom (1972); Lindblom (1986)), which states that vowels in a given space are located in such a way as to maximize contrasts and facilitate perception. This work is agnostic as to whether the TAD prediction that the vowel space increases with the number of vowels of the language: for instance, Al-Tamimi & Ferragne (2005) study French and two varieties of Arabic, and show the prediction is likely to be borne out, whereas Gendrot & Adda-Decker (2007), using the vowel inventories of eight languages, and do not find larger vowel spaces for languages with greater counts of vowels. This agnosticism does not entail that the relationship between vowel space and vowel inventory should not be investigated in SLA. After all, in order to acquire the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts, French learners of English need to create some space in order to add sounds to their vowel inventory. It therefore seems legitimate to see if their vowel space evolves in time. For these calculations, the vowel spaces were calculated using mid-temporal formant values taken from the main corpora (SPPAS-aligned and P2FA-aligned). The values were then normalized using the BDM method (*c.f.* section 2.4. The vowel space is here defined as the area of the convex hull formed by the outermost vowels on

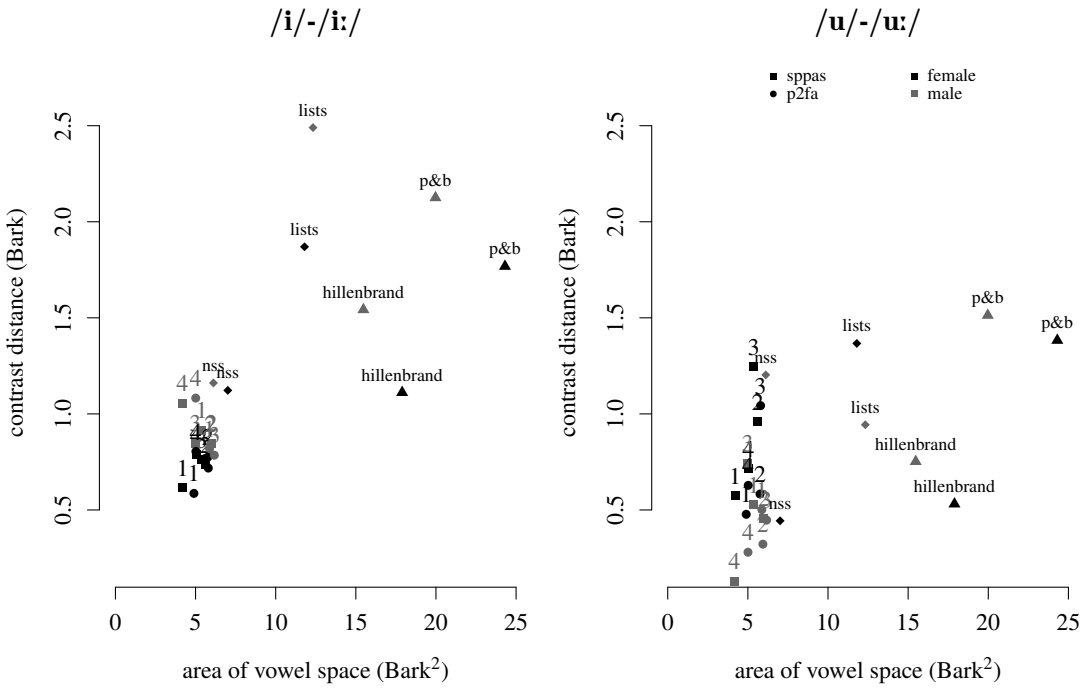


Fig. 2.16: Evolution across all sessions of the distances of the SPPAS-aligned and P2FA-aligned /ɪ/-/iː/ and /ʊ/-/uː/ contrasts, against the vowel space, measured as the polygonal area formed by the outermost vowels on the BDM F_1/F_2 axes. *Digits:* session numbers; *NSS:* (British) native speakers' subcorpus; *Lists:* subcorpus of lists of words; *P&B:* Peterson & Barney (1952) data; *Hillenbrand:* Hillenbrand et al. (1995) data.

the F_1/F_2 axes, as was presented above in figure 2.5. As a reminder, in the BDM method, F_1 values are calculated by subtracting the F_1 value in Bark from the F_3 value on each datapoint; likewise for F_2. The reason why the BDM method was used is that the results using it are very similar to those using Bark values, while factoring in more information, *i.e.* F_3[25].

The results are presented in figure 2.16, which plots the Euclidean distances of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts against the area of the vowel space. Values for female speakers are represented in black, and in dark grey for male speakers. Three native sets have been used for purposes of comparisons: the Peterson & Barney (1952) data; the Hillenbrand et al. (1995) data; and the NSS, with vowels pronounced by the British speakers only. The first

[25] In the case of TAD as in so many other instances, studies do not use standardized procedures. For instance, Jongman et al. (1989) use Hertz F_1, F_2 and F_3 values; Bradlow (1995) uses Hertz F_1 and F_2; Al-Tamimi & Ferragne (2005) use Bark F_1 and F_2; Gendrot & Adda-Decker (2007) use F_0-F_1 *x* F_2-F_3 on a Bark scale.

observation is that the type of vowel production (lists of words *vs.* spontaneous conversations) seems to have an effect on the size of the vowel space: both the Hillenbrand et al. (1995) and Peterson & Barney (1952) data feature the biggest polygonal areas (on the *x*-axis), but, rather surprisingly, the learners' lists of words rank third, in front of the British speakers' corpus. SPPAS-aligned and P2FA-aligned values show very few significant differences, as has often been the case so far. Values for the /ɪ/-/iː/ contrast distance are remarkably more consistent than for /ʊ/-/uː/. This consistency may well indicate awareness among learners of the targets to aim for in the case of the /ɪ/-/iː/ contrast, while the results for /ʊ/-/uː/ are much more chaotic. One argument supporting this assumption of a greater awareness of the /ɪ/-/iː/ target contrast distance is supported by the fact that sessions 4 are the sessions with the highest contrast distances; as time went by, the values became closer and closer to native values, especially when looking at NSS values in female speakers. Another argument comes from the values of the subcorpus of lists of words. The comparatively high /ɪ/-/iː/ contrast distances may indicate over-correction: the differences between the two sounds were exaggerated. A counter-argument to this hypothesis is that the formant values in the corpus have the highest standard deviations across all corpora (*c.f.* section 2.3.1). Another note-worthy observation is that distributions across the *x*-axis for either contrast remain roughly the same, *i.e.* the vowel space does not seem to expand over time. The extent to which acquisition depends on the expansion of the vowel space in order to acquire new contrasts could be a venue of research to explore: for instance, Iverson & Evans (2009) showed that new contrasts were easier to acquire the bigger the vowel inventory in the L1 was. Assuming a connection between vowel space and vowel inventory (with the *caveats* mentioned in the first paragraph of this section), one key of phonological and phonetic teaching might therefore be to work on the expansion of the vowel space when either the L1 vowel inventory or the L1 vowel space is smaller than in the L2: the surface of the French vowel space based on the Gendrot & Adda-Decker (2005) data are 8.05 $Bark^2$ for female speakers, and 5.69 $Bark^2$ for male speakers; using the French

subcorpus, with per-sex, per-phoneme averaged formant values, the surface of the vowel space is 7.19 Bark^2 for women, and 5.83 Bark^2 for men – *i.e.* in either datasets, the surface of the vowel space is smaller than the English native ones. The corpus of Englishspeakers with the smallest surfaces is the NSS, with 7.01 Bark^2 for women, and 6.12 Bark^2 for boys.

2.6 Conclusion

This chapter looked at the specifics of phonemic data regardless of speakers' idiosyncrasies. It was shown that monophthongs feature differences in counts (*c.f.* section 2.1), per-centile extraction quality (*c.f.* section 2.2), differences in standard deviations (*c.f.* section 2.3.1, and frequencies in the words where they appear (*c.f.* section 2.3.2). This heterogeneity, which stems from the very nature of the corpus (*i.e.* learners' spontaneous speech) makes it difficult to apply vowel-extrinsic normalization methods (*c.f.* section 2.4). Finally, it was tentatively proved (*c.f.* section 2.5) that there may well be a relationship between Euclidean distances of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts and the vowel space.

Nonetheless these complexities, induced by the nature of the corpus and by the sheer amount of methods available to process the data, all seem to converge towards deep-rooted differences in the acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts. But the speaker-independent procedures implemented in this chapter may also conversely have contributed to creating artificial differences which obscured similarities. The only way to find out whether this is the case is by looking at the data speaker by speaker – this is what the next chapter undertakes.

Chapter 3

Speaker-dependent analyses

After analyzing the data from the main learner corpus from a systemic and holistic point of view, the focus in this chapter is on the learners' idiosyncrasies, and the specific evolution of their pronunciations of monophthongs, and more particularly of /ɪ/, /iː/, /ʊ/ and /uː/. Two main concerns will underlie the various studies undertaken here. The first concern regards the existence, or not, of cross-speaker patterns of evolution. Can somehow the ability to emulate native-like sounds be predicted? how fluctuating is it from one speaker, one session, one vowel, to another? In order to answer these questions, and in the face of the wealth of data that was collected (and only part of which wich will be exploited in this work), choices of how to process the data had to be made. These choices are at the heart of the second concern: the risks are high that the methods used to select and analyze the data condition the results more than the data themselves. In other words, the likelihood to have the results sway one way or another according to what pieces of information were selected, and how those pieces of information were modelled, exists. In order to preserve neutrality, how the data are going to be processed, and why they are going to be processed in a given fashion, is specified beforehand. Let it be clear that the conclusions drawn at the end of the previous chapter are here maintained: formant values are BDM-normalized because this method of normalization integrates F_3 values in a two-dimensional manner (F_1 and F_2). But

measurements other than mid-temporal ones will be explored, in an attempt to capture as much of the original signal as possible, and compare this information to native values when possible.

After a few preliminary remarks about the selected datapoints and their distribution, this logic of endeavouring to retain as much information as possible is first applied to the study of vowel-inherent spectral changes. By focusing on the offset and onset of the vowel, this theory challenges the traditional approach based on mid-temporal measurements. In section 3.2, it is applied to the main learner corpus and the NSS. Native and learner values are compared, along with their standard deviations. In section 3.3, a machine-learning algorithm, the *k*-Nearest Neighbours is run on the corpus in order to explore the extent to which this classification method manages to categorize the learners' monophthongs accurately. Because running the algorithm several times on the same data does not return the same classification results, and because the training set used, *i.e.* the NSS, is different from the test set, making cross-validation impossible, a procedure is devised to figure out the optimal *k*, *i.e.* the optimal number of nighbours enabling the highest proportion of classification accuracy. Section 3.4 is an attempt to model longitudinal acquisition by mixing continuous and categorical predictors. Linear mixed-effect regressions provide the mathematical framework to do just that, and different longitudinal models are compared. Finally, an attempt at studying the entire signals is made by modeling them using discrete cosine transforms in section 3.5. Once again, the learners' datapoints are compared to the natives' using another type of classification method, quadratic analysis. This method is ultimately applied to models based on both discrete cosine transforms and mid-temporal formant values, in order to establish their comparative added values.

3.1 Preliminary remarks

Because not all speakers in the main corpus took part in the last session, during which the learners also read the lists of words (*c.f.* section A.2.1) and a text in French (*c.f.* section A.2.2), this chapter will only investigate the specific evolution of the acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts for students who took part in all sessions. The reasons behing this reduction of the numbers of speakers are the following: *(i)* in terms of longitudinal study, a fourth session makes the data more robust, and the analyses of possible evolutions more reliable. *(ii)* If the idiosyncrasies of the learners' pronunciations of vowels are to be assessed, comparing their vowel productions in spontaneous conversations with their vowel productions in French and in lists of words is extremely useful – if not necessary for the accuracy of the assessment. It therefore makes more sense to focus on the learners for whom such data is available.

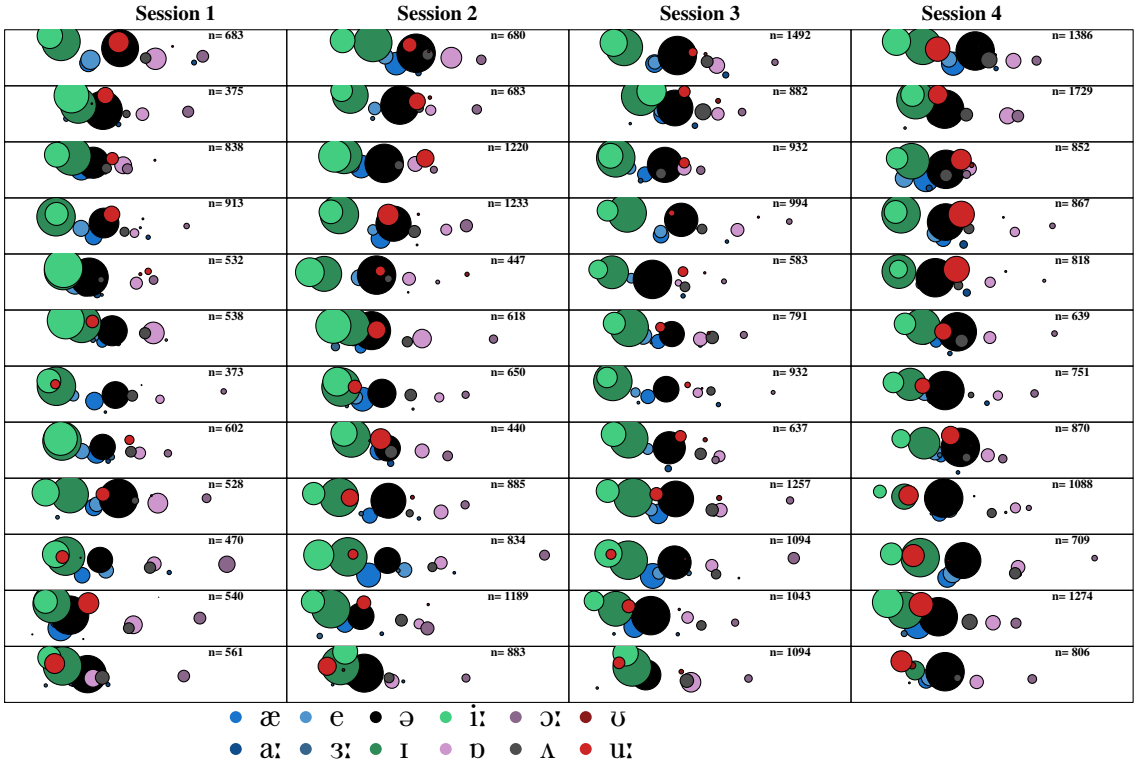


Fig. 3.1: Per-session, per-speaker mean F_1/F_2 values for monophthongs in the BDM-normalized vowel space; the size of the circles is proportional to the number of occurrences (sizes are relative). The total number of monophthong for each session is indicated in each panel.

The calculations will therefore be made for 9 female speakers and 3 male speakers (*c.f.* section 1.1.1 for a reminder of the distribution of the participants). In order to present readable results, the default aligner used for the data is SPPAS. Very few differences, if any, appeared in the previous chapter between the two aligners. All *R* scripts, however, are P2FA-ready, and interested readers may run them to see for themselves if any significant variations between the two aligners might have been overlooked. The datapoints used throughout this chapter come from the NSS, and the main learner corpus. When the focus of the different experiments is on specific phonemes (most frequently on /ɪ/, /iː/, /ʊ/ and /uː/), the subsets are taken from those two datasets. Because they form the basis of the study, it seems in order to have a look at how the monophthongs are broken down across speakers and sessions. Figure 3.1 plots the average mid-temporal, BDM-normalized F_1 (y-axis) and F_2 (x-axis) formants for each speaker and each session. The size of the dots of each monophthong is proportional to its number of occurrences. The same graph for native speakers can be found in figure 3.2. The difference is that each panel does not provide the values of individual speakers, but rather the average values of native speakers of the same sex and accent. Less standard accents present in the NSS (*c.f.* section 1.3), such as the Irish and the Scottish accent, will not be used as reference points in this chapter. Taking a look at the two figures (figure 3.1 and figure 3.2) reveals an arguably very similar distribution of monophthongs: two categories dominate the number of occurrences, /ə/ and /ɪ/. Other dominant phonemes, although to a lesser extent, and in that order, are /iː/, /uː/ and then /æ/. These figures make a defining feature of spontaneous speech (already described at length in the previous chapter) clear: the distributions of tokens are unequal. This unavoidable characteristic is both an advantage in that it most likely embraces the seemingly chaotic structure of natural spoken language, and a disadvantage in that a bias is necessarily introduced when resorting to classification algorithms. How well those algorithms manage to classify underrepresented monophthongs such as, crucially for this study, /ʊ/, is a major point of interest.

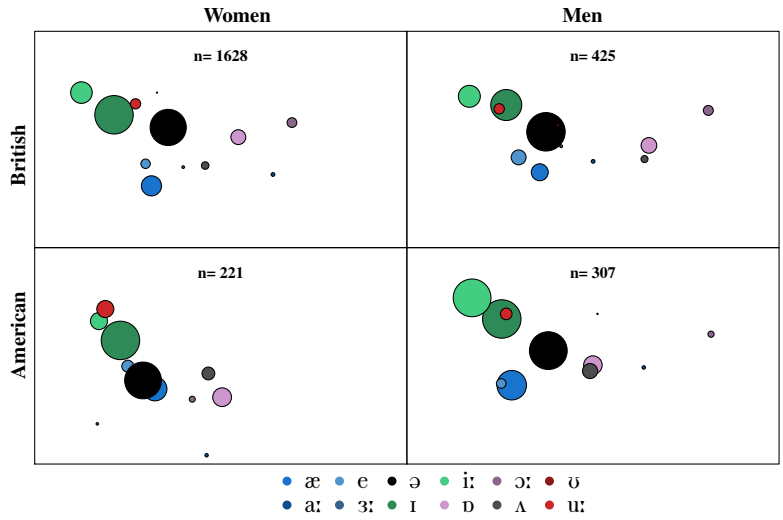


Fig. 3.2: Native speakers' mean F_1/F_2 values for monophthongs in the BDM-normalized vowel space; the size of the circles is proportional to the number of occurrences (sizes are relative). The numbers in each panel correspond to the total number of monophthong.

After these preliminary remarks, it is now time to turn to the first speaker-dependent analysis: vowel-inherent spectral changes.

3.2 Vowel Inherent Spectral Change

Vowel Inherent Spectral Change is the theory that states that vowel quality is a function of the trajectories of formants throughout the duration of the vowel rather than their means or mid-temporal values (*c.f.* Nearey & Assmann (1986), Nearey (2012), Assmann et al. (2012) and the references therein). Morrison (2012) in particular shows how the onset+offset model, *i.e.* the analyses of F_1 and F_2 both at the beginning and at the end of the vowel's duration, provides a better account of how vowels are perceived. Vowels are identified by their vector in the F_1/F_2 space, with measurements often taken at 20% and 80% of the duration. Investigating VISCs in learners' productions may therefore shed light on the clarity and quality of their vocalic pronunciations.

Figure 3.3 provides an example of VISC across all four sessions for speaker DID0014. For each vowel, the initial and final coordinates of the vector in the F_1/F_2 space corresponds

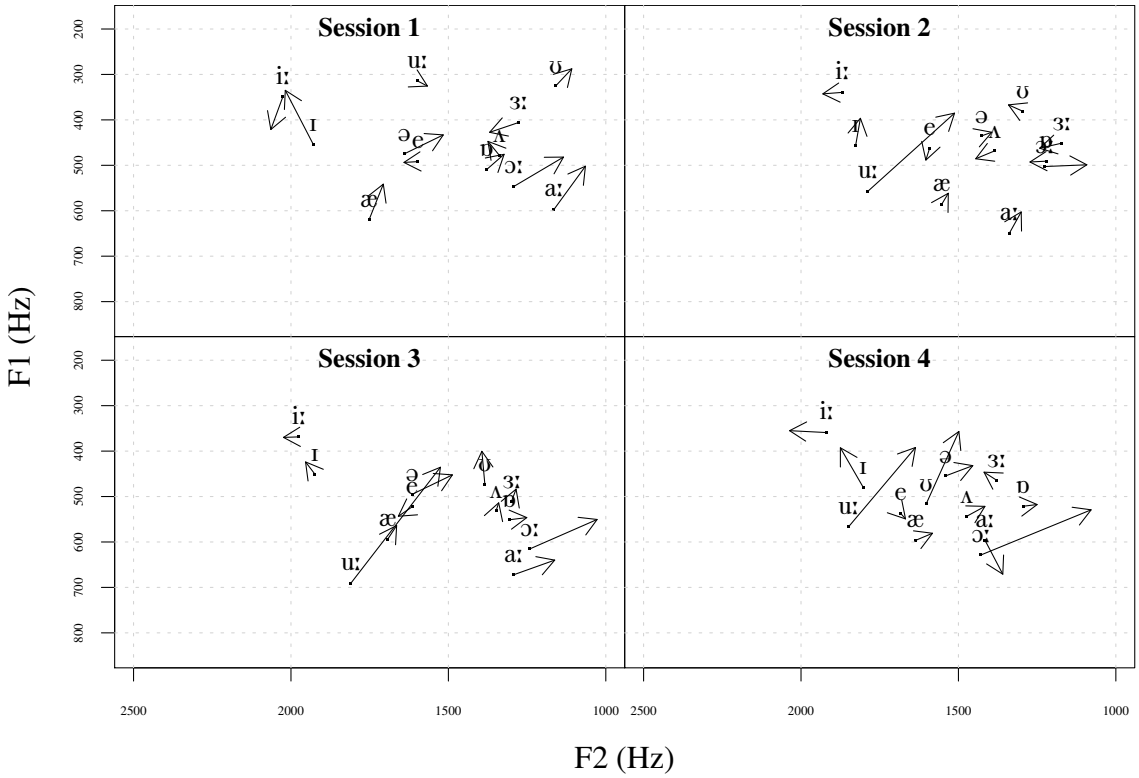


Fig. 3.3: Speaker DID0014's Vowel Inherent Spectral Change for monophthongs (Aligner: SPPAS).

to the F_1 and F_2 means at 20% and 80% of the vowel's duration. These values were those obtained from the SPPAS-aligned TextGrids. Although only an example, it is worth noting that the /ʊ/-/uː/-/u/ vectors are rather chaotic in either their locations (*e.g.* /u/ in session 1) or their lengths (*e.g.* /u/ and /uː/ in sessions 3 and 4). Another back rounded vowel, /ɔː/, also features comparatively longer F_1/F_2 vectors, especially in sessions 3 & 4. Note that the primary purpose of figure 3.3 consists in showing how relative to vocalic quality the stability of formant values is over the vowel's duration: recall that these values are calculated regardless of the consonantal environment; that the number of occurrences varies greatly from one phoneme to the other; and that so do the type/token ratios (TTRs) of the words featuring these phonemes. These caveats paradoxically strengthen the validity of the representation in figure 3.3: if certain means are found to be *consistently* higher or lower, then it could arguably mean that consonantal environments, numbers of occurrences and lexical variety only exert limited influence on VISCs, so that these VISCs may well provide

a metric of vocalic invariance in learners' pronunciations. The question therefore arises whether when adopting this method of calculation (*i.e.* by conflating vowels independently of the words they appear in), patterns emerge across speakers and sessions.

The answer to this question is investigated in section 3.2.1, where the means of onset-to-offset distances (henceforth, OOD) for each vowel, each speaker and each session are compared to their respective numbers of occurrences. The OOD is defined here as the Euclidean distance separating the 20th centile (the onset) from the 80th centile (the offset) in the F_1/F_2 space. For equivalent levels of phonemic acquisitions, the average OODs are expected to show values in the same ranges across speakers. Anomalies such as the lengths of the /uː/ and /u/ vectors in figure 3.3 will thus be detected. section 3.2.2 will take a look at the standard deviations of the OODs, more specifically those of the /ɪ/-/iː/ and /ʊ/-/uː/ phonemes. Dispersion provides an accurate way to assess acquisition, and its per-speaker consistency, or absence thereof, will be studied with respect to the number of tokens and types – in order to account for the influences of the various consonantal environments.

3.2.1 Onset-to-offset distances and vowel tokens

Does measuring the OODs reveal any useful information regarding phonemic acquisition, especially in a skewed corpus featuring a wide array of consonantal environments? Some theoretic stability of the OODs regardless of these environments has got to be posited: if none existed, then the link between phoneme-based understandability (arguably the very foundation of oral communication) and acoustic information would be severed. The problem is of course accrued when studying learners' production, but investigating their OODs, and how these might vary from one phoneme to another, is likely to shed light on the state of their level of acquisition. In order to assess that state, the OODs were measured speaker by speaker and phoneme by phoneme, and plotted against the number of types of syllables the phonemes appear in. The formant values were normalized beforehand, using the Bark

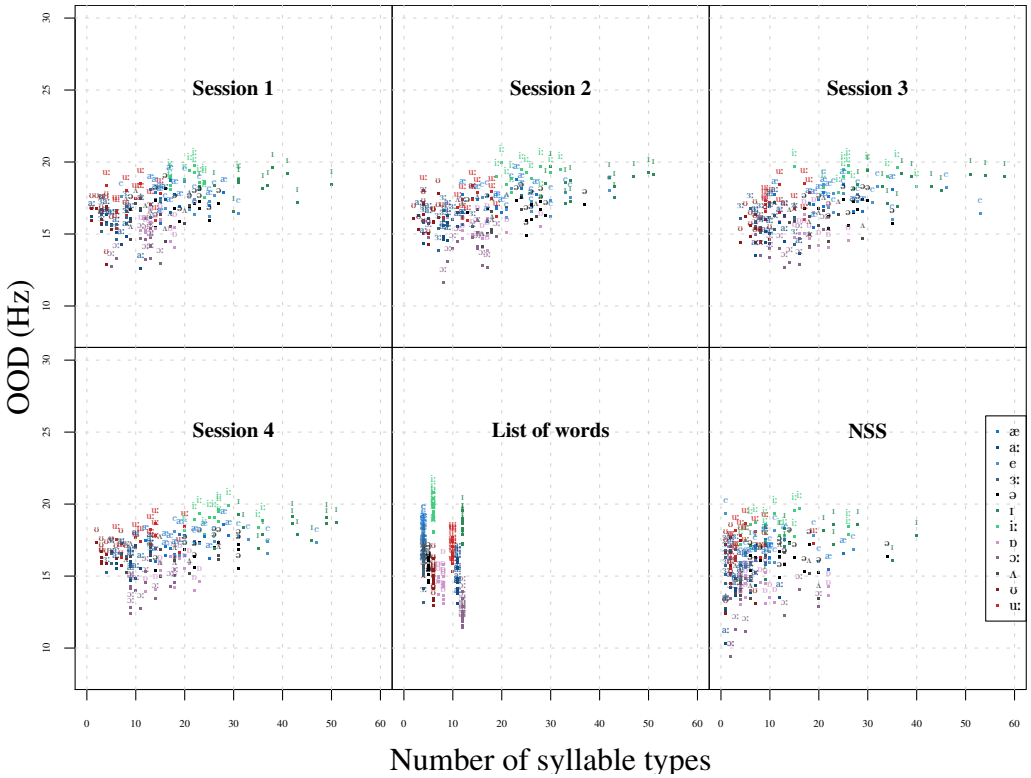


Fig. 3.4: Per-phoneme, per-speaker and per-session onset-to-offset distances (OODs) against the number of syllable types.

Difference Metric (*c.f.* section 2.4.2). This method was chosen here because it factors in the F_3 values, and therefore reduces to two the number of dimensions needed to visualize the three F_1, F_2 and F_3 values. The number of types of syllables is a metric that makes it possible to quantify the potential influence of different consonantal environments on the OODs. The results were then compared to native values. They are presented in figure 3.4. For these calculations, the /u/ and /i/ were merged into the /uː/ and /iː/ values respectively. The data from the recorded list of words has been included too, even though the number of syllable types is by definition static and dependent on the given list of words to read. Perhaps more interestingly, the bottom right panel shows the values for native speakers. Arguably the most striking facts from observing the figure are the similarities between the distributions and values in the learners' data, and those in the natives', in spite of the reduced number of datapoints: for most monophthongs, the relationship between the OODs and the number of syllable types is the same, with /ɪ/, /ə/ and /e/ being the phonemes with the highest

number of syllable types, and the lowest OODs. Back vowels seem to present longer OODs and fewer syllable types. Two native OODs for /ɜː/ and /ɔː/ are higher than 1,000Hz, a value no learner OODs reach. In terms of acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts, figure 3.4 does not reveal any significant differences, and even seems to confirm that both contrasts have been acquired, as the values are close to native values. The same calculations were carried out with the numbers of tokens rather than the number of syllable types, and the overall picture is the same.

However, a closer look at the data sheds light on a much more heterogeneous landscape. To investigate further, the per-gender native OODs (*i.e.*, from the NSS) were calculated. This returned the following values for /ɪ/-/iː/ and /ʊ/-/uː/ respectively: 18.1 Bark, 18.9 Bark, 17.1 Bark and 17.9 Bark for native women; and 16.8 Bark, 17.9 Bark, 14.9 Bark and 16.5 Bark for native men. The same procedure, averaging the OODs for each monophthong was the applied to each learner in each session. The next step was to calculate the difference between the native OODs for the four phonemes and each learner's OOD for each session. The results are presented in figure 3.5. Each panel corresponds to one learner. The dotted line marks the zero-difference value, *i.e.* the reference native line: above it, the learner's average OOD for a given phoneme is longer than the corresponding OOD; shorter when it is below the dotted line. Values above the dotted lines may therefore indicate a degree of articulatory overshooting and conversely, a degree of articulatory undershooting in the case of values below the dotted lines. Native values are sex-dependent, so that the lines indicate variations from the average native values of the corresponding sex. The first nine panels feature the OOD lines of the nine female learners, with the last three panels featuring those of the three male learners.

All /ɪ/ and /iː/ OOD lines across all speakers are above native values. There is a tendency among female learners to present shorter OODs for the /ʊ/ and /uː/ phonemes. Overall, there is much greater consistency in the /ɪ/ and /iː/ OOD lines than in those for /ʊ/ and

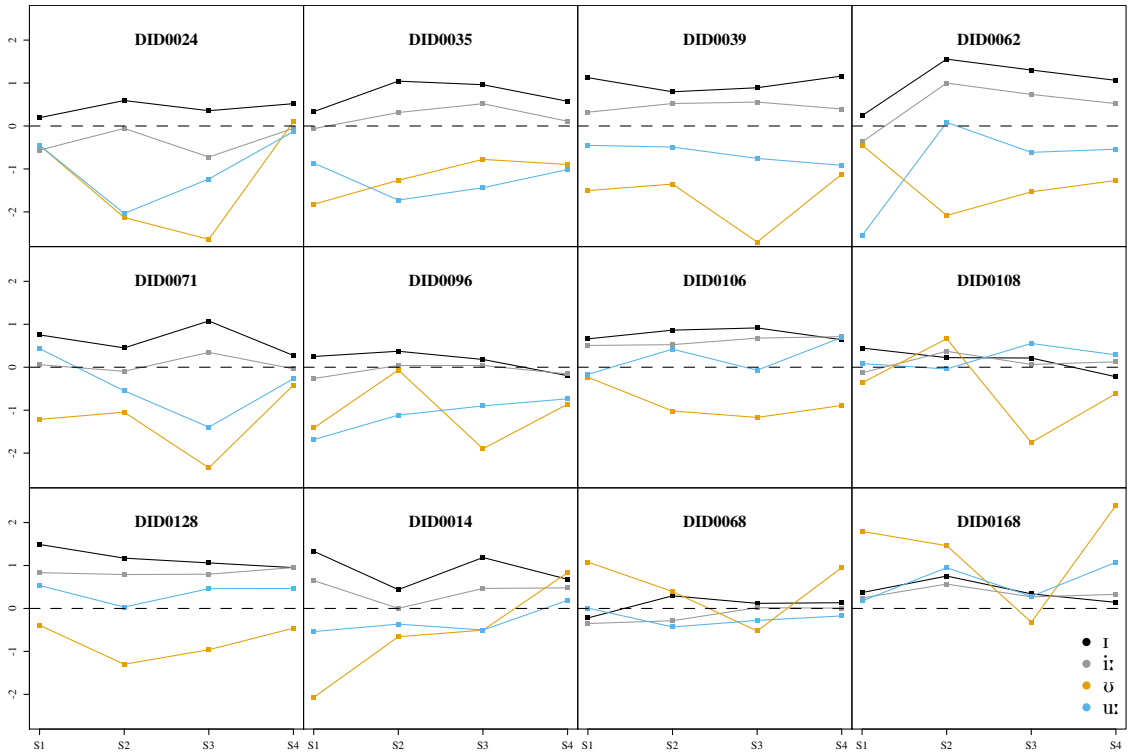


Fig. 3.5: Per-session differences in average OODs between the learners' and the natives' for /ɪ/, /iː/, /ʊ/ and /uː/.

/uː/. Leaving per-session differences aside, the standard deviations across learners of the same sex are the following, for the /ɪ/, /iː/, /ʊ/ and /uː/ phonemes respectively: 0.44 Bark, 0.42 Bark, 0.76 Bark and 0.80 Bark for women; 0.45 Bark, 0.32 Bark, 1.25 Bark and 0.54 Bark for men – the data for the latter being arguably less robust because of the lower number of participants. All of this seems to indicate a difference in the acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts. Looking at the absolute values of the distances to the native OOD values further confirms this statement: the average absolute OOD difference to the native values for /ɪ/, /iː/, /ʊ/ and /uː/ are the following: 0.70 Bark, 0.40 Bark, 1.14 Bark and 0.73 Bark for female speakers; 0.50 Bark, 0.31 Bark, 1.08 Bark and 0.41 Bark for men. It is contended here that these numbers constitute at least tentative evidence that the null hypothesis posited by theories in SLA, whereby there should be no difference in the acquisition of the two contrasts, can be rejected. However, further research can still be carried out to confirm or infirm these findings using VISCs, for instance by adjusting the centiles used to calculate OODs. Similar

results were obtained by using the 25th centile as the onset, and the 75th as the onset. The *R* scripts were also run using P2FA as the main aligner, and no major differences with the statements above were worth reporting.

However, these procedures and results do not really address the issue of the variety of the consonantal environments, which are bound to affect the OODs in different ways. How to take them into account while still using VISCs and trying to assess the state of acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts is the object of the next subsection.

3.2.2 Standard deviations of OODs

This section takes a look at the relationship between the per-phoneme, per-speaker and per-session standard deviations of the OODs, and the number of types and tokens of the consonantal environments embedding the monophthongs. The focus will be exclusively on /ɪ/, /iː/, /ʊ/ and /uː/. The expected relationship is the following: as the number of types and tokens of syllabic templates increase, so should the standard deviations of the OODs. This is because consonants preceding and succeeding a vowel exert influence on formant transitions, and that OODs (admittedly depending on the centile where the measurements are taken – here the 20th and 80th) are likely to be affected by those transitions: a higher number of syllabic types, *i.e.* of consonantal environments likely to exert influence on formants at the 20th and 80th centiles, should therefore induce greater standard deviations of the F_1/F_2 Euclidean distance between the onset and offset of the vowel. A consistent distance between the two centiles on each formant in the same consonantal environment can reasonably be assumed to be evidence of some acquisition. In other words, a higher number of syllabic types, *i.e.* of consonantal environments likely to exert influence on formants at the 20th and 80th centiles, should induce greater standard deviations of the F_1/F_2 Euclidean distance between the onset and offset of the vowel. For this specific study, and just as in the previous section, formant values were normalized beforehand using the BDM method. Figure 3.6 plots the standard

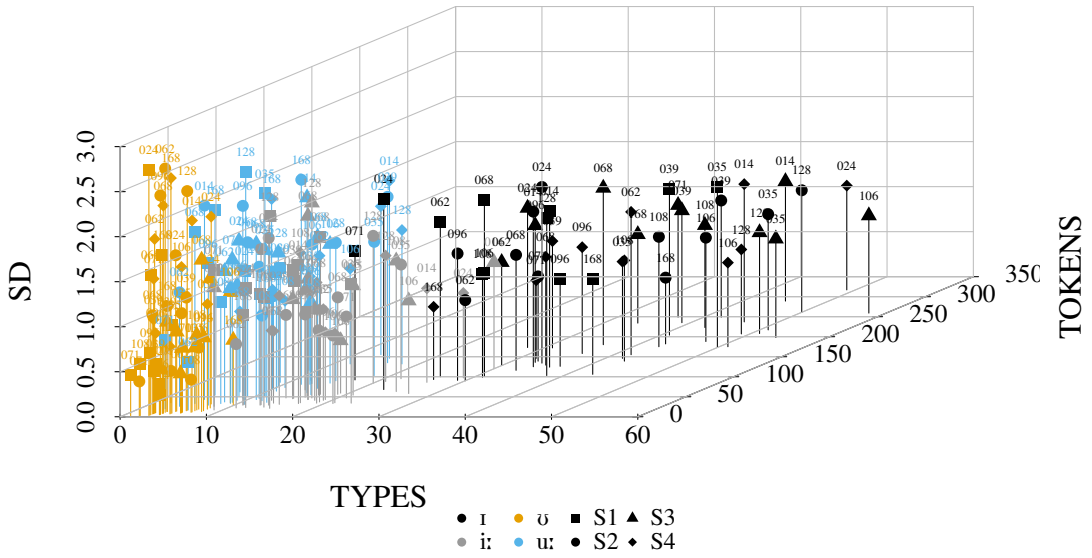


Fig. 3.6: Standard deviations of the per-speaker, per-session OODs for /ɪ/, /iː/, /ʊ/ and /uː/ against syllable tokens and types.

deviations of the OODs for each speaker in each session against the number of syllable types and tokens embedding /ɪ/, /iː/, /ʊ/ and /uː/. The shape of the dots indicate the session, while the numbers on top of them give the speakers' identification numbers. Looking at the figure, the distribution of each phoneme is clearly dependent on the number of syllable types and tokens. At first sight, the standard deviations of the OODs look more widespread in the case of /ʊ/ and /uː/ than for /ɪ/ and /iː/. This is however only partially borne out by taking a closer look: the global per-sex dispersions of the OOD standard deviations are similar for all phones (around 0.35 Bark for women, and 0.42 Bark for men) except /ʊ/, which in both instances display higher dispersion values (0.69 Bark for female learners and 0.74 Bark for male learners). Considering that syllables containing /ʊ/ have the lowest numbers of types and tokens, there is probably a frequency effect happening with this phoneme: the NSS features the same global distribution of phonemes, with /ʊ/ having the lowest numbers of syllable types and tokens too. A reasonable assumption could therefore be that a lower input

of /ʊ/ sounds embedded in fewer syllable types and tokens leads to a less stable state of acquisition, the scarcity of the input entailing higher dispersion. Surveying the per-sex means of the dispersions and comparing them between learners and natives[1] does not reveal any significant differences likely to imply specific levels of acquisition between the phonemes: the mean OOD standard deviations are mostly equivalent between the two categories of speakers, with only /ʊ/ in women, and /iː/ in men, which show substantial differences (respectively 1.12 Bark among learners and 0.64 Bark among natives in the former case, and 1.26 Bark and 0.85 Bark in the latter case): /ʊ/ features greater dispersion overall among female learners, and so does /iː/ among male learners… Still, the data at hand and the

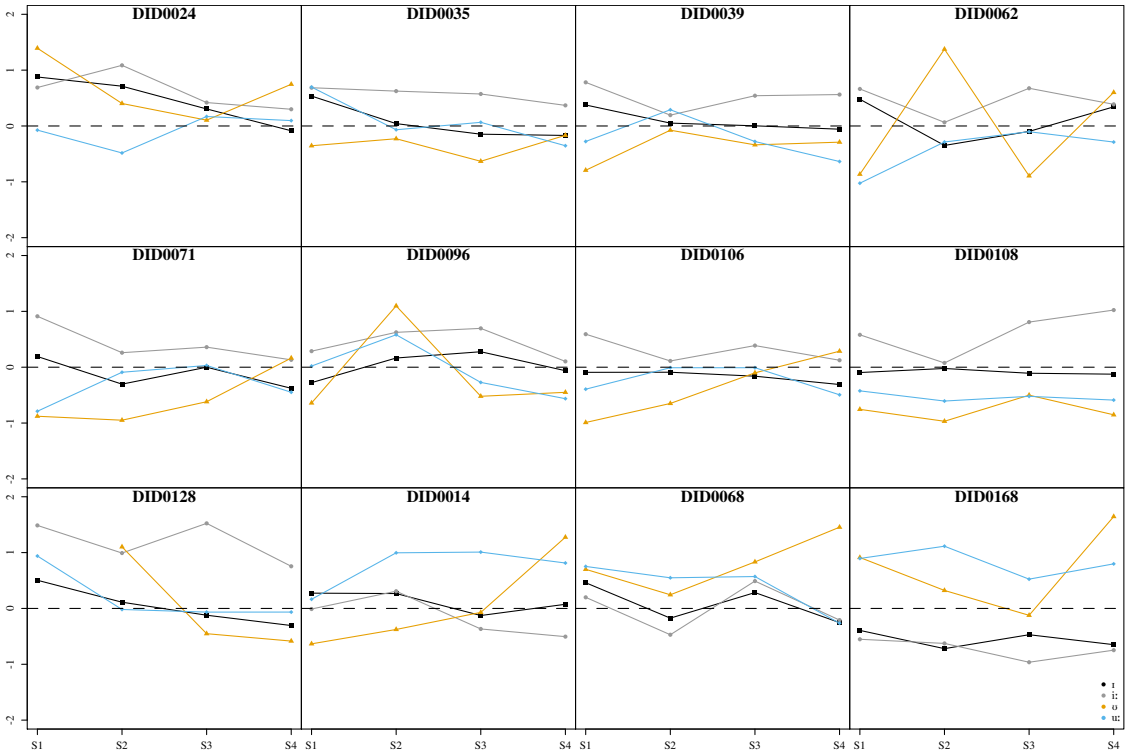


Fig. 3.7: Per-session differences in OOD standard deviations between the learners' and the natives' for /ɪ/, /iː/, /ʊ/ and /uː/.

focus on OOD standard deviations seem to reveal much less stability in the realizations, and therefore arguably in the acquisition, of /ʊ/. In a similar way to figure 3.5, figure 3.7 plots

[1] The same plot as figure 3.6 for native speakers can be found in section E.2 (figure E.2).

the evolution of the OOD standard deviations for each learner across the four sessions[2]. A cursory graphic inspection leads to the conclusion that /ʊ/ OOD standard deviations do seem higher than the counterparts of other monophthongs. The means of the absolute values of the differences between learners' and native speakers' OOD SDs across all speakers and sessions are the following, for /ɪ/, /iː/, /ʊ/ and /uː/ and repsectively: 0.23 Bark, 0.30 Bark, 0.58 Bark and 0.34 Bark for women; 0.35 Bark, 0.46 Bark, 0.62 Bark and 0.70 Bark for men. Once again, the orders of magnitude vary substantially between the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts. The latter contrast features greater standard deviations on average, regardless of sex. Do sessions have an effect? One way to answer this question could be to look at the evolution of the distance to native OOD standard deviations. The mean differences from native OOD standard deviations tend to decrease across female learners for all for monophthongs: /ɪ/ (from 0.38 Bark to 0.20 Bark, 0.14 Bark and 0.20 Bark for sessions 1, 2, 3 and 4 respectively); /iː/, although the values are higher and session 3 counters the tendency (0.74 Bark, 0.45 Bark, 0.67 Bark, 0.42 Bark); /ʊ/ (0.83 Bark, 0.76 Bark, 0.46 Bark, 0.46 Bark); and /uː/ (0.52 Bark, 0.27 Bark, 0.17 Bark, 0.39 Bark), in spite of a spike in session 4, possibly due to the much greater number of tokens. The data for male learners, which is less substantial, does not further validate the small effect of sessions – the mean distances remain stable across the sessions, except for /ʊ/: /ɪ/ is still the most stable monophthong (0.38 Bark, 0.39 Bark, 0.29 Bark, 0.33 Bark); /iː/ presents slightly increasing dispersions (0.25 Bark, 0.47 Bark, 0.61 Bark, 0.49 Bark); /uː/ has comparatively higher mean differences (0.60 Bark, 0.89 Bark, 0.70 Bark, 0.63 Bark), while /ʊ/ starts at 0.75 Bark, drops to 0.31 Bark in session 2 and 0.34 Bark in session 3, only to rise again up to 1.46 Bark in the last session[3].

The dispersion of the OODs which was looked at in this section reveals that /ʊ/ is by far the most volatile monophthong of the four vowels specifically under study. This is all the more unexpected as /ʊ/ is the monophthong embedded in the fewest types of

[2] Speaker DID0128 only pronounced /ʊ/ one time in session 1, hence the absence of a datapoint in panel n°9.

[3] These values are plotted against the number of tokens in figure E.3 of appendix F.

syllables, and occurring the smallest number of times. This section also shed light on the comparable instabilities of /iː/ and /uː/, with /ɪ/ having the least dispersion. It is clear that the realizations of the four phonemes present varying degress of dispersion, which begs the question of a unified analysis. When it comes to acquisition, however, greater dispersion can arguably signify a less stable state of acquisition. From that perspective, the greater instability of OODs for /ʊ/ and /uː/ against that of /ɪ/ and /iː/, when compared to native dispersion, is another indication that the two contrasts under study follow different learning curves. To investigate this issue further, methods of pattern recognition could also be explored: this exploration is the object of the next section.

3.3 *k*-Nearest Neighbours

Possibly one interesting way to assess phonemic acquisition would be to consider it as a classification problem: considering the vowels in the BDM-normalized F_1 / F_2 space, to what extent could machine-learning algorithms be trained to label them, and to what extent would the predictions differ first from one phoneme to another, then from one session to the other? One argument that makes this approach particularly appealing is that having a corpus of native values (albeit a small one) is ideal if one is to consider methods of supervised learning: those native values constitute the ideal labelled training data to infer the learners' values from. This section explores the success rate of a simple classification method, the *k*-Nearest Neighbours (henceforth, KNN).

3.3.1 Method

There are several ways to implement classification on the main data: some key questions to answer are the scope of the test sets (*e.g.* across all speakers or not, or across all sessions), the number of classes (*i.e.* how many different phonemes should the classification operate

on?), and of course the value of k. Here the dimensions taken into account are the BDM-normalized F_1 and F_2 taken at the mid-temporal point because F_3 is also factored in in the computations. Mid-temporal measurements also limit potential influences of consonantal environments. To keep those in check, and as will be the case in section 3.4, the original goal was to take into account only the phones occurring in syllable templates also present in the NSS. In order to assess longitudinal acquisition, and in order to preserve potential idiosyncrasies but also to detect possible cross-speaker patterns, the test sets were split between both speakers and sessions; no subsets of monophthongs were selected. This means that a typical test set consists of all the monophthongs produced by a given speaker during a given session, and embedded in a consonantal environment also present in the NSS. The non-linear nature of KNN makes it possible to implement multiclass classification, provided (as is the case here) that the classes do not overlap (*i.e.* not multilabel classification). The training set was divided into two separate sets, one for each sex: the per-speaker, per-session test sets were therefore classified along the labels of the training sets of the corresponding sex.

Another key issue to address is of course that of which k value to select. Cross-validation, a common method in medical sciences, is in this case not possible: the training sets are distinct from the test sets. Randomly sub-sampling the learners' datasets would besides make very little theoretical sense – or at least a better solution exists, namely the resort to the NSS as the training set. Cross-validation not being a viable option here, another procedure to determine the optimal value of k had to be implemented. One way to go about doing so is to operate KNN classifications with all values of k up to the rounded value of the square root of the number of datapoints in both the training and test sets. $\sqrt{n}$, with n the number of instances in both sets, is an empirical value commonly used (probably after (Duda et al., 2000, Chapter IV)). For each KNN classification with $k \in [1, \sqrt{n}]$, the proportion p of correctly

labelled phonemes was stored, in each session for each speaker. The process[4] is represented in figure 3.8 for speaker DID0068 in Session 2, where the blue line plots the proportion p of accurately classified monophthongs against each value of k.

One issue with scanning all the values of k from 1 to $\sqrt{n}$ is that optimal values, *i.e.* values of k for which the proportion of accurately labelled phonemes in the test set is the highest may vary from one pass to another. This instability of results is likely due to even-numbered values of k, where the majority vote can return a tie – and therefore arbitrary decisions must be taken –, equal distances between datapoints, or all neighbours are from different classes. In order to counter this instability of results, 1,000 parses were carried out, *i.e.* the values of k were scanned from 1 to $\sqrt{n}$ 1,000 times for each speaker in each session. For each pass, the *most frequent* optimal k-value was selected: in other words, the k-value from 1 to $\sqrt{n}$ returning the highest proportion of accurately labelled phonemes was computed 100 times, and stored, and the most frequent of these values was then retained for the final comparisons. Before proceeding on to the results, another issue has to be addressed, that of scaling. The space where the datapoints are located is the BDM-normalized F_1/F_2 space – two dimensions with different measurement scales: although F_1 and F_2 feature roughly equivalent standard deviations (1.42 and 1.45 Bark respectively), there are non negligible discrepancies between their minimal values (2.4 *vs.* 0), maximal values (15.34 and 10.65), and more importantly, their means (10.22 against 2.95). These different scales mean that when computing the Euclidean distance between neighbours for the KNN algorithm, F_1 will have a greater influence on the calculated distance. Scaling is therefore required, and the method chosen for standardization is the z-score: $z = \frac{x-\mu}{\sigma}$, with z the normalized value, x the raw value, μ the mean of all values, and σ their standard deviations. All BDM-normalized F_1 and F_2 values were thus standardized.

[4] The process to determine optimal values of k and the design of figure 3.8 were taken directly from daviddalpiaz.github.io.

An example of the entire process can be found in figure 3.8, which shows the case of speaker DID0068, a male speaker, in session 2. In this instance, there are 618 phones embedded in a syllable present in the NSS. The number of phones pronounced by native male speakers is 1,289, so the total number of datapoints in both the training set and the test set is 1,907. The top panel of figure 3.8 shows how the KNN algorithm was run on those two sets with values of k varying from 1 to the integer value of $\sqrt{1,907}$, *i.e.* 43. The y-axis indicates the global proportion of correctly labelled phonemes for each k-value. The highest proportion in this example is 0.33, corresponding to a number of neighbours $k = 42$. This value of k can be called "optimal" because it is the value returning the highest classification accuracy. "Classification accuracy" is here to be understood as the proportion of learners' monophthongs correctly recognized by the KNN algorithm as the monophthong that should have been produced given the word in which it appears; whether this identification is "correct" or not is assessed by using the datapoints of the natives' monophthongs as references. However, as mentioned above, the optimal value of k can change if the process is repeated in exactly the same way, so a procedure[5] had to be found in order to determine a more robust optimal k-value for each speaker in each session. The solution adopted here was to loop the process of scanning k-values from 1 to $\sqrt{n}$ 1,000 times, and to store the optimal k-values after each pass. The different optimal values for speaker DID0068 in session 2 over 1,000 passes are plotted in the bottom panel of figure 3.8. How then to select a final, optimal k-value? Arguably the logical thing to do is to take the value towards which k converges, if any. In this case, the converging value is the most frequent over 1,000 passes. Figure 3.8 shows how relevant this method is: the optimal k-value obtained in the top panel (42) is different from the most frequent value over 1,000 passes (23). Figure 3.9 shows how many times k-values between 1 and $\sqrt{1,907}$ have been considered optimal after the 1,000 passes. Interestingly, the value 42 was only found optimal 60 times. 4 values competed for the top position as the most frequent value, 21, 22, 23 and 29. 23 came first with 182 occurrences,

[5] The reader will remember that cross-validation is not a viable option.

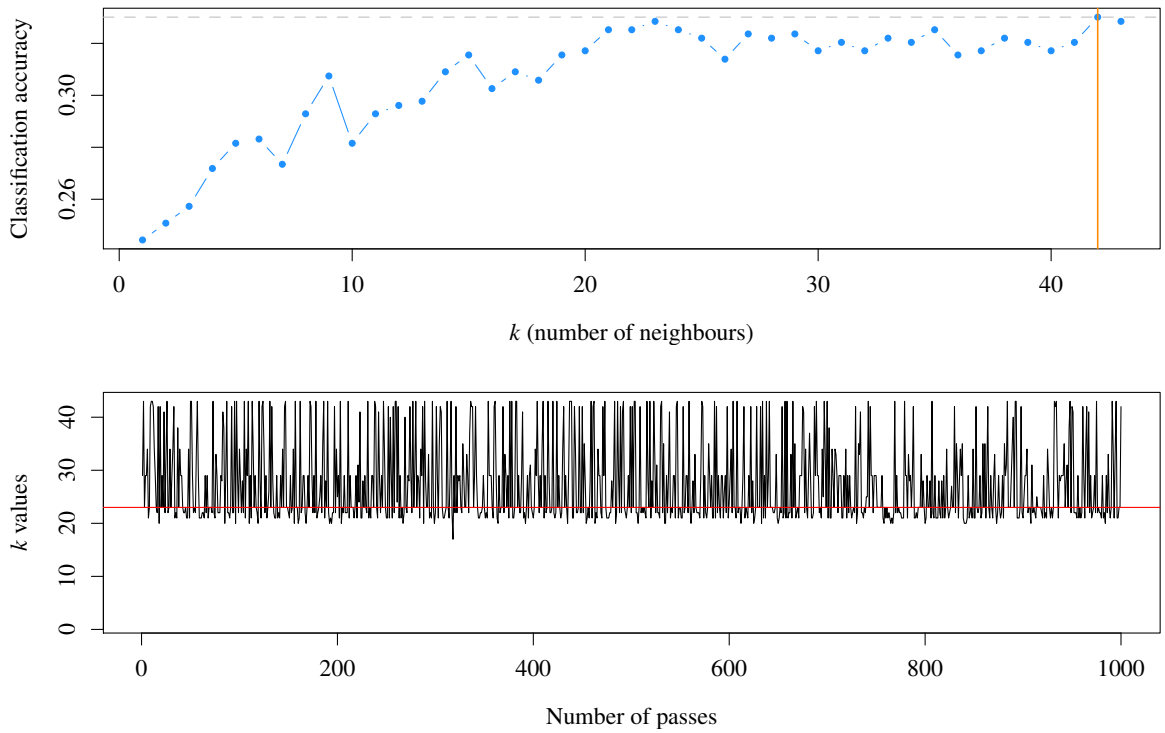


Fig. 3.8: *Top panel:* Proportions of correctly labelled phonemes against different values of k (with $k \leqslant \sqrt{n}$) for speaker DID0068 in Session 2 . The grey dotted line indicates the highest proportion of correctly classified phonemes (regardless of their categories); the vertical orange line indicates the optimal k value. *Bottom panel:* variations of optimal k values across 1,000 passes; the red line indicates the most frequent value (here: 23).

against 174 for 21. This means that for speaker DID0068 in session 2, the optimal k-value is 23.

What sort of accuracy can be expected? There are many reasons why a low classification accuracy, *i.e.* a low number of learners' monophthongs being identified as what the learners meant to pronounce based on the natives' references, can occur. The first reason is that the training sets and the test sets are separate. Another reason is the dispersion of learners' values, and the discrepancy between their realizations and the native targets. An overall low accuracy rate will dismiss the method as not efficient to detect potentially different rates of phonemic acquisition. On the other hand, clear differences in proportions of accurate labels from one monophthong to another should constitute evidence that discrepancies exist in the state of acquisition for each phoneme. A possible method to assess classification

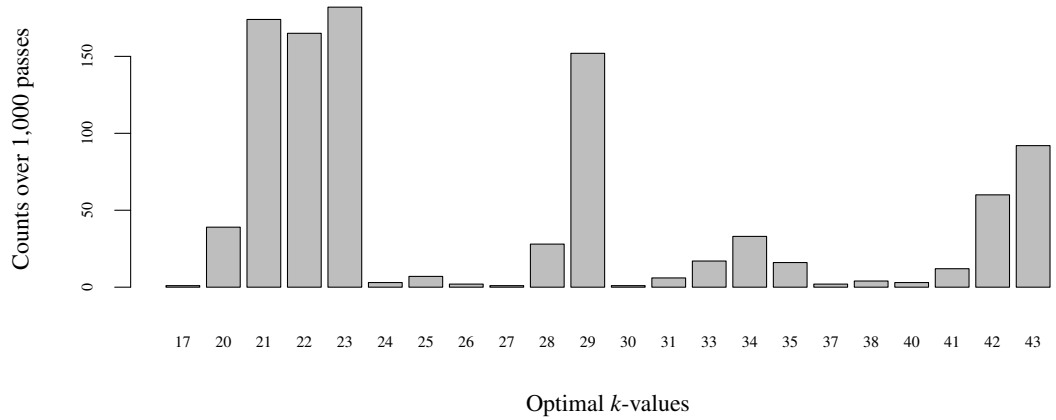


Fig. 3.9: Barplot of the counts of optimal k-values over 1,000 KNN passes for speaker DID0068 in session 2.

accuracy *a priori* is to run the algorithm on the native values, using, this time, cross-validation. Originally, as briefly mentioned above, the NSS was chosen to serve as the training set. The way cross-validation was performed on the datapoints is the following: the NSS was split into two subsets, one for female speakers, another for male speakers. These two subsets were then themselves separated according to the variety of spoken English. Only British and American English were retained. The BDM-normalized F_1 and F_2 values were then scaled using the z-score method described above. These subsets were then split into ten folds, using the *R* `caret` package. Table 3.1 provides an example of a random sampling of the female British speakers' data into ten folds. The distribution of phonemic targets within each fold matches that of the entire dataset. Within the folds, k was allowed to vary from 1 to $\sqrt{n}$, n being the number of phones in the fold. As the folds were parsed, the global proportion of accurately predicted phonemes was stored, along with the individual proportions for each category. After the algorithm was run on all ten folds, and with $k \in [1, \sqrt{n}]$, the optimal k value, *i.e.* the number of nearest neighbours for which the classification accuracy was the highest, was selected. The highest classification accuracy corresponds to the mean of the correctly labelled phonemes of each fold. Then this entire process, starting from splitting

	Fold01	Fold02	Fold03	Fold04	Fold05	Fold06	Fold07	Fold08	Fold09	Fold10
æ	18	19	19	18	19	18	19	19	18	19
ɑː	4	3	3	3	3	4	3	4	4	4
e	9	8	9	8	9	9	8	9	9	9
ɜː	3	3	3	2	2	2	2	3	3	2
ə	33	34	33	34	33	34	33	34	34	33
ɪ	36	36	35	36	36	35	35	35	36	36
iː	20	20	20	20	20	20	20	20	20	20
ɒ	14	14	14	14	14	14	13	14	13	14
ɔː	9	9	9	9	9	9	9	9	9	9
ʌ	7	7	7	7	7	7	7	7	7	7
ʊ	1	1	1	2	1	1	1	1	1	1
uː	10	10	9	9	10	9	9	10	10	9

Table 3.1: Example of 10 folds for the female British speakers: distribution of phonemic targets. This extremely uneven distribution eventually led to forfeiting the NSS as the training set.

one of the four sex- and accent-dependent datasets into ten folds, was repeated 100 times. The mean global proportion of accurately predicted phonemes in the case of native British women is 0.43; for British men, 0.37; and in the case of American native speakers, 0.30 and 0.34 for women and men respectively. These very low scores can be accounted for first by the low number of occurrences (1628, 425, 221 and 307 for British women and men, and American women and men respectively); then by the very unequal numbers of occurrences from one phoneme to another. However, because these overall proportions of classification accuracy were so low, and because the numbers of phonemes were so different across categories (*c.f.* figure 3.2, with no /ʊ/ in the dataset of male American speakers), it was decided that another set of native monophthongs should be resorted to. The reader interested in the results obtained from the NSS can refer to figure E.4 in section E.4, which is the NSS-based equivalent of figure 3.10, detailed below, and to an example of a confusion matrix based on the NSS (table E.1). What training set to use, then? Because of its easy availability, and because it is based on various speakers whose individual data have not been compiled, the Peterson & Barney (1952) set was used[6]. Cross-validation on the P & B corpus

[6] Clearly the issue of what training to use is not a trivial one, and the question is far-reaching. Either option at our disposal introduced a bias: resorting to the NSS corpus presented the advantage of remaining within the realm of spontaneous conversation. After all, the gaps in the number of occurrences between monophthongs all the more reflect, or so we argue, the reality of natural English as this gap is mirrored in the main learner corpus. A cursory look at figure E.4, however, quickly reveals how of little use the obtained results are: the proportions of accurately predicted /ʊ/ and /uː/ are null in 2, out of 48, instances (4 sessions × 12 speakers). The issue is that it is unclear whether these low ratings are the consequences of variable realizations on the learners' part, or of the limited numbers of native /ʊ/ and /uː/ in the first place – which make it too exceptionally possible for the

yields much more robust results: over 1,000 passes, the overall proportion of accurately labelled monophthongs is 0.787 in both male and female speakers. One major difference with the NSS is that the P & B corpus does not contain data for /ɒ/ (as it is an American corpus) and /ə/.

3.3.2 Results

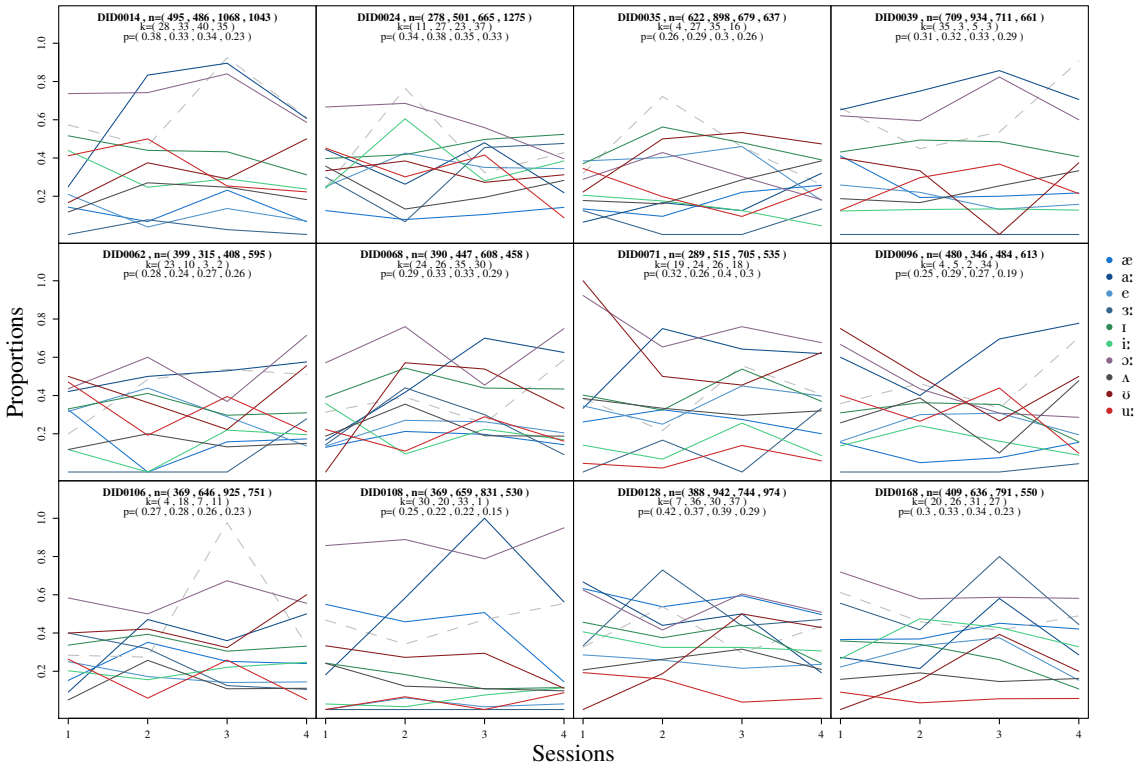


Fig. 3.10: Proportion of correctly labelled phonemes in the BDM-normalized F_1 / F_2 space using the KNN classification method. In each panel, the total number of tokens n for each session is indicated, along with the optimal k-values and the global proportion of accurately labelled phonemes. The grey dotted line indicates the frequency of occurrence of the optimal k value over the 1,000 passes.

The results of the study are summarized in figure 3.10. The proportion of each phoneme correctly labelled by the algorithm after 1,000 passes is plotted speaker by speaker and

algorithm to ascribe those predictions to the learners' productions. On the other hand, the P & B corpus imposes a variety of accent (American) and uses lists of recorded words with controlled consonantal environments and balanced numbers of items in each phonemic category – features which do not reflect the spontaneous conversation under study here. This is admittedly another form of bias, but at least the results (*c.f.* figure 3.10) provide information on the learners' productions only, and not on the possible flaws of the training set.

session by session. The grey dotted line plots the frequency with which the optimal k value between 1 and $\sqrt{n}$ was selected over 1,000 passes. It provides a measure of robustness of the results: the closer to one this value is, the less variation there is in the classification accuracies of the phonemes. The lowest value is 0.199 (for speaker DID0062 in session 1), *i.e.* the optimal k value of 23 was selected 199 times over the 1,000 passes. The maximal value is 0.977 (speaker DID0106 in session 3). On average, the optimal k values were selected 471 times; the dispersion is at 0.18. Such numbers can arguably show that the results are robust: even the minimal number of occurrences is evidence of stability, since each value competes with all other values from 1 to $\sqrt{n}$. The lowest value for n is 278 (speaker DID0024 in session 1), *i.e.* the smallest interval of possible k values is $[1,16]$. How correlated is the frequency of occurrence of the optimal k value and the number of phonemes? One assumption could be that the lower the number of phonemes to classify, the higher the frequency of the optimal k value: with fewer values to test against, an optimal k value is more likely to emerge. It turns out this is not the case, however: the Pearson coefficient is $r = 0.35$ – too weak a value to assume a relationship between the phoneme counts and the frequency of occurrence of the optimal k value over 1,000 passes. The panels figure 3.10 also display the total count of monophthongs across all sessions (n), the most frequent k value returning the highest global classification accuracy (k), along with the mean proportion of accurate labeling regardless of phonemic categories (p). No clear pattern emerges from figure 3.10, except perhaps the visibly higher proportions of /ɑː/ and /ɔː/. No evolution across sessions is observable. How do monophthongs fare when compared to one another? A synthesis of these findings can be found in table 3.2, which shows the per-phoneme means and standard deviations of the classification accuracies across speakers and sessions. The proportion of accurately labelled /ɔː/ is much higher (at 60%) than its closest competitor, /ɑː/. That these two monophthongs are low back vowels may explain such relatively high rates of accurate predictions. Their relative isolation in the vowel space, and their closeness to French sounds, may also contribute

	æ	ɑː	e	ɜː	ɪ	iː	ɔː	ʌ	ʊ	uː	Mean
μ	0.25	0.49	0.24	0.18	0.37	0.21	0.60	0.22	0.37	0.20	0.31
σ	0.16	0.23	0.12	0.22	0.11	0.13	0.18	0.10	0.20	0.14	0.16

Table 3.2: Per-monophthong means (μ) and SDs (σ)of the KNN classification accuracies across speakers and sessions

to these high proportions. This assumption cannot be generalized to back vowels, since /uː/ features a rather low chance of correct prediction, with 20%. Quite interestingly, the two contrasts under study, /ɪ/ – /iː/ and /ʊ/ – /uː/, present the same proportions: 37% for the short vowel, and 20% for its long counterpart. At first sight, then, the KNN algorithm does not establish any clear differences in the acquisition of the two contrasts. However, the standard deviations of /ɪ/ (0.11), /iː/ (0.13), /ʊ/ (0.2) and /uː/ (0.14) confirm what has already been detected in the other sections: /ɪ/ and /iː/ are consistently more stable than /ʊ/ and /uː/ in spite of a greater number of occurrences (7959, 5294, 578 and 3358 respectively). The slightly higher standard deviations in classification accuracies, along with the gaps in the numbers of occurrences, reveal disparities across speakers and sessions. But these disparities should not be exaggerated, as they are still quite subtle. Following these observations, the question naturally arises whether a correlation exists between the number of occurrences and the proportion of accurate predictions for each monophthong. Figure 3.11 gives graphical evidence that it is not the case: the proportions of accurately labelled phonemes for each speaker in each session were plotted against their respective numbers of occurrences, and no relationship between the two parameters is visible. The Pearson correlation coefficient is $r = -0.09$. No session effects are observable either. Within-phoneme values are randomly distributed along the *y*-axis, which confirms the absence of correlation even at a finer-grained level. These findings arguably demonstrate the validity of the procedure and its results: had there been a correlation between classification accuracy and the number of occurrences, cross-comparisons between monophthongs would have been unreliable. In other words, the analysis of the classification is made possible because the differences in counts have been

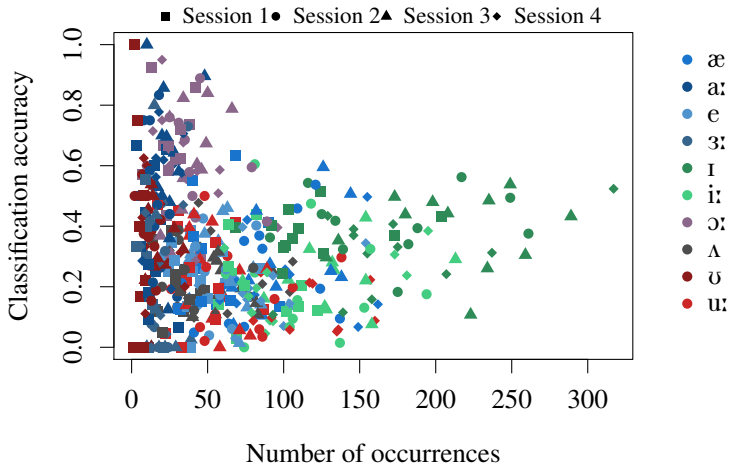


Fig. 3.11: Per-session, per-speaker proportions of correctly labelled phonemes (using optimal *k*) against their numbers of occurrences.

effectively neutralized. Coming back to the results from figure 3.10, and their summary in table 3.2, it looks as if the differences in proportions of monophthongs accurately predicted by the algorithm are only marginal. Is this really the case, though?

Another approach worth adopting to assess whether resorting to the KNN algorithm reveals any differences in the acquisition of /ɪ/, /iː/, /ʊ/ and /uː/ is to see what values the monophthongs were predicted to have – in other words, to look at the confusion matrices. Figure 3.12 plots the confusion matrices of all 12 speakers. The numbers inside each square give the proportion of actual phonemes on the *x*-axis predicted to belong to the corresponding category on the *y*-axis. The way these confusion matrices were computed is the following: for each session and each speaker, the algorithm was run using the optimal *k* value, following the procedure described in section 3.3.1. The resulting actual and predicted values were then listed together for each speaker, and the confusion matrices were then computed over the four sessions. No clear pattern emerges either upon observing the confusion matrices. However, looking at the second most predicted phoneme which is not the actual phoneme for /ɪ/, /iː/, /ʊ/ and /uː/ provides some insight: table 3.3 lists the best alternative predicted phonemes for actual /ɪ/, /iː/, /ʊ/ and /uː/ for each speaker, and compares their percentages of predictions

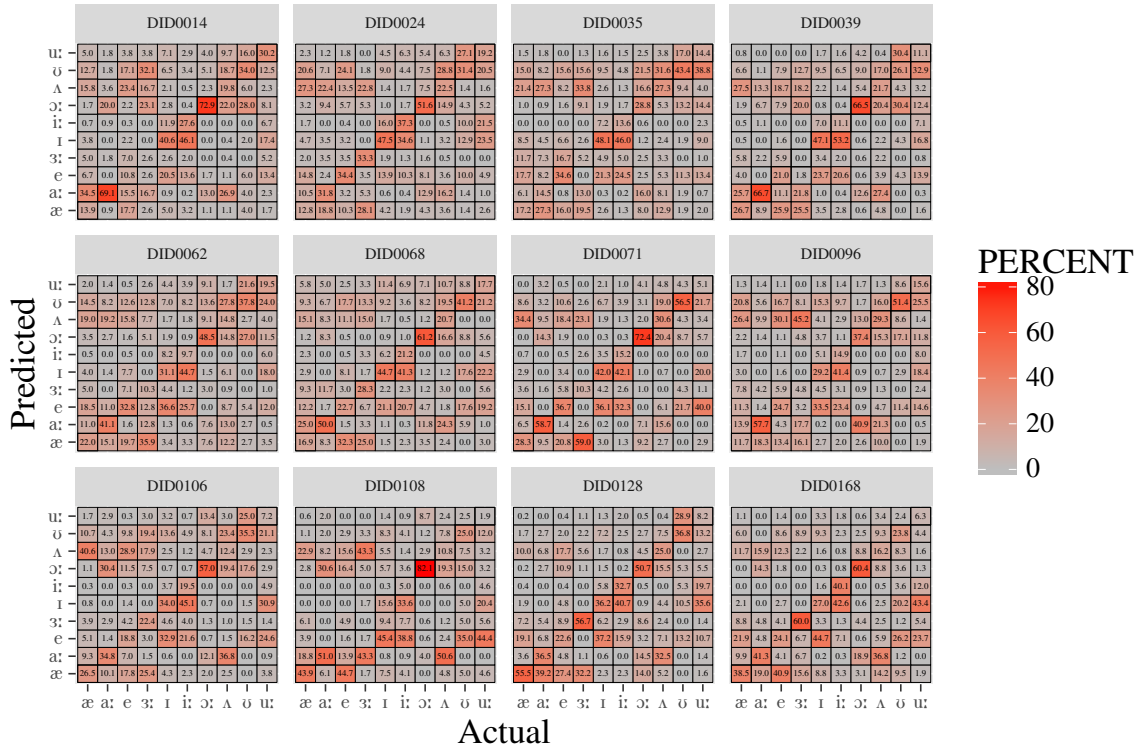


Fig. 3.12: Learners' confusion matrices for KNN classification: the optimal *k* values were retrieved for each session and per-session confusion matrices were then merged for each speaker. The numbers inside the tiles indicate proportions (*c.f.* main text).

with those of the actual four phonemes. The number of occurrences of each phonemic target for each speaker across the four sessions are also given as a reminder, in order to investigate whether the total count exerted influence on the actual and alternate percentages. From here onwards, "actual percentage" refers to the percentage of correctly identified phonemes, while "alternate percentage" refers to the highest proportion of inaccurately predicted phonemes. The phonemes different from the actual ones with the highest alternate percentages are refered to as the "best alternates".

Arguably, the most striking differences between the four phonemes lie in the consistency of the lists of best alternates. In cases of inaccurate classifications, /ɪ/ is overwhelmingly predicted as /e/, with only one exception in speaker 24, for whom the best alternate is /iː/. Likewise, the best alternate for /iː/ is /ɪ/, except for speaker 108, whose best alternate for /iː/ is /e/. Even when including the two exceptions, these best alternates are phonologically

				Predictions (%)						Predictions (%)	
Speaker	**Actual**	**Count**	**Alternate**	**Actual**	**Alternate**	**Speaker**	**Actual**	**Count**	**Alternate**	**Actual**	**Alternate**
014	ɪ	775	e	40.66	20.52	014	iː	557	ɪ	27.65	46.14
024	ɪ	692	iː	47.54	16.04	024	iː	474	ɪ	37.34	34.60
035	ɪ	734	e	48.09	21.25	035	iː	522	ɪ	13.60	45.98
039	ɪ	860	e	47.09	23.72	039	iː	504	ɪ	11.11	53.17
062	ɪ	473	e	31.08	36.58	062	iː	331	ɪ	9.67	44.71
068	ɪ	535	e	44.67	21.12	068	iː	392	ɪ	21.17	41.33
071	ɪ	626	e	42.01	36.10	071	iː	387	ɪ	15.24	42.12
096	ɪ	511	e	29.16	33.46	096	iː	350	ɪ	14.86	41.43
106	ɪ	723	e	34.02	32.92	106	iː	426	ɪ	19.48	45.07
108	ɪ	636	e	15.57	45.44	108	iː	443	e	4.97	38.83
128	ɪ	779	e	36.20	37.23	128	iː	511	ɪ	32.68	40.70
168	ɪ	615	e	26.99	44.72	168	iː	397	ɪ	40.05	42.57

				Predictions (%)						Predictions (%)	
Speaker	**Actual**	**Count**	**Alternate**	**Actual**	**Alternate**	**Speaker**	**Actual**	**Count**	**Alternate**	**Actual**	**Alternate**
014	ʊ	50	ɔː	34.00	28.00	014	uː	344	ɪ	30.23	17.44
024	ʊ	70	uː	31.42	27.14	024	uː	307	ɪ	19.22	23.45
035	ʊ	53	uː	43.40	16.98	035	uː	299	ʊ	14.38	38.80
039	ʊ	46	uː ~ ɔː	26.09	30.43	039	uː	380	ʊ	11.05	32.89
062	ʊ	37	ɔː	27.03	37.83	062	uː	200	ʊ	19.50	24.00
068	ʊ	34	e ~ ɪ	41.18	17.65	068	uː	198	ɪ	17.68	22.22
071	ʊ	23	e	56.52	21.74	071	uː	175	e	5.14	40.00
096	ʊ	35	ɔː	51.43	17.14	096	uː	212	ʊ	15.57	25.47
106	ʊ	68	uː	35.29	25.00	106	uː	346	ɪ	7.23	30.92
108	ʊ	40	e	25.00	35.00	108	uː	216	e	1.85	44.44
128	ʊ	38	uː	36.94	28.95	128	uː	365	ɪ	8.22	35.62
168	ʊ	84	e	23.81	26.19	168	uː	316	ɪ	6.33	43.35

Table 3.3: Comparison of the per-speaker percentages of the most frequently alternative predicted phonemes (Alternate) for actual /ɪ/, /iː/, /ʊ/ and /uː/ with the proportions of accurate identifications (Actual).

close to their actual counterparts: /e/ is a front vowel just like /ɪ/, but with a sligtly wider opening of the mouth. This process of a possible excessive aperture is reproduced with actual /iː/, predicted as /ɪ/ 11 out of 12 cases. The overall picture is much less consistent, from a phonological point of view at least, in the case of /ʊ/ and /uː/. /ʊ/ has /uː/ as best alternate in 4 out of 12 cases; /ɔː/ in three cases, with a draw between those two phonemes in one instance. These eight cases present some phonological consistency, with /uː/ and /ɔː/ both being rounded back vowels. With /ʊ/ being a near-close, near-back rounded vowel[7], /uː/ and /ɔː/ can be considered as predictions reasonably close to the actual phoneme. The four remaining best alternates are /e/, with /ɪ/ being an equally best alternate in one instance. The

[7] The extent to which /ʊ/ is rounded will not be discussed here. One tempting explanation for these two best alternates is that learners may over-round their lips when pronouncing /ʊ/. The data here do not seem to support this assumption. BDM-normalized F_2 values for female native speakers are much lower on average than learners' values: 2.88 Bark in the NSS and 3.67 Bark in the P & B corpus, against 3.93 Bark and 4.95 Bark in the conversations and reading lists respectively. For male speakers, the means are 4.30 Bark and 3.68 Bark for natives in the NSS and the P & B data respectively, and 3.28 and 4.90 for learners in conversations and reading lists. With lip-rounding lenghtening the vocal tract, which in turn leads to a lowering of F_2, it cannot be said that learners tend to round their lips more than the natives, at least from these casual pieces of evidence. Predictions of /ʊ/ as either /uː/ or /ɔː/ may therefore not be attributable to excessive lip-rounding by learners.

phonological distance, defined here as the number of shared phonological features, is great between /ʊ/ and /e/: apart from the similar aperture of the mouth, the lip-rounding feature and the places of articulation are different, if not opposite. Such a variety in the selection of best alternates for /ʊ/, along with the phonological inconsistencies of these best alternates, starkly contrasts with the uniform arrays of best alternates for /ɪ/ and /iː/. This situation is pointedly not unlike that of /uː/: one expected best alternate for this monophthong, mirroring the best alternate for /iː/, would be /ʊ/. /ʊ/ is however the best alternate for /uː/ in only 4 out of all 12 cases. In the remaining cases, two front vowels, /e/ and /ɪ/, come out as best alternates, on 2 and 6 occasions respectively. The same remarks as /ʊ/ can be made regarding the phonological distance separating the best alternates from the actual monophthong. If the overall proportions of accurate predictions are similar for /ʊ/ and /uː/ on the one hand, and for /ɪ/ and /iː/ on the other, the phonological consistency of best alternates is greater for the latter couple of phonemes, than for the former. This statement should be slightly qualified: the gaps in consistencies only happen along the F_2 axis, *i.e.* in terms of places of articulation, rather than along the F_1 axis, in degrees of mouth aperture. Conversely, consistent best alternates only vary from their actual referent in terms of mouth aperture.

Another piece of information likely to reveal whether underlying differences exist between /ɪ/, /iː/, /ʊ/ and /uː/ when it comes to classification accuracy using the KNN algorithm is the proportion with which the best alternates were predicted. Whether the proportions of prediction of the best alternates are higher than those for the actual phonemes may also provide insight into those potential underlying differences. The proportions of actual predictions are inferior to those of best alternates in: 5 cases for /ɪ/; 11 cases for /iː/; 4 cases for /ʊ/; and 11 cases for /uː/. Based on these observations, a conclusion could be that the two lax vowels are pronounced in a way which is closer to native values than their tense counterparts. However, such a conclusion might be considered hasty if the two following parameters are taken into account: *(i)* the count of phones, which is much lower for /ʊ/,

makes the comparison between /ɪ/ and /ʊ/ somewhat fragile. *(ii)* likewise for /iː/ and /uː/, adding the actual and alternate predictions returns significant differences: for /uː/, the summed proportions are higher than 50% of all predictions in only one instance. This means that actual occurrences of /uː/ are predicted as phonemes other than /uː/ and the best alternate in the majority of instances. The situation is radically different for /iː/, where the summed proportions of actual and alternate predictions fall under 50% in one instance, and are higher than 60% in 7 cases out of 12. Bearing in mind the consistency of best alternates with /ɪ/ and /iː/, it looks as if the experiment with the KNN algorithm confirms the existence of differences in the levels of acquisition of /ɪ/, /iː/, /ʊ/ and /uː/.

Conclusion & future research

The experimental design of this section is unconventional: using a separate dataset as training set is not something commonly done in the fields (*e.g.* behavioural sciences or biology) that resort to KNN algorithms for research. It is however hoped here that the theoretical reasoning underlying the choice to use the native P & B dataset will be found sound. The results, *i.e.* the extent to which the actual occurrences of the monophthongs under study were correctly predicted by the algorithm, reveal the complexity of the processes at hand in conversational speech: no clear cross- or inter-speaker patterns, along with cross- or inter-phonemic patterns have emerged. It is contended here that this absence of patterns only partially due to the unconventional experimental set-up.

Once again, the results for /ɪ/ and /iː/ are more robust and consistent than those for /ʊ/ and /uː/ – albeit in a subtle way. These differences are far from being clear-cut, but the look at the best alternates in the predictions make them apparent, and, arguably, significant. Once again too, the differences in the number of occurrences across the four phonemes, with /ʊ/ featuring the lowest count, question the validity of cross-comparisons. Let it be emphasized

that such is the nature of conversational speech – a skewed distribution of tokens, and that this bias must also be taken into account.

Finally, this study could be extended by including the corpus of the recorded text in French: one way to go about investigating the quality of the learners' realizations would be to use the values of the French /i/ and /u/ as the training set, combined with those of /ɪ/, /iː/, /ʊ/ and /uː/ of the P & B corpus in turn, and examine the way the occurrences in the main corpus are predicted. Of course, another venue of research could also be to carry out the exact same experiment, only with a considerably more substantial native corpus of conversational speech.

Having resorted to the KNN algorithm to investigate potential differences in the states of acquisition of /ɪ/, /iː/, /ʊ/ and /uː/, it is now time to take a look at linear mixed-effects regressions, which are the object of the next section.

3.4 A longitudinal effect? An LMER analysis

This section uses linear mixed-effects regression (henceforth, LMER) to assess the evolution of the acquisition of the /ɪ/-/iː/ and /ʊ/-/uː/ contrasts. The focus here will be on longitudinal acquisition, *i.e.* the `SESSION` parameter in the main corpora will be used as a time predictor. As there are no reasons to assume temporal change is linear, models computing non-linear change with polynomials are also explored. However, with four sessions, the number of fixed effects must be kept under that of observed values to avoid saturated models (*c.f.* (Long, 2012, :119)). Because the order must be "*at least* two *less than the number of possible time points*" (p. 323), only quadratic polynomials will be investigated to model change over time. All calculations were made using the `lme4` package (Bates et al. (2015)). The response variables in the following sections are the Euclidean distance in the vowel space from native values to learner values, and the difference between native and learner OODs. Because they were extracted following the same procedure as the data in

the main corpora, the native reference points are the gender-dependent means of the NSS datapoints. Depending on the method used to analyze the fromant values (*i.e.* mid-temporal measurements, VISCs or DCTs), the native values have been converted accordingly. Once again, the formant values have been normalized using the BDM method, because it is vowel-intrinsic, and because it factors in F_3 values. The Euclidean distance between native mean values and each learner's datapoints provides an arguably reliable metric to assess actual acquisition. LMERs have been chosen as the privileged method of multimodel analysis for the following reasons: *(i)* simple linear models cannot be used because of the longitudinal nature of the data. With per-subject repeated measures, the fundamental assumption that datapoints must be independent is violated. *(ii)* the unequal number of items under each level means there are a lot of missing values. LMERs can handle missing data. *(iii)* the nested, hierarchical structure of the data calls for the analyses of between-groups (*i.e.* `SEX` or `LPDPHONEME`), within-groups, between-subjects and within-subjects effects. LMERs make these analyses possible. When comparing models, package `AICcmodavg` (Mazerolle (2017)) has also been used to compare the Akaike Information Criterion (AIC), the Deltas and the weights of evidence.

In this study, two response variables, which echo the work done in sections 3.2.1 and 3.2.2, are going to be investigated in turn: first, the distance between the BDM-normalized /ɪ/, /iː/, /ʊ/ and /uː/ learners' points in the F_1/F_2 vowel space, and the natives', is investigated. The natives' formant values were averaged across all occurrences for each sex. The second response variable is the difference in OODs between natives and learners (independently from the location of the monophthongs in the vowel space). The datasets used to compare the models work along the same lines: each datapoint includes the specific OOD difference and the distance from the corresponding, sex-dependent, but also *syllable-dependent* mean native vowel. As an example, each occurrence of /iː/ by a male learner will be compared to the mean male /iː/ native values pronounced *in the same syllable* as the one it appeared in

in the main corpus. The reason why learner values were compared to those obtained by the natives in the same syllable rather than in the same word is because some words may contain the same mophthong twice or more, *e.g.* "artistically" (/aːˈtɪstɪkəli/), present in the corpus. The two /ɪ/ appear in syllables with different structures in terms of coda, onset and stress. All these differences have not been formally modelled (*i.e.* categorical variables encoded in the corpus such as *x*SYLLSTRUC, STRESS or *x*SKELS have not been used as effects in the LMERs), but their potential influence was at least partially taken into account by calculating learner-to-native distances from values for phones embedded in the same syllables – rather than words only.

3.4.1 Purpose and issues: a warning

The main purpose of resorting to LMERs here is to carry out a longitudinal analysis, and to investigate whether differences exist between the /ɪ/, /iː/, /ʊ/ and /uː/ vowels. If the time predictor to use, *i.e.* SESSION, is straightforward, the highly nested and heterogeneous nature of the data requires caution when selecting predictors. Emulation of longitudinal analyses in other fields (*e.g.* medicine and behavioural sciences especially) was implemented as rigorously as possible, but the unequal numbers of datapoints, the likelihood of hidden interactions (within and between words, regardless of subjects, for instance), the potentially high number of predictors likely to increase the model fit, and the corresponding exponential increase in slope and intercept effets – all of this needs to be checked and controlled thoroughly. With these provisos in mind, the key issue here lies in the status of the categorical variables LPDSYLL and WORD in the equation of the models. If these two variables could arguably be considered as a static predictors, their potential independence from subjects IDs, *i.e.* their non-nested nature, but also their sheer number of categories, make it unsafe and difficult to use them as predictors. This state-of-affairs underpins the decision to restrict the corpus. In an attempt to work around, if not solve, these issues, the corpus was

restricted to occurrences of /ɪ/, /iː/, /ʊ/ and /uː/ embedded in *syllables* which were *also in the NSS*. Choosing `LPDSYLL` over `WORD`, *i.e.* only selecting the occurrences of /ɪ/, /iː/, /ʊ/ and /uː/ in syllables also present in native occurrences of the matching sex, allows for finer-grained analysis - and more datapoints (14,433 for matching syllables against 11,205 for matching words). This feature factors in, and neutralizes the issue of, the variety of consonantal environments likely to exert influence on the formant values. This means that the distance from each datapoint in the (BDM-normalized) F_1/F_2 space to the native datapoints is based on syllable-specific, sex-dependent, values. This makes it possible for `LPDSYLL`, *i.e.* the categorical variable listing all the syllables containing the occurrences of the four monophthongs, not to be included as a predictor in the equations of the models, thereby considerably reducing the amount of calculations. The extent to which this work-around to account for the great variety of the data is a crucial point that will be discussed later. The

	Session 1				Session 2				Session 3				Session 4			
Speaker	/ɪ/	/iː/	/ʊ/	/uː/	/ɪ/	/iː/	/ʊ/	/uː/	/ɪ/	/iː/	/ʊ/	/uː/	/ɪ/	/iː/	/ʊ/	/uː/
DID0014	98	74	5	50	102	68	5	43	259	173	18	53	191	158	8	142
DID0024	56	61	2	23	122	72	12	55	158	123	14	44	288	150	16	150
DID0035	144	97	8	46	180	142	7	92	165	103	8	48	116	70	11	76
DID0039	172	107	8	78	211	140	10	117	202	120	7	34	161	87	5	112
DID0062	91	87	7	7	74	68	11	21	110	70	3	28	118	68	4	114
DID0068	92	84	4	31	97	92	6	41	163	89	9	36	92	62	1	58
DID0071	77	51	2	17	108	86	2	38	220	117	5	28	107	77	4	61
DID0096	103	86	4	28	76	57	5	42	122	74	5	37	126	79	6	81
DID0106	86	69	3	27	159	97	12	69	216	134	24	67	134	76	7	123
DID0108	82	57	3	29	156	117	3	34	197	130	6	49	105	63	6	72
DID0128	74	54	0	51	219	143	14	77	182	104	12	72	169	142	6	110
DID0168	84	59	7	53	143	97	13	70	204	139	21	59	55	43	0	99
TOTAL	1159	886	53	440	1647	1179	100	699	2198	1376	132	555	1662	1075	74	1198

Table 3.4: Number of occurrences of /ɪ/, /iː/, /ʊ/ and /uː/ in the dataset of words common to both the main corpus and the NSS.

per-speaker, per-session number of occurrences of each of the four monophthongs under study is presented in table 3.4. Once again, the number of /ʊ/ items is comparatively much lower than the numbers of its counterparts, with no occurrences in two sessions, speakers DID0128 and DID0168 in sessions 1 & 4 respectively. The total number of datapoints across speakers and sessions is 14,433, which can be broken down as follows: 6,666 occurrences of /ɪ/, 4,516 occurrences of /iː/, 359 occurrences of /ʊ/ and 2,892 occurrences of /uː/. The total number of syllables common to both native and learner data is 168.

In the next subsections, the details are provided of the theoretical questions and choices underlying the designing of the LMER models. Issues regarding the response variables, the fixed effects and the random effects will be dealt with in turn.

The response variables

Originally the purpose was to find response variables likely to capture the essence of "phonemic acquisition". As was hopefully shown, if only tentatively, in the previous sections, this essence is somewhat elusive, from both a theoretical and practical point of view. The two response variables chosen here, *i.e.* the mid-temporal distance in the BDM-normalized F_1/F_2 space between learner and native values taken from monophthongs appearing in matching syllables on the one hand, and the difference between native and learner OODs, with the same constraints, on the other, aim to solve the issue by complementing one another and reducing the number of parameters with possible intercept or slope effects, while still mirroring the intrinsic complexity of the data. Technically the response variables were calculated using the following procedure (only the steps common to the two response variables are described here):

1. Selecting the speakers with four sessions, and the four phonemes (/ɪ/, /iː/, /ʊ/ and /uː/);
2. normalizing the F_1, F_2 and F_3 values at the 20^{th}, 50^{th} and 80^{th} centiles using the BDM method in the main corpus and the NSS;
3. merging the obtained datasets by their `LPDSYLL` columns, *regardless of session occurrences:* this means that *any* phoneme from the learner data will be included provided it has a syllabic match in the native data. There is no condition that the phoneme and its syllable structure should appear in all four sessions (this would in effect reduce the dataset to a non-workable array of the most frequent words such as "it" and "too");

4. averaging over the per-speaker, per-session values. This returned a 190-row dataset: one single /ɪ/, /iː/, /ʊ/ and /uː/ value per session and per speaker, with 2 missing rows for speakers DID0128 and DID0168 (*c.f.* above). Averaging over all the different syllable types is a necessary step to prevent nesting and uncontrollable subsampling. Theoretically speaking, averaging captures what a phoneme is, *i.e.* the specific acoustic signature common to all the phoneme's occurrences regardless of contextual information.

With all that done, however, one key issue from the perspective of this work remained: the discrepancy between the numbers of occurrences from one phoneme to another is distinctive enough to be highly problematic. How justified is it *not* to take into account this information? The top panel of figure 3.13 plots the per-phoneme number of types against the per-phoneme

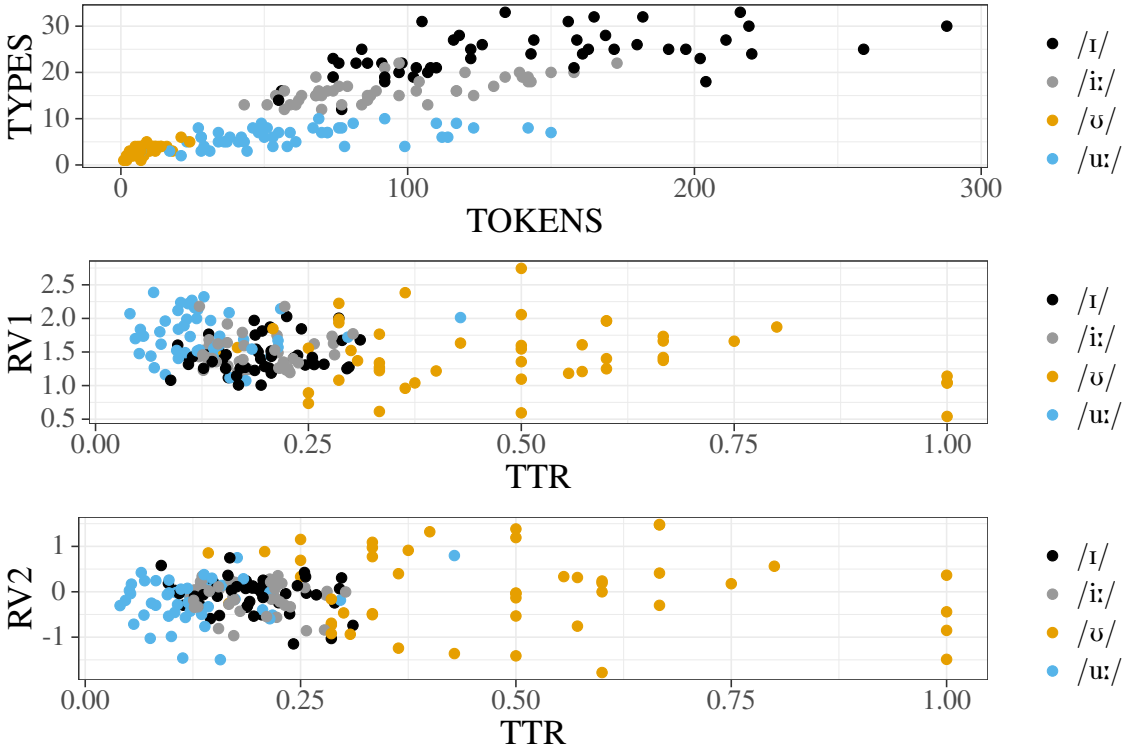


Fig. 3.13: *Top panel:* Number of types against number of tokens; *middle panel:* distance to native values in Bark (**R**esponse **V**ariable 1) against the TTRs; *bottom panel:* differences in OODs (**R**esponse **V**ariable 2) against the TTRs.

number of occurrences. In all panels, each dot corresponds to a learner's mean number of

types and tokens of a given phoneme for a given session. The figure clearly shows that the distribution of the dots in the Types/Tokens space is dependent on the phoneme type. The major pitfall at this stage is the risk of self-confirmatory bias: if the researcher believes that frequencies and varieties of uses matter, then somehow the information will be added in the models; if she does not, *i.e.* if she believes only information at the phonemic level matters when dealing with phonemic acquisition, then the information will be left aside. This stance is that of most theories in SLA today. Two questions therefore arise: *(i)* how to include linguistic, extra-phonemic information, such as types and tokens, knowing that they are continuous variables? *(ii)* Can models based on LMERs disprove one or the other position mentioned above?

At this stage of the research, in order to answer those questions, the only solution seems to be to make TTRs, or tokens, or types, part of the response variable. However, if such a solution can be argued to make some mathematical sense, it makes little theoretical sense: not only would the response variable be too opaque, but the nature of the operation linking the TTRs or their components to the response variable is unclear: if choosing types over tokens, or vice versa, then part of of the linguistic information is left aside; if choosing TTRs, then should they be multiplied, or divided, by the response variables? The middle panel and bottom panel of figure 3.13 indicate that the TTRs of /ʊ/ are in general higher, and more distributed along the *y*-axis. Considering that the values of the two response variables are a direct indication if not of acquisition, at least of closeness to native values, and that the lower both these response variables are, the better, two options seem viable: one the one hand, multiplying the response variables by the TTRs; on the other hand, dividing the response variables by *both* the number of tokens and the number of types. In both cases, the response variables of /ʊ/ would be penalised. From a theoretical point of view, the first solution is preferable: TTRs are a well-known parameter in linguistics. If multiplying TTRs by the response variables makes them too opaque, at least that opacity is limited by the resort

to another readable parameter. Besides, mathematically, the second solution would lead to arguably extreme penalisation of /ʊ/ values, as shown in figure E.5 (an optional graph in section E.5). The first solution is therefore retained (it is more readable theoretically and less extreme mathematically). But because of the resulting opacity, which may lend itself to not illegitimate accusations of "data torture", models using the plain, original response variables will also be analyzed. From now on, the following terminology will be used:

1. RV1, response variable n°1, will refer to the distance between native and learner values in the BDM-normalized F_1/F_2 space;
2. RV1*c* stands for the TTR-*corrected* RV1, *i.e.* the original RV1 value multiplied by the TTR;
3. RV2, response variable n°2, will refer to the difference between the BDM-normalized native and learner OODs;
4. RV2*c* refers to the TTR-corrected RV2, *i.e.* the product of RV2 and the corresponding TTRs.

All non-corrected variables Figure 3.14 plots the resulting mean individual curves of the four response variables for each phoneme in grey, with the mean curve for all speakers in a thicker black line. The top row plots RV1, the second row, RV1*c*, the third row the absolute value of RV2, and the fourth row the absolute value of RV2*c*. Examination of the graphs shows greater dispersion for RV1, corrected or not, than for RV2. /ʊ/ also clearly stand out against the other phonemes for its much greater variability between speakers. Corrected response variables also show less dispersion than their uncorrected counterparts, without however totally changing the overall profiles of the responses: the level of applied correction does not modify the data to a point where a link between the original data and its corrected version could not be established. The effects of TTR-correction change from one response variable to the other: in the case of RV1, correction increases discrepancies between phonemes, especially for /ʊ/, both in terms of overall curve shapes and values of

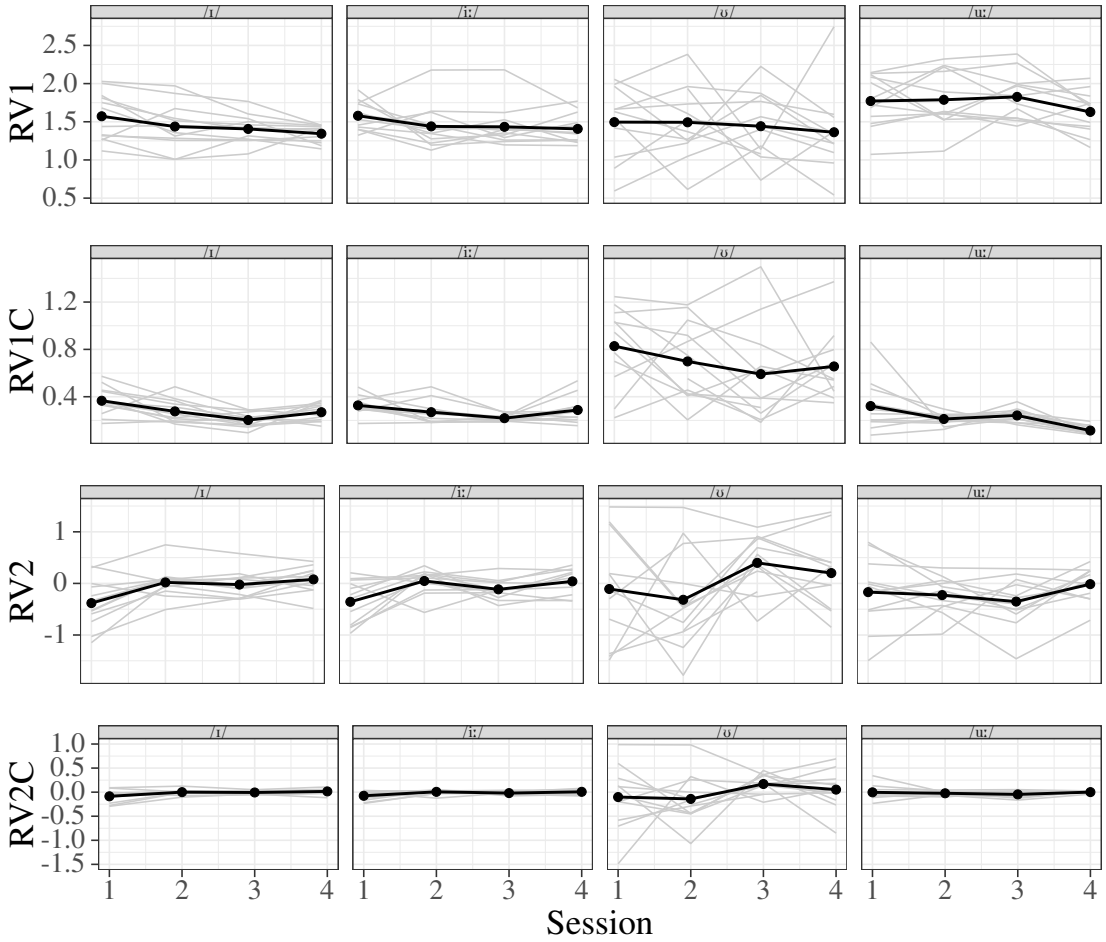


Fig. 3.14: Per-session means of the 4 response variables, with mean individual lines (in grey).

RV1. Conversely, RV2*c* flattens the mean curves and levels the differences in values between phonemes. Both corrected response variables, however, seem to display less individual dispersion than their uncorrected counterparts. This is especially visible with /uː/, whose individual curves show great disparities between one another with the uncorrected response variables; these disparities are leveled out with RV1*c* and RV2*c*. The graphs do not reveal any unified temporal evolution. /ɪ/ and /iː/ curves look very similar, with /uː/ featuring a very similar outlook. When a slope is visible, as for instance with RV2 or the /ʊ/ and /uː/ curves, the general trend seems to be decreasing. The RV2*c* curves for /ɪ/ and /iː/ look like flat lines. This cursory visual analysis begs the question whether a longitudinal effect exists, and if there is one, to what extent is it similar from one phoneme to another? These

questions are the ones our resort to LMERs will be trying to answer. Let us finish this section by mentioning that at the very least, the /ʊ/ curves look dissimilar enough from their peers for other phonemes to say suggest that the acquisition of /ʊ/ is very likely to undergo a specific evolution.

Having provided details about the response variables, it is now time to turn to fixed effects.

3.4.2 Fixed and random effects

The main purpose here is to find out whether there exists a longitudinal effect, and the extent to which a phoneme-dependent effect exists. Fixed effects need to be established from a linguistic, theoretical point of view. The study being longitudinal, the time predictor is `SESSION`. The `SESSION` values, originally "`S001`", "`S002`", "`S003`" and "`S004`" were converted to a dummy numeric variable with values 0, 1, 2 and 3 respectively, so as to have intercepts at Session 1. Because response variables are already sex-dependent, `SEX` is not included as a fixed effect. The issue now is to determine the status of `LPDPHONEME`. It is argued here that `LPDPHONEME` is *not* a fixed effect, as it does not apply to the population under study (*i.e.* the learners). This is in keeping with longitudinal studies in other fields: treatment studies in medical sciences typically compare the evolution of a response variable in a group following a given treatment against a base-line group either following none or another. In those instances, `TREATMENT` is used as a fixed effect because it exhausts the population under study, *i.e.* the patients: one half of the population is following the treatment, the other half is not. Likewise, in behavioural sciences, repeated measures of a response variable such as reading scores will be modeled with fixed effects providing information on the students' academic, social and/or ethnic backgrounds. Once again, these fixed effects categorize the the population under study. This is (emphatically) not the case for`LPDPHONEME`, which categorizes response variables. However, it still makes sense to assume that there might be

differences both in intercepts and in slopes caused by either SESSION and/or LPDPHONEME. The way this will be assessed is by modeling various time changes (none, linear or quadratic) with per-phoneme response variables, *i.e.* by pre-selecting the datapoints corresponding to a given phoneme and comparing models predicting various sorts of time changes. No fixed effects have therefore been selected: LPDSYLL, pertains to the same sort of factors as LPDPHONEME, and so do variables such as SYLLSTRUC, CVSTRUC or SKELS.

When it comes to random effects, random intercepts and slopes for every speaker are posited by default when it is possible. The focus of the study being the influence of the time predictor and/or of the static predictor on the response variables, random effects are selected on two bases: *(i)* enabling model comparison; *(ii)* theoretical viability. The first provision rests on the following advice by (Long, 2012, p. 324): "*when considering the selection of time transformations, the number of static predictors and random effects is held constant among the models. (. . .) The reason is that interpretations are clearer if there is one influence on model fit*". This entails that a pure intercept model can only be compared to models with intercept random effects. When the correlation between subjects and sessions is posited, it is assumed that each learner starts with a given distance from the mean native values (intercept), and as sessions unfold, this distance evolves in an idiosyncratic manner (slopes). Let it be clearly stated at this point that both the method and the findings are *exploratory*, if only because of the robustness of the NSS values, whose relatively small number implies their means are but indicative.

It is now time to describe the models in detail.

3.4.3 Models

In order to steer clear from any potential risk to "torture the data", the approach here has resolutely been deductive rather than inferential: it is our view that the complexity of the data should preclude results-driven approaches. The terminology in the equations below matches

the conventional notation in algebraic formulas, with β being a component of fixed effects and as such a regression coefficient; b a component of random effects marking individual deviations from fixed effects; and ε being the random error, *i.e.* the regression error term. The subscripts $_i$ and $_j$ indicate the speakers' ID and the session time points respectively.

Table 3.5: Working hypotheses and statistical models

Name	Working Hypothesis	LMER Model
M1	Intercept effect of phonemes; no over-time change	$y_{ij} = (\beta_0 + b_{0i}) + \varepsilon_{ij}$
M2	Linear change with random intercepts	$y_{ij} = (\beta_0 + b_{0i}) + \beta_1(SESSION_{ij}) + \varepsilon_{ij}$
M3	Quadratic change with random intercepts	$y_{ij} = (\beta_0 + b_{0i}) + \beta_1(SESSION_{ij}) + \beta_2(SESSION_{ij}^2) + \varepsilon_{ij}$
M4	Linear change with random slopes	$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})SESSION_{ij} + \varepsilon_{ij}$
M5	Quadratic change with random slopes	$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})SESSION_{ij} + \beta_2(SESSION_{ij}^2) + \varepsilon_{ij}$

Table 3.5 summarizes the models compared. The *R* code snippets for the models can be found in section C.8. Model n°1 predicts a phonemic, intercept effect only, with no over-time changes (*i.e.*, a flat line). M2 models linear change, with no random slopes. Likewise, M3 corresponds to quadratic time change with no random effects on curvature. M4 and M5 introduce random effects on slope and curvature. M1, M2 and M3 on the one hand, and M4 and M5 on the other, will be compared together. Following (Long, 2012, p.246 and ff.), in order to compare the models, the weight of evidence is used. the weight of evidence is the probability that a model is the best approximating one in the set of models being compared. Table 3.6 lists the best fitting models with respect to phonemes and response variables. The weights of evidence are given in the column named AICcWt. The p-value is the one obtained by carrying out a Shapiro-Wilk test on the residuals of the best fitting model. p-values here correspond to the probability for the distribution of the residuals of the best fitting model to follow a normal distribution. Low values, for instance inferior to $\alpha = 0.05$, indicate that even the best fitting model in the multimodel comparison somehow fails to capture the data in a satisfactory manner. Low p-values are observed with corrected response

Table 3.6: Per-phoneme, per-response variable results of the multimodel comparisons. *AICcWt:* weight of evidence; *p-value:* p-value of the Shapiro-Wilk test carried out on the residuals of the fitted models.

Phoneme	Response Variable	Model Number	AICcWt	p-value
ɪ	RV1	M2	0.71	0.38
ɪ	RV1	M4	0.70	0.84
ɪ	RV1c	M3	0.95	0.58
ɪ	RV1c	M5	0.99	0.24
ɪ	RV2	M3	0.71	0.99
ɪ	RV2	M5	0.80	0.91
ɪ	RV2c	M3	0.60	0.04
ɪ	RV2c	M5	0.73	0.86
iː	RV1	M2	0.55	0.11
iː	RV1	M4	0.68	0.14
iː	RV1c	M3	0.81	0.00
iː	RV1c	M5	0.95	0.00
iː	RV2	M3	0.51	0.44
iː	RV2	M5	0.51	0.35
iː	RV2c	M3	0.53	0.07
iː	RV2c	M5	0.56	0.14
ʊ	RV1	M1	0.65	0.70
ʊ	RV1	M4	0.79	0.85
ʊ	RV1c	M1	0.41	0.05
ʊ	RV1c	M4	0.70	0.30
ʊ	RV2	M2	0.46	0.47
ʊ	RV2	M4	0.80	0.23
ʊ	RV2c	M1	0.51	0.02
ʊ	RV2c	M4	0.79	0.04
uː	RV1	M1	0.37	0.68
uː	RV1	M5	0.53	0.18
uː	RV1c	M2	0.77	0.00
uː	RV1c	M4	0.79	0.00
uː	RV2	M1	0.55	0.18
uː	RV2	M5	0.60	0.17
uː	RV2c	M1	0.63	0.00
uː	RV2c	M5	0.65	0.32

variables exclusively: for /ɪ/, the selected quadratic model for RV2c^i[8] features residuals that most likely do not follow a normal distribution; likewise, for /iː/, RV1c^i and RV1c^s; for /ʊ/, RV1c^s and RV2c^i and RV2c^s; for /uː/, RV1c^i, RV1c^s and RV2c^i. One commonplace, if opaque, way to solve the issue of non-normally distributed residuals, is to log-transform the response variables. The obtained results can be found in table E.2: no improvements

[8] The superscript letter will indicate what random effects were tested on the response variables: ^{i}, as in RV1c^i will refer to intercept random effects, *i.e.* tested in models M1, M2 & M3; ^{s}, as in RV1^s, will refer to intercept and slope random effects.

have been made with respect to the distribution of the residuals. More models even show non-normally distributed residuals[9], so that the log-transformation does not improve the models fits. 10 out of the 16 models using a corrected response variable present residuals whose distribution is very likely not normal. In the rather conservative view held here, it appears that the models as they have been defined fail to capture corrected data. At this point of the research, this means that the considerable discrepancies in types and tokens that are observed among the four phonemes under study cannot be modelled. How to do so is best kept for future research, let it simply be reminded that it is contended here that these discrepancies are linguistically significant, and should somehow be modelled. The (hopefully temporary) failure to do so leads us to discard the models with corrected response variables, and focus exclusively on the models predicting the plain response variables. Even without factoring in the types and tokens of the phonemes, can differences in the modelled rates of acquisition between the different phonemes be observed? This is the object of the next section.

3.4.4 Results

This section discusses the results obtained from the predictions of the models using uncorrected response variables. Figure 3.15 plots the observed (square dots) and fitted (lines) responses variables for each session and each phoneme. The columns give the obtained results for /ɪ/, /iː/, /ʊ/ and /uː/ respectively. Rows 1 & 3 correspond to response variables 1, rows 2 & 4, to response variables 2. The top two rows show data with models using random intercepts only; the bottom two rows show data using random slopes. As a reminder, the lower the RV1, the closer it is to native values in the BDM-normalized F_1 / F_2 space. For RV2, the closer to 0 (materialized by a dotted lines in panels in rows 2 & 4), the smaller the difference between native and learner OODs. Odd-numbered rows are expected to be

[9] Or rather, with very low probabilities of presenting normally distributed residuals.

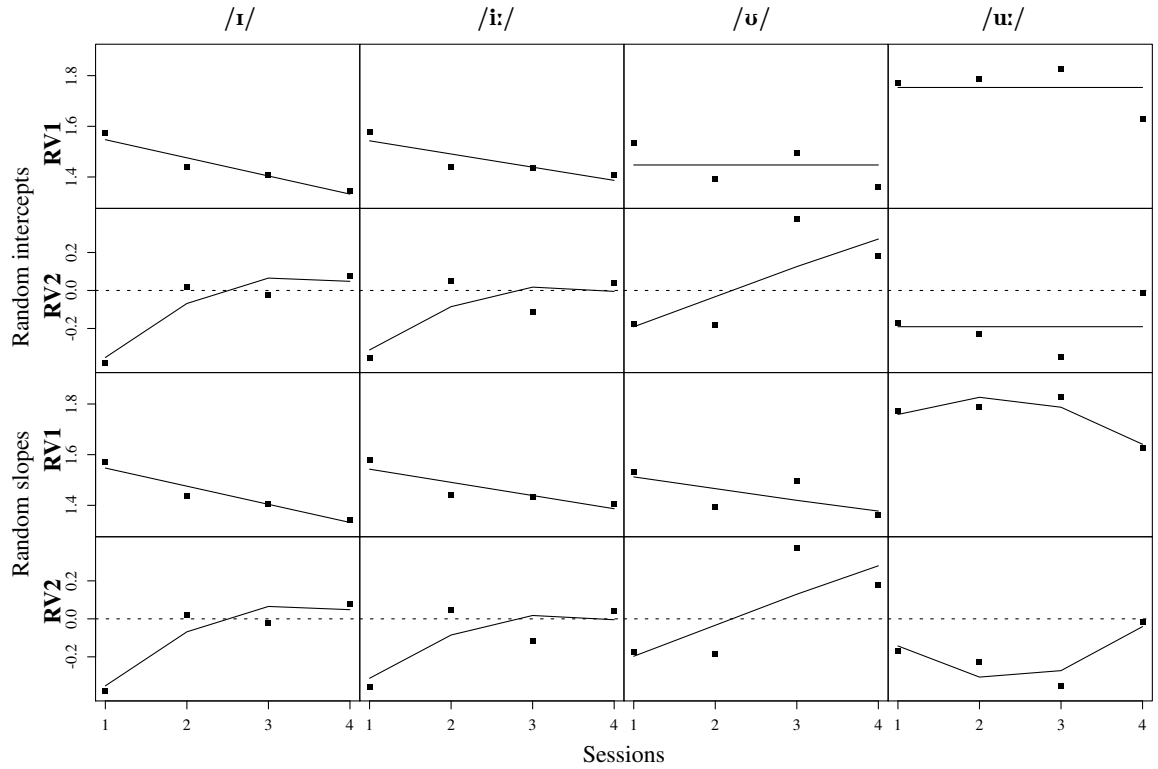


Fig. 3.15: Plots of the mean fitted response variables. *Top row:* RV1 with random intercepts; *row 2:* RV2 with random intercepts; *row 3:* RV1 with random slopes; *row 4:* RV2 with random slopes. Square dots indicate the mean observed values across speakers for each session. Lines show the fitted values. The dotted lines in rows 2 & 4 indicate a null distance between native and learner OODs.

rather similar to one another, and so are even-numbered rows: the change from random intercepts to random slopes in the models should only return different predictions for highly dispersed data. This is indeed the case for /ɪ/ and /iː/, whose values are fitted by the same linear or quadratic models regardless of random effects. /uː/ on the other hand has values so dispersed (*c.f.* figure 3.14) that models change between models with random intercepts and models with random slopes. /ʊ/ however features no difference in choices of linearity for RV2. This is surprising, considering the extreme variations between- and within-speakers. It is hypothesized in this case that past a certain degree of dispersion, the deviation from fitted values is so high that changes in random effects are only marginal. The evolutions of /ɪ/ and /iː/ seem pretty robust. They are all evolutions that tend towards native values. In the case of RV1, the distance from native values decreases regularly with time. The mean

observed values are close to the fitted line, with the variance of residuals at $\sigma^2 = 0.025$ and $\sigma^2 = 0.031$ for /ɪ/ and /iː/ respectively. The fitting deviation is slightly smaller with random slopes models for the two phonemes, $\sigma^2 = 0.015$ and $\sigma^2 = 0.030$. The two types of best fitting models, with either random intercepts or slopes, predict linear decreasing change for /ɪ/ and/iː/. The cases of /ʊ/ and /uː/ are different. They feature differences between the two sorts of models. Models with random intercepts predict no evolution over time for either phoneme: unlike the regular decrease for /ɪ/ and /iː/, a stagnation is predicted. Variances of the residuals are also higher: $\sigma^2 = 0.17$ for /ʊ/ and $\sigma^2 = 0.052$ for /uː/. The predictions of models with random slopes offer different interpretations, however: the fitted values of /ʊ/ show the same decreasing line as /ɪ/ and /iː/, albeit with a slightly gentler slope ($\beta_1 = -0.046$ against $\beta_1 = -0.072$ and $\beta_1 = -0.052$ respectively). The residual variances are similar, $\sigma^2 = 0.17$ for /ʊ/ and $\sigma^2 = 0.044$ for /uː/. The overtime evolution for the latter is quadratic for models with random slopes, however, with $\beta_1 = 0.12$, et $\beta_2 = -0.054$. With residual variances comparable in the two models, yet with different predictions, it is difficult to give either solution prevalence and draw even tentative conclusions about the acquisition of the two phonemes and the evolution of their distances from native values in the BDM-normalized F_1 / F_2 space. A trend towards a reduction might exist, but the results are arguably less robust than for /ɪ/ and /iː/. With three phonemes (/ɪ/, /iː/ and /ʊ/) presenting the same predictions across the two sorts of models (quadratic with a negative second-order coefficient for /ɪ/ and /iː/, linear with a positive slope coefficient for /ʊ/), the results for these three phonemes may be considered as more robust for the measurements of OODs (rows 2 & 4 of figure 3.15). If the findings are correct, then the evolution and state of acquisition of these phonemes are different: both /ɪ/ and /iː/ start with OODs inferior to native values, and those OODs get nearer to native OODs from the second session onwards, and remain close to native values. The fitted model of /ʊ/ shows no such adjustment towards native values: sessions 1 & 2 feature undershot OODs, with values in sessions 3 & 4 overshooting native

values. No convergence towards native values is visible for the OODs of /ʊ/ unlike those of /ɪ/, /iː/ – or /uː/: if the model with random intercepts predicts no change over time for /uː/, with predicted values at $\beta_0 = -0.19$, the mean observed value for session 4 sits at -0.014 Bark (against 0.078 and 0.040 for those of /ɪ/ and /iː/ respectively). The model with random intercepts is therefore probably more robust, with a residual variance of $\sigma^2 = 0.12$ against $\sigma^2 = 0.17$ for the model with random intercepts. The fitted curve is quadratic with a positive second order coefficient, $\beta_2 = 0.099$, lending itself to the interpretation that after a lapse in sessions 2 & 3, the OODs tended towards native values in the last recording session.

This coarse study may serve as basis for further exploration using LMERs. The focus was here on time change, and on whether the rate of phonemic acquisition changed between /ɪ/, /iː/, /ʊ/ and /uː/. The rate of acquisition was measured here both by the distance between learner and native mid-temporal, BDM-normalized formant values in the F_1 / F_2 space, and the differences between learner and native OODs. The findings seem quite robust with /ɪ/ and /iː/, as very few differences exist between random intercept and slope models. For these phonemes, the tendency is clear for measurements to get closer and closer to native values with time. The fits of the models in the cases of /ʊ/ and /uː/ are looser, with differences in the predictive slopes between random intercept and slope models for the two measurements with /uː/ and for the mid-temporal distance for /ʊ/. The consistent predictions for OODs with /ʊ/ across the two types of models do not indicate better acquisition, but rather divergence away from native values. However, the fitted decreasing slope of distances in both sorts of models may indicate a degree of acquisition, but the substantial dispersion makes those predictions less robust than in the case of /ɪ/ and /iː/. With higher distances from native values, but a possible improvement in OODs in the last session according to the random slope model, the case of /uː/ offers a complex landscape. Added to the similarly contrasted outlook of fitting models for /ʊ/, it seems safe to assert that the LMER models used in this

study contribute to show that the /ɪ/–/iː/ distinction is better acquired than the /ʊ/–/uː/ contrast.

The next and final section of this chapter deals with Discrete Cosine Transformations.

3.5 Discrete Cosine Transformations

This section uses Discrete Cosine Transformations to model formant tracks, and investigates the differences with native values from the NSS. The first subsection justifies the resort to these mathematical transformations; the second subsection describes the experimental set-up and presents the results. Finally, a conclusion and venues of future research are given in a third subsection.

3.5.1 Presentation and justification of DCTs

One key issue that this whole work has been trying to address is that of finding out ways to represent and visualize unwieldy data. The structural complexity of conversational speech, with vowels embedded in a wide array of consonantal environments, combined with the wealth of information which scripts of automatic extraction enable to retrieve, make it necessary to resort to mathematical methods simplifying the data while retaining as much information as possible. The corpora under study here all feature formant measurements at every centile of the vowels' duration. For each formant, the measurements along the time axis form a signal as in figure 1.15 in Chapter 2, or figure 3.16 below. This signal is a formant track which can be mathematically approximated in several ways, by using for instance quadratic polynomials, fractional polynomials, splines or trigonometric functions. Provided the error rate is kept at a moderate level, modeling complex signals with such known mathematical functions reduce the number of parameters to study – a considerable advantage when dealing with a high number of parameters and datapoints.

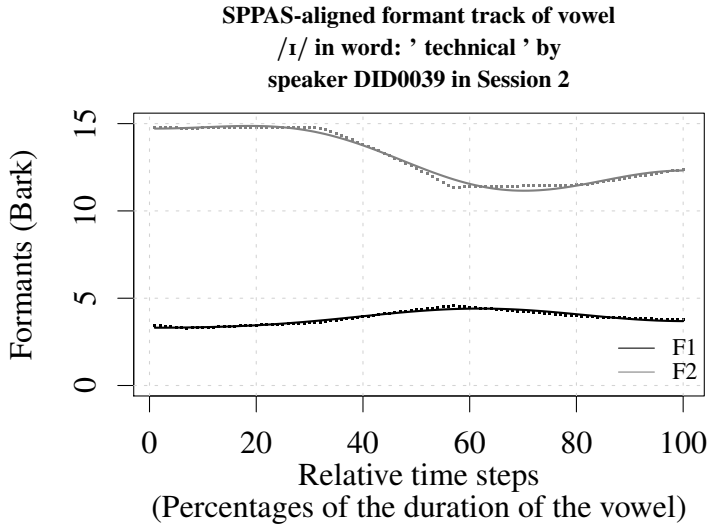


Fig. 3.16: Example of BDM-normalized F_1 and F_2 formant tracks (dots) with the superimposed DCT-smoothed signal (lines).

With such an array of modeling options to choose from, the risk exists again to torture the data, *i.e.* to adopt *post hoc* analyses that will return the desired results. Therefore the modelling method adopted here is that described in Harrington (2010), the Discrete Cosine Transformation (henceforth, DCT). For mathematical details, the reader is referred to Harrington (2010) (pp. 304-305). In a nutshell, the idea is to sum sinusoids to reconstruct the original signal, like discrete Fourier transforms. Technically, in DCTs, the sinusoids have no phase, and so are cosine waves. Such a way to model the signal is meaningful because the amplitudes of the first three cosine waves, named k_0, k_1 and k_2, are "*proportional to the signal's mean, slope and curvature respectively*" (Harrington, 2010, p.305). From the data-streamlining perspective developed here, this reduction of a one-hundred-point signal to three coefficients is extremely valuable, when these signals must themselves be integrated in further calculations, such a BDM-normalizations, themselves iterated over speakers, phonemes and sessions. From a more theoretical and phonetic point of view, using DCTs is in keeping with more recent research such as VISCs, already mentioned in section 3.2: the VISC theory contends that the onset and offset of vowels are perceptually crucial to their identification. By taking the entirety of the signal into account, DCTs make

it possible to represent the specificities of each vowel more precisely. The extent to which the *common ground* of each category of vowels, *i.e.* the particular features that differentiate them from other categories, is preserved, especially in learners' conversational speech, has yet to be assessed.

Figure 3.16 shows the DCT-smoothed signal of BDM-normalized F_1 and F_2 in one occurrence of /ɪ/ for speaker DID0039 in Session 2. The dotted curve is the raw signal, *i.e.* the 100 centiles that were collected when `PRAAT03` parsed the vowel. The smoothed curves can be reconstructed from the three amplitudes of the cosine waves, the DCT coefficients k_0, k_1 and k_2. When combined to BDM normalization, which includes F_3 centiles, these coefficients make a drastic reduction of datapoints possible: where 900 datapoints would otherwise have to be computed (100 datapoints for each formant), BDM-normalized, DCT-smoothed formant tracks for F_1 and F_2 only require…6. Figure 3.17 details how the smoothing procedure operates. It was generated using the *R* package `emuR` by Winkelmann et al. (2016), and snippets from Harrington (2010). All DCT-related codes used later on in this study were written using functions from the `emuR` package. In all panels, the x-axis corresponds to the centiles of the duration of the vowel, while the y-axis indicates the formant values in Bark. The left column plots the half-cycle cosine waves whose sums reconstruct the signal: the more cosine waves are added, the more smoothing is obtained, but also the more coefficients are needed. In this study, only the first three coefficients are retained. The top-left panel shows that the first DCT coefficient, k_0, corresponds to the mean value of the raw signal (here, $k_0 = 3.88$). The raw signal is the F_1 raw signal from figure 3.16. The scaling in figure 3.17 zooms in on F_1, which explains the seemingly different shapes of the curves. But the two F_1 signals in both figures are the same. The right column of figure 3.17 displays the progressive smoothing of the raw signal, plotted in dots. In the top-right panel, the smoothing is coarse and virtually non-existent, since the first cosine wave is a flat-line corresponding to the mean of the raw signal. In the middle right panel, the first two cosine waves from the

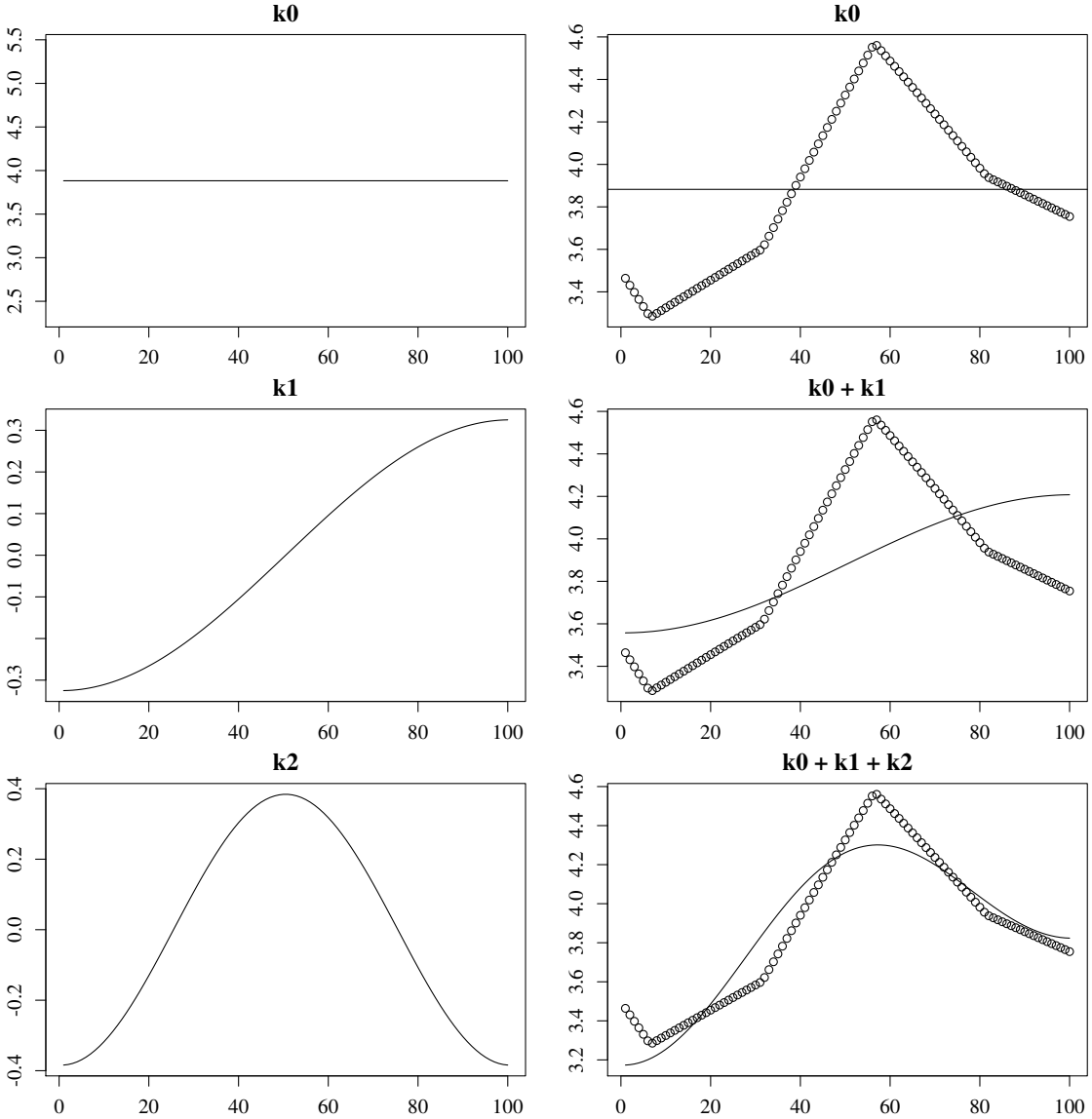


Fig. 3.17: *Left column:* Half-cycle cosine waves after applying a DCT to the to the F_1 raw signal of figure 3.16. *Right column:* Raw signal and DCT-smoothed signal with incremented summing of the cosine waves. *x*-axis: centiles; *y*-axis: Bark.

left column have been summed, yielding a smoother curve. Finally, the bottom-right panel features the curve corresponding to the sum of all the cosine waves from the left column. The obtained smoothing, with only three coefficients, returns a satisfying fit of the raw signal. The procedure to study each vowel is therefore the following: the raw formant tracks of F_1, F_2 and F_3, corresponding to 900 datapoints, are BDM-normalized into two F_1 and F_2 formant tracks, which include the F_3 information; then a DCT is applied to these two formant

tracks, limiting the number of cosine waves to three. Each of the BDM-normalized formant track can be retrieved from the three DCT-coefficients from which the cosine waves can be inferred, returning a total of 6 numbers to account for the original 900 datapoints of each vowel.

It looks as if DCT-smoothing and BDM-normalization have the potential to reduce the size of the data in a way that might still retain their finer-grained specificities. Having explained how the two procedures synergize, it is now time to examine whether they are relevant to assessing phonemic levels of acquisition.

3.5.2 Experimental design

This subsection explores the DCT coefficients of all speakers across all sessions for phonemes /ɪ/, /iː/, /ʊ/ and /uː/, and compares them with the native values from the NSS. Just like in previous sections, only monophthongs occurring in consonantal environments also present in the NSS were retained. This condition reduced the number of tokens under study to 11,323. They can be broken down as follows: 6,666 /ɪ/; 1,511 /iː/; 359 /ʊ/; and 2,787 /uː/. Each of these phonemes was then processed as per the guidelines described in section 3.5.1, *i.e.* the formant tracks were BDM-normalized, and then a DCT was applied to them. This returned $6 \times 11,323 = 67,938$ DCT coefficients, distributed across speakers, sessions and phonemes in the same way as the original datapoints.

A first step was then to average the DCT coefficients of each phoneme over each speaker and each session, yielding a total of 190 values for each DCT coefficient. A total of 192 (12 *speakers* $\times$ 4 *sessions* $\times$ 4 *phonemes*), but there were no occurrences of /ʊ/ in session 1 of speaker DID0128 and session 4 of speaker DID0168. The left column of figure 3.18 plots the mean values of k_0 (first row), k_1 (second row) and k_2 (third row) for F_1 (x-axis) and F_2 (y-axis). Each dot corresponds to the values of a given phoneme (/ɪ/ in black, /iː/ in grey, /ʊ/ in yellow and /uː/ in blue) for a given session (squares for session 1, circles for session

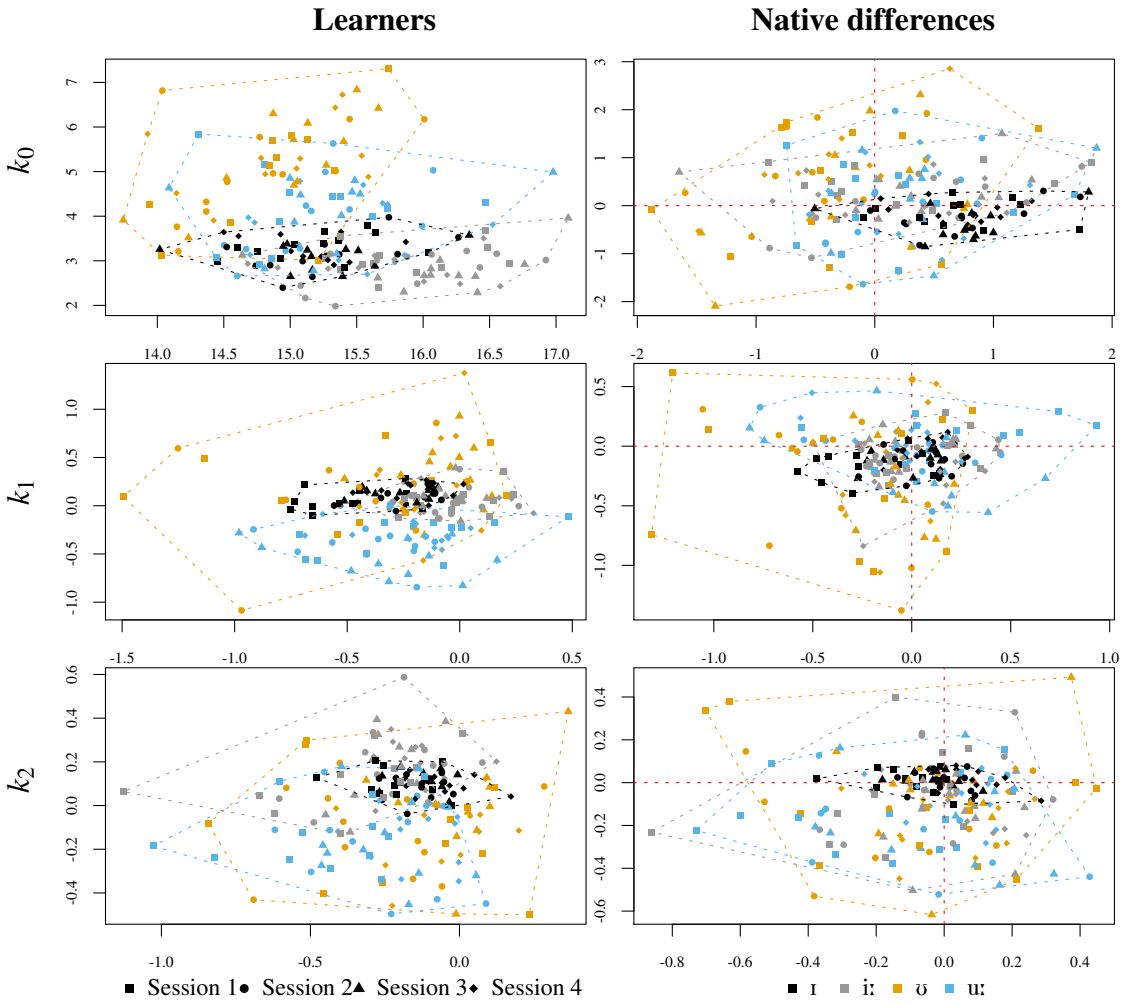


Fig. 3.18: Per-speaker, per-session means of the DCT coefficients for F_1 (x-axis) and F_2 (y-axis) in each phoneme. *Top row:* k_0; *middle row:* k_1; *bottom row:* k_2. *Left column:* values for learners; *right column:* differences between learners' and native speakers' means. *Dotted lines:* convex hulls linking extreme values.*Red dotted lines:* 0-difference point (origin) with native values.

2, triangles for session 3 and diamonds for session 4). Extreme values for each phoneme and each DCT coefficient are linked into a polygon to make these extreme values more visible. A look at the area of these convex hulls reveals discrepancies between phonemes. For k_0, which corresponds to the average value of a signal throughout the duration of a vowel, the area are 2.00, 3.00, 7.13 and 6.24 Bark^2 for /ɪ/, /iː/, /ʊ/ and /uː/ respectively. These numbers indicate a much greater dispersion across speakers for /ʊ/ and /uː/ than for /ɪ/ and /iː/ regarding average signal values. The same observations can be made on the distribution of

points for k_1 values: the areas of the convex hulls for /ɪ/, /iː/, /ʊ/ and /uː/ are 0.21, 0.30, 2.52 and 0.75 Bark^2 respectively. Compared to k_0, however, the gap between /ɪ/ and /iː/ on the one hand, and /ʊ/ on the other, is less wide in the case of k_1. The situation is different with k_2, where the size of the convex hull for /ɪ/ stands alone against those of /iː/, /ʊ/ and /uː/. Bearing in mind that k_0, k_1 and k_2 refer to the mean, slope and curvature of the signal respectively, it is striking to see how consistently smaller the convex hulls of all three DCT coefficients are for /ɪ/ – in spite of its much greater number of occurrences. Means and slopes are also much more consistent across speakers for /iː/ than for /ʊ/ and /uː/. However, looking at the areas of convex hulls is only a valid methodology if the extreme points of the polygons are *not* isolated outliers. Graphic observation indicates that this is probably not the case for the /ʊ/ k_1 and /iː/ k_2 convex hulls, meaning that they could be considered smaller than they are. Could the observed higher dispersions be the consequence of specific distributions of consonantal environments? DCT coefficients encoding the entire signal of each vowel for each BDM-normalized formant, and the signal beign subject to the consonantal environment, a direct relationship between dispersion and at least the variety of consonantal environments can be expected: the higher the number of different consonantal environments, the greater the dispersion – the bigger the convex hulls. It looks as if this is however not the case: the numbers of different consonantal environments for /ɪ/, /iː/, /ʊ/ and /uː/ respectively are the following: 72, 39, 9 and 28. These numbers should be measured against the counts of each vowel. A coefficient assessing the relative dispersion for each vowel and each DCT coefficient could be the following: $c_k^i = \frac{n_{syll}^i \times \mathscr{A}^{ik}}{n_{phon}^i}$, with n_{syll}^i the number of different syllabic environments in which the vowel appears, n_{phon}^i the number of occurrences of phoneme i, and $\mathscr{A}_k^i$ the area of the convex hull of the sets of per-speaker, per-session means for a given DCT coefficient k. The smaller coefficient c_k^i would be, the less dispersed the values are. Table 3.7 shows the 12 dispersion coefficients[10]. An observation common to all three DCT coefficients is that the dispersion coefficients c_k^I are smaller than

[10] These dispersion coefficient were all multiplied by 1,000 to make them more readable.

c	/ɪ/	/iː/	/ʊ/	/uː/
k_0	18.14	51.77	111.08	54.12
k_1	1.91	5.19	39.31	6.49
k_2	0.85	7.67	13.02	4.03

Table 3.7: Dispersion coefficients for each vowel and each DCT coefficient

their counterparts of other vowels. Likewise, c_k^U is much higher, while $c_k^{i:}$ and $c_k^{u:}$ reach similar values. Arguably the most crucial DCT coefficients when it comes to accurate perception of the phonemes are the first two – k_1, *i.e.* the slope of the signals, potentially being a major component of a VISC-based theoretical framework (*c.f.* section 3.2). Assuming, as is contended here, that the dispersion coefficients disclose information on the stability of the realizations of /ɪ/, /iː/, /ʊ/ and /uː/ across speakers and sessions, it looks as if three stages can be distinguished: one of a more stable and advanced level of acquisition, indicative of the situation for /ɪ/; /iː/ and /uː/ might reveal a second, intermediary level, while /ʊ/ seems much less stable across speakers and sessions, thereby revealing a possible, lower and more fragile state of acquisition.

These results, however, should be compared to native data. This is the object of the right column of figure 3.18, which shows the per-speaker, per-session differences between their average DCT coefficients for each BDM-normalized formants and the average native values. The differences took the speakers' sex into account, *i.e.* the learners' values were subtracted from native values of the corresponding gender. The red vertical and horizontal lines in each panel indicate a null difference: the closer to them the dots are, the more native-like the DCT coefficients are. Cursory observation would seem to bear out the findings established in the previous paragraph: the /ɪ/ values look more concentrated than those of the three other phonemes. They are particularly centered around the 0-difference point for k_1 and k_2. Once again, /ʊ/ values are widely scattered in all directions. The validity of these impressionistic remarks needs to be tested, however. The method used in order to do that is the following: first, rather than dealing with means, the dataset with all 11,323 phonemes

served as the basis for this part of the study. For each phoneme, the distance from the native F_1 and F_2 values for k_0, k_1 and k_2 was listed. As was the case above, these distances were measured from the average values of native speakers of the same sex as the learners. This procedure returned F_1 and F_2 coordinates for all phonemes, with specific coordinates for each DCT coefficient (along the same principle as the right column of figure 3.18). The points the furthest away from the origin were then stored. This returned three coordinates, one for each DCT coefficient. For k_0, the most distant phoneme from native values was an /iː/, pronounced by speaker DID0035 in session 1; it is 10.66 Bark away from the origin. For k_1, the longest distance is 5.92 Bark, with an /iː/ pronounced by speaker DID0014 in session 3. For k_2, an /ɪ/ by speaker 0096 in session 3 is 4.54 Bark away from the origin. The idea was then to create circles around the origin, with varying radii, and to investigate the points these circles included. The lengths of the radii are incremented proportions of the longest distance, which gives both an upper limit to the lengths of the possible radii, and an estimate of how far the outliers may be. The radial lengths were allowed to slide from 5% to 25% of the pre-stored maximal distance, in increments of 0.1%. All the points within these circles around the origin were then selected in turn. From these subsets of datapoints, the intra-phoneme proportions of DCT coefficients were calculated: the intra-phoneme proportions refer to the per-phoneme percentages which the subsets of within-range datapoints represent in comparison to the entire set of the phoneme's datapoints (including, then, those outside the circles). Likewise, among all the datapoints within the circle, the inter-phoneme proportions, *i.e.* the distribution of phonemes in those datapoints within the circle, were also calculated. The results of such a procedure are presented in figure 3.19. The right column plots the inter-phoneme proportions against the varying lengths of the radii, expressed as percentages of the maximal distance from the native values. For a given radius, *i.e.* a given x, the proportions of each phoneme in the subset of datapoints within the circle are calculated. Figure E.6 in section E.7 provides a graphic explanation of the

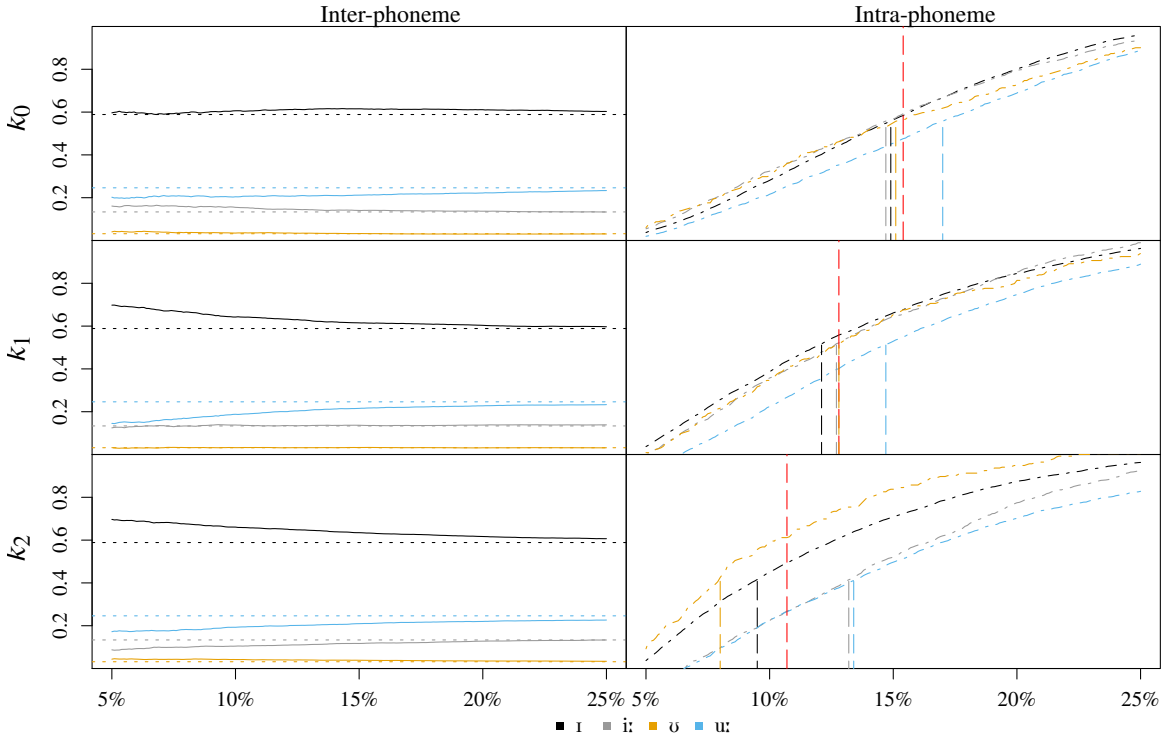


Fig. 3.19: Inter- (*left column*) and intra- (*right column*) phoneme proportions for k_0 (*top row)*, k_1 (*middle row*) and k_2 (*bottom row*). The *x*-axis increments the percentage of the maximal distance from the sex-dependent native reference points within which, or from which, the proportions are calculated. The proportions are given on the *y*-axis.

process at hand. The flat dotted lines provide the base proportions, *i.e.* the percentage of each phoneme in the entire dataset: /ɪ/ thus accounts for 58.9% of all monophthongs (6,666 occurrences); /iː/, 13.3% with 1,511 occurrences; /ʊ/, 3.2% with 359 items; and /uː/ 24.6% with 2,787 occurrences. Whether the inter-phoneme proportions are higher or lower than the base proportions may give an indication of the state of acquisition of the phonemes: if higher, then it means that more phonemes than expected in a given category are present in the circle, *i.e.* the category is over-represented in the subset of datapoints around the origin – their differences from native values are smaller, and therefore their acquisition may be deemed better than the other categories. The inter-phoneme proportions of /ɪ/ and /ʊ/ are consistently above the baseline for the three k_0, k_1 and k_2 coefficients. The /ʊ/ curve for k_0 dips slightly under the baseline as the radius of the circles increases, a counter-intuitive trend possibly caused by the smaller number of occurrences. The inter-phoneme proportions

of /iː/ for k_0 and k_1 are also above the baseline, although to a much lesser extent for the latter coefficient. The curve is below the baseline, however, for k_2. For all three coefficients, the inter-phoneme proportions of /uː/ are under the baseline, meaning that the category is under-represented in areas around the origin. The distance of its values for k_0, k_1 and k_2 to native values are overall greater than expected, and may therefore indicate a lower level of acquisition. The right column plots the intra-phoneme proportions for each DCT coefficient against the varying length of the radius of the circle subsetting the datapoints. In these cases, the subset of points of a given phonemic category is compared to the size of all the points of that category, including those outside the circle. The dotted vertical segments with matching colors indicate at which radius length the cut-off points of 50% of the datapoints of the category have been reached. The red segments mark the same cut-off point of 50%, but for the entire dataset, regardless of the phonemic categories. This point will be refered to as the Global Cut-Off Point (henceforth, GCOP). For the differences from native values for the means of the signals, *i.e.* k_0, all inter-phoneme proportions reach the cut-off points before the GCOP except for /uː/. The same pattern takes place for the differences in the slopes of the signals, *i.e.* for k_1: /uː/ is singled out because it takes a circle with a longer radius to capture half of its datapoints. For k_2, both /iː/ and /uː/ have cut-off points with longer radii than the GCOP. Combined with the findings on inter-phoneme proportions, it looks as if this experimental design evidences a less robust state of acquisition for /uː/: the differences with native values for the three coefficients reveal more dispersed values than could be expected from the number of occurrences of this vowel: the differences with native values for the three coefficients reveal more dispersed values than could be expected from the number of occurrences of this vowel.

Is there, then a longitudinal effect, and do the speakers display the same patterns altogether? One way to answer those questions is by looking at the per-session, per-phoneme evolutions across the four sessions of each learner's average DCT coefficients of /ɪ/, /iː/, /ʊ/

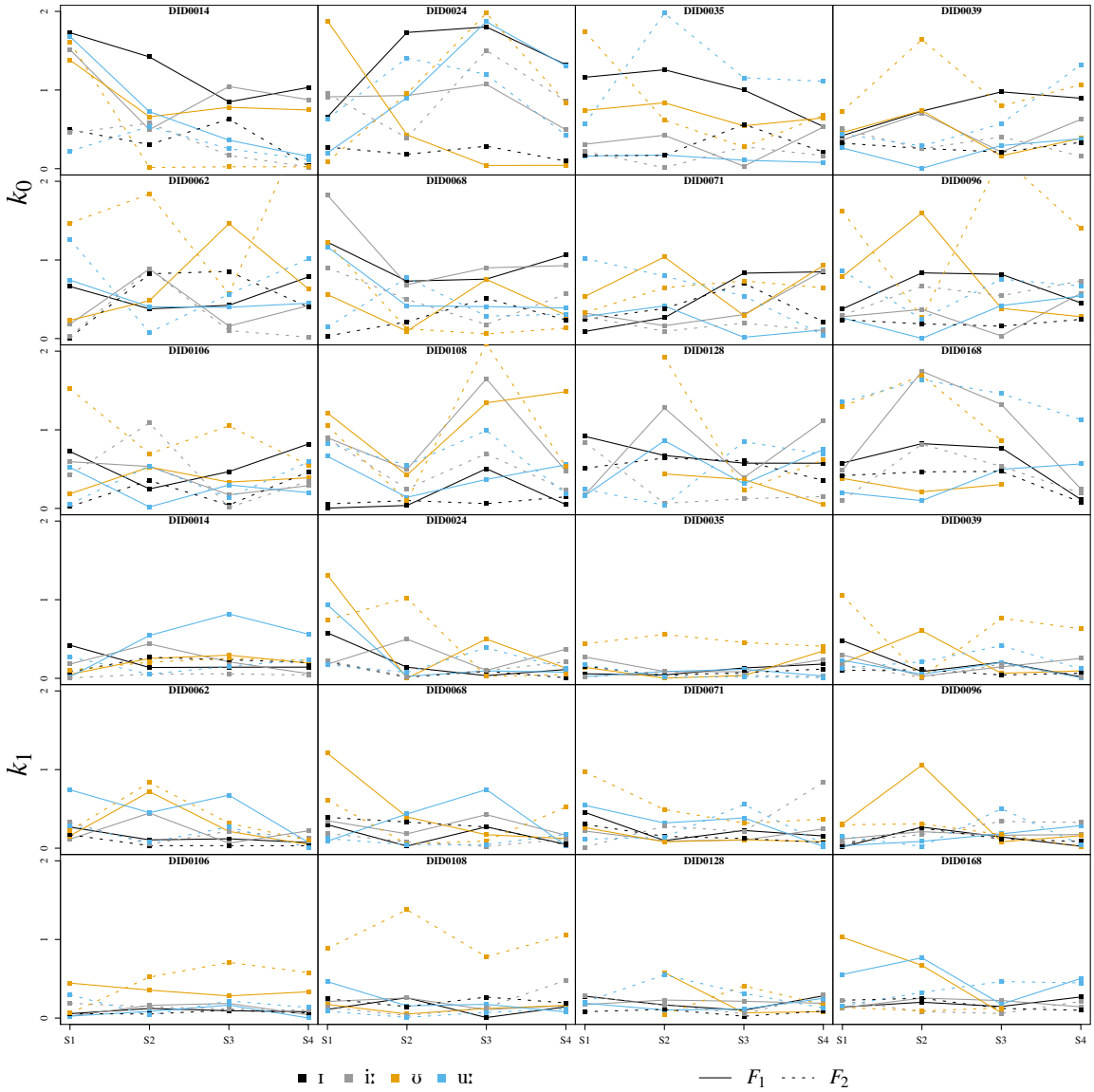


Fig. 3.20: Per-session, per-speaker evolution of the absolute values of the differences from native values for k_0 and k_1. *First three rows:* k_0; *last three rows:* k_1.

and /uː/. This is the object of figure 3.20, which plots the absolute means of the differences from native values for the first two DCT coefficients k_0 and k_1. The reason why only the first two coefficients were represented is twofold: *(i)* first, a 36-panel plot is harder to make sense of and to read than a mere 24-panel one – that number already being high enough to give an impression of clutter; *(ii)* k_2 is arguably of lesser importance when it comes to phonemic analysis: to the best of our knowledge, the curvature of the signal matters less than its mean and slope for correct identification. The same dataset of 11,323 datapoints has

been used. The interested reader can have a look at the same figure for k_2 in figure E.7[11]. F_1 is plotted in full lines whereas F_2 is plotted in dotted lines. The first three rows plot the differences from native values for k_0, the last three rows for k_1. The scales on the y-axis for the two coefficients were deliberately kept identical in order to make cross-comparisons possible. Clearly the amplitude of the curves are greater for k_0 means than for k_1, regardless of the phonemic categories. Regarding k_0, no cross-speaker, or phoneme-specific patterns seem to emerge: the evolutions of the differences from native values are most likely purely idiosyncratic. However, a longitudinal effect may well argue to take place in a few speakers, namely speakers DID0014, DID0035, DID0068, and possibly DID0168. The situation is somewhat different with k_1, where patterns may be argued to exist. The F_1 and (to a greater extent) F_2 curves for /ʊ/ (*i.e.* the yellow ones) are distinguishably higher than their counterparts in other phonemic categories. This seems to be the case for either formant in seven out of twelve cases (DID0024, DID0035, DID0039, DID0062, DID0068, DID0106, DID0108). In no instances are the curves for /ɪ/ or /iː/ among the higher ones. It therefore looks as if their signal slopes are closer to native ones – possibly the sign of a more advanced state of acquisition.

3.5.3 DCTs *vs.* mid-temporal values

At this stage, one final question that needs to answered is that of the added value of DCTs. After all, they are computationally more intensive than the widespread mid-temporal formant values, so is there a reason to take the extra coding steps to extract the formant tracks and use DCTs to analyze them? And, perhaps more crucially, to what extent do results obtained from them change the analysis of phonemic acquisition, compared to results obtained from mid-temporal values?

[11] Unlike in figure 3.20, in figure E.7 the y-axis ranges from 0 to 1 (not 0 to 2).

One exploratory method before going into more specific speaker-dependent details is to compare phonemic classification using the two forms of data, *i.e.* DCTs *vs.* mid-temporal formant values. These two forms will again be compared after BDM-normalization, using the same dataset of 11,323 /ɪ/, /iː/, /ʊ/ and /uː/ phonemes embedded in syllables also present in the NSS. In order to test the classification accuracy, a Quadratatic Discriminant Analysis (henceforth, QDA) was used. QDAs work like logistic regressions in that they test the effect of continuous dependent predictors (in this case, mid-temporal formant values and DCT coefficients) on categorical response variables. However, logistic regressions are confined to two classes for the dependent measures, a limitation which QDAs do not have. Another option was Linear Discriminant Analysis, but this method has constraints on the equality of covariances which the data at hand does not necessarily comply with. Besides, considering the complexity of the data, it seems best to envisage non-linear classification rather than linear classification. The exploration of the differences between the two forms of data starts with sex-dependent, but speaker-independent, QDA. In order to assess the respective benefits of each element in the analysis of the data, six two-dimensional models have been studied. The second dimension of each model is the model itself with the duration of the phoneme, *i.e.* column `PHONDUR` in the datasheet, included in the set of continuous predictors. All models include the BDM-normalized F_1 and F_2 dimensions. They are summarized in table 3.8. These

Table 3.8: Models subjected to the QDA

Model	Predictors
m_1	Mid-temporal $F_1 + F_2$
m_2	Mid-temporal $F_1 + F_2$ + Duration
m_3	$F_1\ k_0 + F_2\ k_0$
m_4	$F_1\ k_0 + F_2\ k_0$ + Duration
m_5	$F_1\ k_1 + F_2\ k_1$
m_6	$F_1\ k_1 + F_2\ k_1$ + Duration
m_7	$F_1\ k_2 + F_2\ k_2$
m_8	$F_1\ k_2 + F_2\ k_2$ + Duration
m_9	$F_1\ k_0 + F_1\ k_1 + F_2\ k_0 + F_2\ k_1$
m_{10}	$F_1\ k_0 + F_1\ k_1 + F_2\ k_0 + F_2\ k_1$ + Duration
m_{11}	$F_1\ k_0 + F_1\ k_1 + F_1\ k_2 + F_2\ k_0 + F_2\ k_1 + F_2\ k_2$
m_{12}	$F_1\ k_0 + F_1\ k_1 + F_1\ k_2 + F_2\ k_0 + F_2\ k_1 + F_2\ k_2$ + Duration

models are of increasing complexity, and progressively integrate the different combinations

of DCT coefficients. With a maximum number of 7 predictors, the rule of thumb that the number of parameters p should be inferior to the number n of datapoints divided by 5 (*i.e.* $n >= 5 \times p$) is respected. The idea underlying such comparisons was to endeavour to identify, as precisely as possible, the contributions of each parameter – *i.e.* mid-temporal F_1 and F_2, all three DCT coefficients taken separately and together, and phoneme duration – to the identification and classification of the four categories of phonemes under study. Unlike the procedure commonly used, and in the wake of what was done in section 3.3, the training set was made up of the occurrences of /ɪ/, /iː/, /ʊ/ and /uː/ in the NSS. The test set consisted of the learners' phonemes. All the predictors were scaled by z-score standardization within the subset of datapoints for speakers of the same sex. A QDA was run for each model on each sex separately, *i.e.* both the training set and test set were split into two to account for sex differences. The *R* package used was the `MASS` package by Venables & Ripley (2002). The distribution of the occurrences for each set is the following: 703 datapoints for female speakers, 418 for male speakers in the NSS; in the main corpus, 8,508 for female speakers and 2,815 for male speakers. One key issue to solve with such a procedure, *i.e.* using the NSS as a training set, is that of the prior probabilities to input. The default values of prior probabilities for a QDA using `MASS` are the respective proportions of each category in the whole dataset. The priors for each sex in the two datasets are listed in table 3.9. Although

Table 3.9: Prior probabilities of the sex-specific datasets

Dataset	ɪ	iː	ʊ	uː
Female natives	0.66	0.16	0.02	0.16
Male natives	0.63	0.18	0.04	0.16
Female natives	0.60	0.13	0.03	0.24
Male natives	0.56	0.15	0.03	0.25

these proportions are not too dissimilar, the proportions of /uː/ in the learner corpus are 50% higher than in the NSS. What effect then do the priors have on the proportions of accurate predictions after running QDAs? Figure 3.21 plots these proportions against the twelve models presented in table 3.8. The relevant datapoints, *i.e.* the DCT coefficients,

the mid-temporal formant values and the durations were all normalized beforehand within each per-sex subset. The only differences between the different passes consisted in changing the prior probabilities used for the train sets: in turn, the priors for the train and test sets were used. The differences displayed in figure 3.21 are minor: in all instances, the shapes of the curves are parallel, and sit at around the 60% mark of accurate predictions. Prediction accuracy is lower for male speakers than for female speakers, a state of affairs most likely due to the lower number of datapoints in both train and test sets for male speakers than for female speakers. With such minimal differences, it was decided that the rest of the study aiming at comparing the twelve models would resort to the default values of the prior probabilities, *i.e.* those of the NSS, for the sake of simplicity. The next paragraph looks into more details at how the prediction accuracies change across phonemes and models.

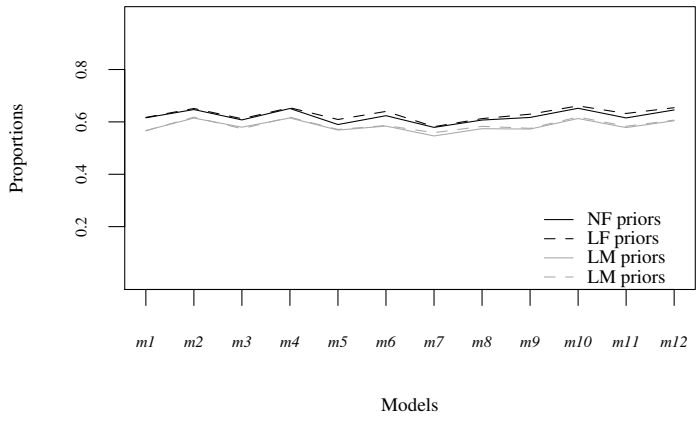


Fig. 3.21: Per-model proportions of accurate identification using different prior probabilities. *FN:* female natives; *MN:* male natives; *FL:* female learners; *ML:* male learners.

Let it be reminded that for the time being, session and speaker differences are not factored in. The results breaking down the proportions of accurate predictions by model, phoneme and sex are displayed in figure 3.22, with each panel corresponding to a QDA run on one of the models presented in table 3.8. The *y*-axis indicates the proportion of phonemes accurately predicted by the QDA. The proportions were calculated by dividing the diagonal of the

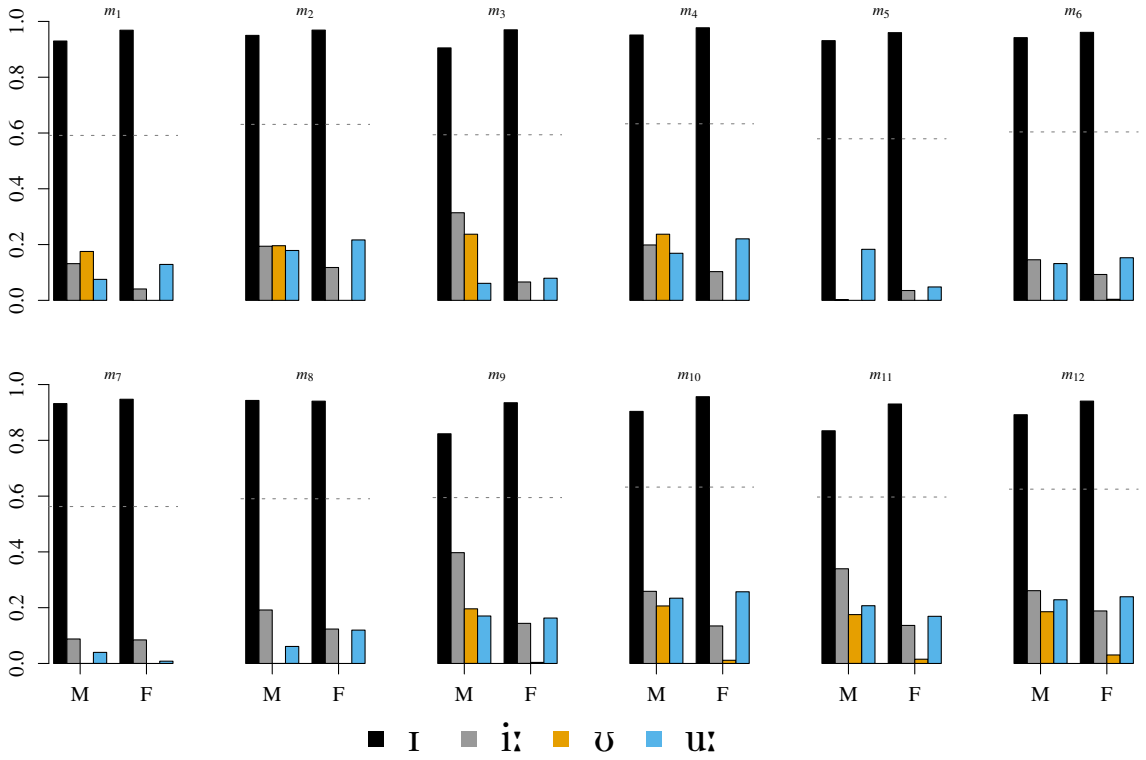


Fig. 3.22: Per-sex proportions of accurately classified phonemes after a QDA. *Dotted line:* global proportion of accurate prediction regardless of sex or phonemic category.

confusion matrices for each sex after running the QDA by the total number of occurrences of each phoneme for each sex as well. The dotted lines in each panel show the global proportion of accurate prediction regardless of sex and phonemic differences. The first striking feature of these results is that duration increases the proportion of correct identification for all models (recall that even-numbered models add `PHONDUR` in their set of predictors). The differences in rates of accuracy, sex and phonemes aside, are tiny: the global proportions range from 56.27% for m_7 (*i.e.* only the curvatures of the F_1 and F_2 signals are factored in) to 63.30% for m_4 (*i.e.* the emulated means of the F_1 and F_2 signals with duration included)[12]. The m_2 model, based on the phonemes' durations and their mid-temporal F_1 and F_2 values, comes a close third with a global proportion of accuracy at 63.09%. In-between m_2 and m_4, at 63.24%, comes m_{10}, based on the first two DCT coefficients and duration. If Ockham's Razor is a principle to be adopted, then clearly a model based on the study of duration and

[12] The table of results averaged over the male and female speakers can be found in table E.3 in section E.7.3.

mid-temporal F_1 and F_2 formant values is the most efficient. The second best model in terms of simplicity is m_4. Of all three coefficients, k_0 seems to be the most crucial one when it comes to the identification of phonemes – admittedly hardly a surprising result since k_0 emulates the mean of a formant's signal. However, looking at the proportions of accurate prediction for each phoneme reveals that the two seemingly most efficient models present a major flaw, at least from the perspective defended here. Quite surprisingly, neither m_2 nor m_4 manages to predict *any* of the 262 occurrences of /ʊ/ by female speakers. This is all the more surprising as they are even fewer instances of /ʊ/ in the male learner corpus – 97 –, yet the QDA correctly identifies 7. Interestingly enough, the first model where occurrences of /ʊ/ pronounced by female speakers are accurately predicted is m_6, based on k_2 only. Further looking at the specifics of the per-phoneme predictions, it turns out that the best model for predicting all phonemic categories is the last one, m_{12}, which factors in duration and all DCT coefficients, and is therefore also the most complex one. Even though it only ranks 4$^{\text{th}}$, m_{12} features the highest proportions of prediction accuracy for phonemes with low numbers of occurrences – especially /ʊ/, with 3.05% among female speakers, and 18.56% among male speakers. The former proportion, more than the latter, is of special interest for the purpose of this discussion: clearly from the data displayed in figure 3.22, occurrences of /ʊ/ pronounced by female learners are highly unlikely to be accurately predicted by an NSS-based QDA. For those particular cases, m_{12} outshines its competitors: the second and third best models, m_{11} and m_{10}, accurately predict /ʊ/ pronounced by female speakers in 1.53% and 1.15% of all cases respectively. Considering that m_{12} is the most complex of all models, and apparently the most able to predict all categories, how does it fare compared to the most streamlined and commonplace one, *i.e.* m_2?

At this stage, the issue here becomes a theoretical one, rather than a practical one. After all, conclusions about phonemic acquisition can be drastically different if applying a classification procedure like a QDA returns drastically different results whether the datapoints

were preprocessed in one manner rather than in another: with no /ʊ/ pronounced by female speakers accurately predicted with the first two most efficient models (m_2 and m_4), it can easily be infered that the acquisition of that particular phoneme is much less advanced than the other three. Clearly here once again the major stumbling block is the difference in counts and the skewed distribution of phonemes pertaining to the very nature of corpora based on spontaneous speech: the fewer occurrences of a phoneme, the less likely it will be correctly identified and predicted. In the view defended here, the case is very strong to adopt whatever model is biased in favour of minority occurrences, at the expense of global accuracy. The trade-off, it is argued, is minor from a purely practical point of view[13]: the gap in the global proportion of accurate prediction is only 0.59% between m_2 and m_{12}. It is, on the other hand, quite major from a theoretical one: m_{12} returns the highest proportion for a phonemes most other models are not able to predict. The price to pay, the second lowest proportion of accuracy for an already overrepresented phoneme, /ɪ/, 0.59% in global predictions, and admittedly considerably added complexity with the need to extract and process all the formant tracks, still seems acceptable – it is in any case one that is strongly recommended to be paid here. Having concluded from this exploratory experimental set-up that four models, *i.e.* m_2, m_4, m_{10} and m_{12}, return competitive, yet contradictory, results, it is now time to investigate how they fare when it comes to longitudinal, speaker-dependent analyses. Because the biggest difference between the first two models, *i.e.* the influence of consonantal environments which m_4, being based on the emulated mean k_0 of the signal, is comparatively more likely to be affected with than m_2, has been neutralized by the design – only phonemes embedded in syllables existing in the NSS; because m_{10}, in spite of its overall better performance at predicting the different categories, shows a rate of prediction for occurrences of /ʊ/ pronounced by female speakers which is 50% lower than m_{12}; and in order not to clutter the graphs with marginally useful comparisons, only m_2 and m_{12} will be investigated in the next paragraph.

[13] Coding hurdles aside, of course.

The comparisons between the per-speaker, per-session proportions of accurate predictions by QDAs based on m_2 and m_{12} are displayed in figure 3.23. The proportions obtained from the mid-temporal formant values and the phonemes' durations are plotted in full lines,

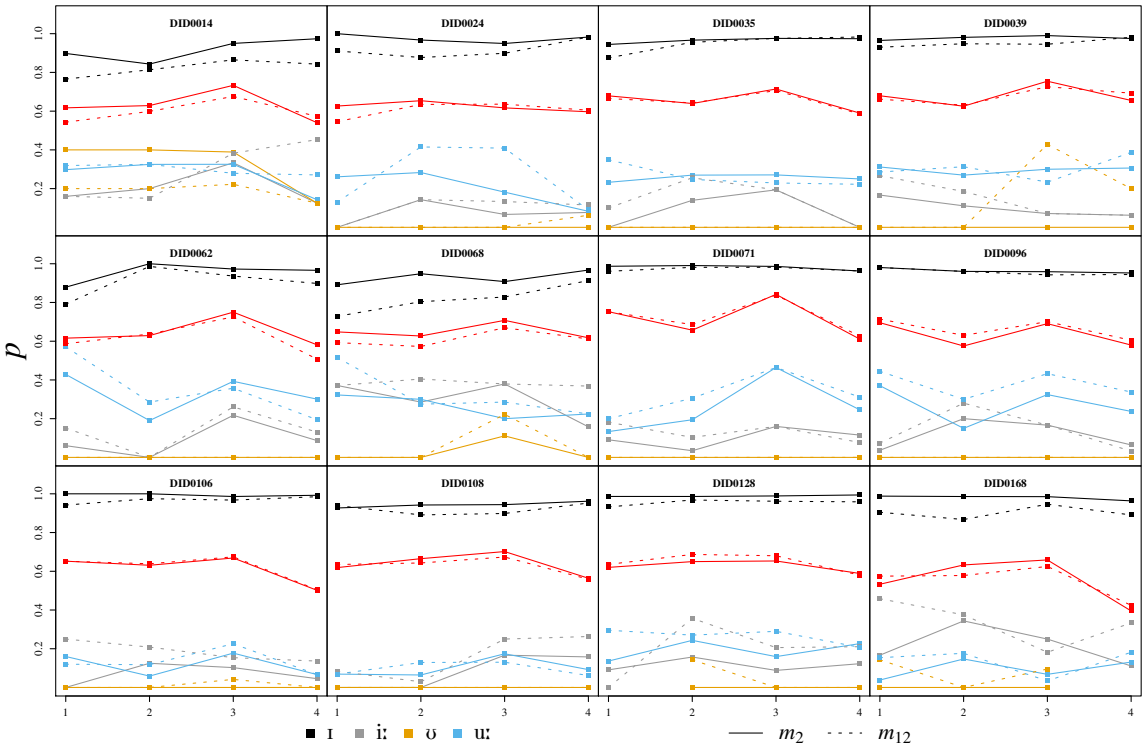


Fig. 3.23: Per-speaker, per-session proportions of accurate predictions by QDA. *Full lines:* QDA based on m_2; *dotted lines:* QDA based on m_{12}; *red lines:* global proportions (regardless of phonemic categories).

whereas those obtained from the durations and the three DCT coefficients k_0, k_1 and k_2 are plotted with dotted lines. As usual, the x-axis and the y-axis indicate the session numbers and the proportion between 0 and 1 respectively. Each panel gives the results obtained by running the QDA on one given learner's datapoints. The red lines mark the overall proportion of correct classification across the four phonemes. Just as in the previous paragraphs, the data was scaled across speakers in each dataset (*i.e.* the natives' or the learners'), within each sex. The prior probabilities were left at their default values in the `MASS` package, meanong they correspond to the proportions of occurrences of each category in the sex-dependent NSS subsets. The features common to all speakers are the clearly higher prediction accuracy

for /ɪ/ regardless of the models, and the globally similar global proportions, ranging from 0.40 for speaker DID0168 in session 4 to 0.84 for speaker DID0071 in session 3 in the case of m_2; and from 0.42 to 0.84 for the same speakers and sessions in m_{12}. The overall similarities between the results obtained from the two models must not conceal the following discrepancies (the proportions for m_2 are given first, then those for m_{12}): /iː/ in speaker

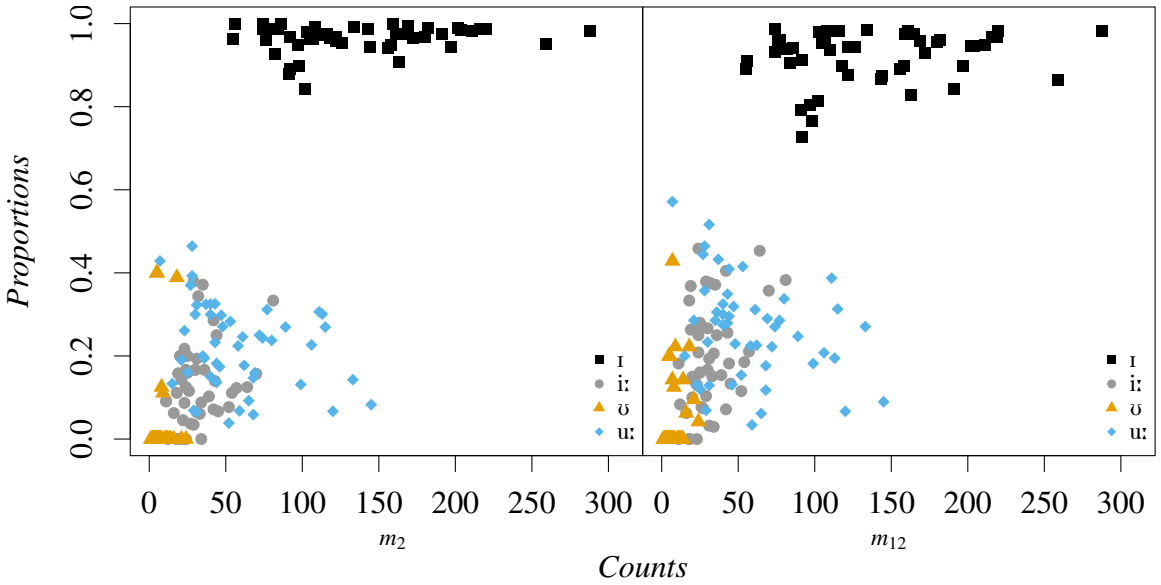


Fig. 3.24: QDA proportions of accurate predictions against the numbers of occurrences of each phoneme in all sessions. *Left:* m_2; *right:* m_{12}.

DID0014's session 4, 0.13 *vs.* 0.45; /uː/ for speaker DID0024's session 3, 0.18 *vs.* 0.41; /ʊ/ for speaker DID0039's session 3, 0 *vs.* 0.43. More generally, the highest absolute difference in proportions for /ɪ/ is 0.16 (speaker DID0068 in session 1), a gap with arguably little consequence on a diagnosis on phonemic acquisition. The differences is higher than 20% in 5 instances (out of 48) in the case of /iː/, and in 4 instances for /ʊ/, with a high mark of 43%, mentioned above; and only two instances for /uː/. All in all, these discrepancies are somewhat circumscribed in frequency of occurrences, and limited in scope. How does the respective counts of each phoneme in each session affect the classification rate? Figure 3.24 plots the per-speaker, per-phoneme proportions of each phoneme against their numbers of occurrences for the two models. The respective Pearson correlation coefficients for /ɪ/, /iː/, /ʊ/ and /uː/ in the m_2 models are $r_I = 0.18$, $r_{i:} = 0.31$, $r_U = 0.08$ and $r_{u:} = -0.26$;

in m_{12}, $r_I = 0.21$, $r_{i:} = 0.39$, $r_U = 0.18$ and $r_{u:} = -0.29$. The plot and these coefficients seem to indicate that the effect of the numbers of occurrences on prediction accuracy is somewhat contained, and that phonemic categories do exert influence on this accuracy: the cases in point are the instances when there are as many occurrences of /ɪ/ and /iː/, above the 50-occurrence mark. The consistency of the predictive proportions even when those two categories share the same numbers of occurrences may also confirm *a posteriori* the validity of this experimental set-up based on QDAs. It goes without saying, however, that other parameters, such as syllable structure and lexical variety, are likely to exert influence on the specificities of each phonemic category – a venue of research to be explored in the future.

3.5.4 Conclusion

This section has attempted two things: *(i)* first, to use DCTs in order to sketch the learners' various states of acquisition, and their evolutions for /ɪ/, /iː/, /ʊ/ and /uː/. *(ii)* to assess the added value of the more computationally intensive DCTs compared to more traditional approaches based on mid-temporal formant values. The method used to achieve those two goals was to carry out QDAs in order to establish how accurately the phonemes were predicted from native values. For the first purpose, the conclusion is not that different from the findings of the previous sections, *i.e.* there is a very strong likelihood that a hierachy exists between the levels of acquisition of the four phonemes under study. This hierarchy may well be the following, in decreasing order: /ɪ/; /iː/; /uː/; and lastly, /ʊ/. One major *caveat* to these results is that this hierarchy follows that of the number of occurrences of each phoneme. The exact extent of this influence will have to be determined in future research, as it is a pre-condition of any study of spontaneous speech that the corpus will end up featuring unequal numbers of datapoints in each category. It will be once again emphasized here that the respective proportions of each monophthong in the main learner corpus emulate those obtained in the NSS. It is also once again contended here that frequency of occurrence

is bound, *somehow*, to affect acquisition – a state of affairs that was tentatively observed here. Regarding the added value of DCTs, it should be noted that mid-temporal formant measurements return results that are predominantly similar to those obtained by using DCTs, and possibly even overall more efficient. In terms of ratio of computational complexity to efficiency, let it be clear that mid-temporal measurements well, win hands down. If the findings of this section are anything to go by, DCTs should however not be discarded, since they seem to be able to allow for more refined classifications and distributions. Further research is needed to bear out the following statement, but it may well be the case that with their modelling of consonantal environments and the better classification rates they enable for underrepresented categories, the first three DCT coefficients are perfectly suited for corpora based on spontaneous speech.

3.6 Conclusion

This chapter had set out to explore the individual evolutions of acquisition, in the hope of finding cross-speaker patterns, along with differences in the learning slopes of the four monophthongs focused on, *i.e.* /ɪ/, /iː/, /ʊ/ and /uː/. Various methods of data processing, from mid-temporal formant measurements to VISCs and DCTs have been used, along with different classification methods such as KNNs and QDAs, or modelization frameworks, such as LMERs. In perhaps all instances, differences in the acquisition patterns of /ɪ/, /iː/, /ʊ/ and /uː/ have been observed: different dispersions for each phoneme in the case of VISCs in section 3.2, gaps in the predictive consistencies of best alternates in the case of KNNs in section 3.3, varying slopes in the LMER models in section 3.4, discrepancies from native values specific to each vowel in the case of DCTs in section 3.5…Let it be clear however that no definitive statement (should such a thing as a "definitive statement" be even possible) regarding the acquisition of /ɪ/, /iː/, /ʊ/ and /uː/ is here formulated. The complexity of the data is a surface that has hardly been scratched: many parameters such

as word frequency, syllabic structure, stress, MoAs and PoAs of preceding and succeeding phonemes, all of which are available in the corpora, have not been factored in. To take but an example, explaining away the substantial residuals observed in section 3.4 will probably require (in future research) the addition of the WORD parameter – thereby demanding much more processing power than what the models investigated here necessitated, because of the unbound nature of the number of categories for this parameter. Another point is that infirming the null hypothesis (*c.f.* the beginning of this chapter) by resorting to such chaotic data as those extracted from spontaneous speech is, well, too easy. If anything, it is surprising that this null hypothesis *held so well*. For in all cases where it was tentatively asserted that differences in the acquisition of the four phonemes might exist, lingering doubts remained: that the frequency of occurrences might make cross-phoneme comparisons unreliable, for instance; or that varying arrays of consonantal environments and of lexical frequencies might challenge the legitimacy of compounding such heterogeneous data in unifying calculations. Let this chapter end, then, not only with the hope that the computations undertaken therein turn out to be worthy and insightful, but also by a tip of the hat to the simple and potent elegance of the null hypothesis and of analyses based on mid-temporal formant measurements.

Conclusion

This study was resolutely quantitative. An unreasonable amount of research time was dedicated to designing procedures to extract relevant data from the 81 recordings of the LONGDALE project. Chapter 1 endeavoured in its first sections to describe the raw recordings and the participants, along with the processes that led to the extraction of the data. Multitier `PRAAT TextGrids` were generated for each participants. They feature tiers aligned by two aligners, SPPAS and P2FA, containing intervals with boundaries for words, English and French syllables, and individual phonemes. Tiers containing the individual transcriptions and syllables of each word as listed in the LPD have also been added, with boundaries defined by each of te two aligners. The same `TextGrids` were also generated on three separate subcorpora of lesser sizes: two sets of recordings, one of a list of English words, the other of a text in French, made as part of the LONGDALE project, were also processed. A homemade corpus of native spontaneous speech was also created. In total, 120 `TextGrids` have been generated. Their phonemic tiers were then parsed by a script that extracted information for each vowel. An ambition of exhaustivity existed, as with each vowel, on top of its label, came 541 datapoints (86 for the supbcorpus in French) – a total to multiply by two, because the process took place one time for each aligner (except for the French subcorpus which was aligned using SPPAS only). The data extracted dealt with extra-linguistic information such as the speaker, the session or the number of days spent abroad, linguistic information, such as the word the vowel appeared in, the syllable, the syllable structure, the various transcriptions (from either the aligner's own dictionary or from the LPD), whether the

syllable is stressed or not, the preceding and succeding phonemes, their places and manners of articulation, and acoustic information, such as the first four formant values at each centile of the vowel's duration, its duration or its intensity... In total, by adding the vowels collected based on the intervals created by the two aligners (*i.e.* these vowels are the same but with slightly varying boundaries between the two alignments), a total number of 199,950 vowels were extracted. Added together, and regardless of the aligners and subcorpora, the grand total of cells available from all the datasheets amounts to 107,052,945. Regardless of the quality of the collected vowels, what possible errors remained, what features based on the collected information did the recordings have? These are the questions which the second part of chapter 1 tried to answer. Failed extractions for certain formant values, explained or unexplained variations in labelling by the aligners for frequent words, errors in syllabification caused by discrepancies between dictionaries or recondite decision processes in the aligners' algorithms were noted, along with detailed studies of the vowels' durations and the learners' speech rates.

All these preliminary analyses having been made regardless of vocalic categories, chapter 2 was an attempt to describe the specificities of each vowel regardless of the speakers' idiosyncrasies, in order to detect the possible existence of cross-speaker patterns of acquisition. The focus was on monophthongs exclusively. Before that, however, it was necessary to try to assess the quality of the automatic extraction carried out along the lines of the previous chapter. It was demonstrated that the formant values obtained through the procedure across all the centiles were mostly within reasonably realistic ranges, from which it is contended that the conclusion that the automatic extraction of those values was reliable can be drawn. Then the average distribution of the monophthongs in the vocalic trapezoid was measured, and compared to native values, and the lexical variety attached to each monophthong was investigated. The overwhelming proportions of a few function words for certain phonemes was noted, a characteristic which future research will have to take into account in a subtler

way than was done here: the approach remained mostly blind to lexical variations. A study of the dispersion of the values of F_1, F_2 and F_3 across all centile led to the discovery of disparities among vowels, with /ʊ/ and /uː/ displaying higher dispersions than their counterparts. A tentative investigation was then carried out in order to devise a procedure that could lead to retaining the most consistent, least dispersed, formant values for each vowel among all the values on each centile. These gaps in dispersion, combined with the differences in frequencies of occurrences between the monophthongs led to the question of how to process the acoustic data. It was asserted that the most common methods of normalization, such as the Lobanov method, were particularly suited for corpora where each phonemic category was represented by the same number of tokens. On the other hand, they were ill-suited for corpora based on spontaneous conversation, in which the skewness of the phonemic distributions is a defining feature. In that respect, the similarities between the distributions of the native corpus and those of the learner corpus suggested that the normalization methods to use for spontaneous speech should be vowel-intrinsic, but formant extrinsic. This suggestion was demonstrated by comparing the various methods of normalization to a corpus with even distributions across phonemes, the P&B dataset (Peterson & Barney (1952)). The comparative advantages of two vowel-intrinsic methods of normalization, Bark and Bark Difference Metric (BDM) were determined, and a recommendation in favour of BDM was made: the method makes it possible to integrate more information, *i.e.* the F_3 signal, in a simple way, by reducing a three-dimensional parameter (F_1, F_2, F_3) to two (Z_1 and Z_2). The chapter ends by investigating the relationship between contrast distances, *i.e.* the length of the /ɪ/-/iː/ vector in the BDM-normalized space on the one hand, and that of the /ʊ/-/uː/ on the other, and the surface of the convex hulls linking the outermost vowels of the entire inventory. This relationship, measured by the ratio of the vector length to the polygonal area, was calculated for all English corpora, along with the P&B data. The consistency of the results for /ɪ/-/iː/ may indicate a greater awareness of the coarticulatory targets to reach

for that contrast than for /ʊ/-/uː/. The extent to which this assumed greater awareness is evidence of a difference in acquisition has yet to be established in a firmer fashion.

Chapter 2 having unveiled a consistent cross-speaker pattern in the higher rates of dispersion of tokens of /ʊ/-/uː/, it was then time to take a look at each learner's evolution of the acquisition of the two contrasts. The first concern of chapter 3 was to compare the vowel inherent spectral changes of each learner in each session, along with those taken from the subcorpus of read lists of English words, to their native counterparts. Besides the added value, with respect to the previous chapter, of focusing on the per-session, per-speaker means of BDM-normalized F_1 and F_2 values, the focus on VISCs made it possible to take a look at values other than mid-temporal ones: the starting and ending points of the VISCs in this work were set at 20% and 80% of the vowels' durations respectively. The lengths of the resulting vectors for each speaker were compared to native values, with a more detailed look at the four phonemes under study in each session. Even by considering each speaker individually, the evidence is consistent that the onset-to-offset distances (regardless, however, of their locations in the vocalic space) are closer to native values for both /ɪ/-/iː/ than for /ʊ/-/uː/. Such results came to be interpreted as a strong argument against a similar acquisition of the two contrasts. The findings were then further corroborated by the analyses of the standard deviations of the OODs for each speaker in each session – as established in chapter 2, and in spite of lower numbers of tokens, /ʊ/-/uː/ feature higher standard deviations, even when comparing with native values and when factoring type-to-token ratios. Could such findings be confirmed by using classification algorithms? The idea was that if indeed some phonemic targets are better acquired by the learners than others, then it can be reasonably assumed that the phones emulating those categories would fare better when subjected to classification algorithm.

The method chosen was the *k*-nearest neighbours. A set-up, later on used for quadratic discriminant analyses, was designed, whereby instead of randomly sampling the data in

even folds, and selecting training and test sets in turn among these folds, the training set was the corpus of native speakers. The variables under study were the mid-temporal BDM-normalized F_1 and F_2 values of each phone, but in order to keep potential consonantal influences in check, only the phones appearing in syllabic structures also present in the native data were selected. Because of the nature of the KNN method, 1,000 passes were carried out on the phones of each speaker in each session in order to select the optimal k – *i.e.* the value of k returning the highest classification accuracy. In each pass, k was allowed to vary from 1 neighbour to $\sqrt{n}$ neighbours, with n being the total number of phones for a given speaker in a given session. Because of the skewed distribution of realizations of the different phonemic categories in the native subcorpus, the training data consisted of the P&B data. The results at first sight did not reveal any clear patterns that would support a rejection of the null hypothesis: the four phonemes under study were classified with very equivalent success rates. A look at the second best alternates, however, revealed phonological inconsistencies, with /ʊ/ and /uː/ quite often predicted to be front vowels. More crucially, for these two phonemes, the prediction rate of the second best alternates were often higher than their own prediction rate, therefore supporting, albeit tentatively, the idea that /ʊ/ and /uː/ have values less similar to native targets than /ɪ/ and /iː/. These studies, however, did not really establish anything regarding the truly longitudinal aspect of phonemic acquisition.

In order to check whether an effect existed, an experiment based on linear mixed-effects regressions. Several models were compared, with temporal effects predicted in turn to be either non-existent (flat curve), increasing or decreasing (slope), or evolving (quadratic). Various response variables were investigated, mostly involving distances from native values – either in the BDM-normalized F_1/F_2 space or in terms of standard deviations. Although the results must be interpreted with great care, the consistency of the evolution towards native values of the response variables in the case of /ɪ/ and /iː/, when compared to the absence of evolution or the greater distances from native values in the case of /ʊ/ and /uː/, once

again seems to suggest that the acquisition over time of /ɪ/ and /iː/ is better than /ʊ/ and /uː/. In a final endeavour to establish whether the acquisitions of the two contrasts were similar, the study of the entire signal of the first three formants throughout the duration of the vowel was undertaken. In order to do so, the signals, emulated by the formant tracks made up of the measurements at each centile of the vowels' durations, had to be modelled. This was done with discrete cosine transform, which allowed the reduction of the number of parameters for each vowel from 300 to 6, after BDM normalization. Comparisons then again were made with native values, after selection of the tokens embedded in syllables also present in the native subcorpus. Using a procedure that takes into account the original distributions of the phonemic categories, /uː/ was consistently found to be underrepresented in the expected proportions of tokens similar to native values. From such findings the conclusion was drawn, tentatively again, that the /ɪ/-/iː/ contrast is acquired in a more robust manner than the /ʊ/-/uː/ contrast. How different would these findings have been if mid-temporal formant values had been used instead? What are the advantages of resorting to a much more intensive, coding-wise and computation-wise, method of analysis? These crucial questions were addressed by comparing models based on either DCT coefficients or mid-temporal values, with durations added as a variable in each combination. How similar to native values the learners' tokens were according to the processing method adopted was established by running quadratic discriminant analyses. The influence across the board of including the durations of vowels was demonstrated extremely clearly. Regarding the efficiency in classification accuracy, the simpler models based on mid-temporal values fared extremely well, but the full DCT model using all three coefficients shone by recognizing tokens with very low prior probabilities.

Should, then, the null hypothesis that both /ɪ/-/iː/ and /ʊ/-/uː/ contrasts are acquired at the same rate, be rejected or accepted? This work ultimately suggests, from a body of tentatively corroborating evidence, that it should be rejected. A definitive answer, if there

ever was to be one, should, it is contended, use spontaneous speech as data, because the frequency of occurrences of tokens of given phonemic categories is bound to influence acquisition. Because they maximize the quantity of information available for analysis, and make it possible to handle skewed distributions (a defining feature of spontaneous speech), DCTs should be the processing method of choice for further study. This is the conclusion this work most vehemently asserts.

What further steps to take, then? In the face of the data collected, it looks as if the present study has barely scratched the surface of what can be investigated. The influence of syllabic structures, the alignment of which took so much research time; the role of word frequencies; of lexical nature; of the tasks the students were accomplishing when recorded; none of these parameters, although they are readily available in the data, have been investigated. Other methods of visualization, such as kernel density plots, could be used[14]. LMERs could help establish the role of tasks in the pronunciation. Diphthongs could also be analyzed, along with the differences in realizations between allophonic /iː/ and /i/, or /uː/ and /u/. This work ends, then, on a simple hope: that the collected data will be put to much better use than it has been so far.

[14] Ballier & Méli in Appendix F gives an attempt to use them.

References

Abrahamsson, N. (2003). DEVELOPMENT AND RECOVERABILITY OF l2 CODAS. *Studies in Second Language Acquisition*, *25*(03).
URL https://doi.org/10.1017/s0272263103000147

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*(5), 3099.
URL http://dx.doi.org/10.1121/1.1795335

Al-Tamimi, J.-E., & Ferragne, E. (2005). Does vowel space size depend on language vowel inventories? evidence from two Arabic dialects and French. In *Proceedings of Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology*, (pp. 2465–2468).

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC map task corpus. *Language & Speech*, *34*, 351–366.

Assmann, P. F., Nearey, T. M., & Bharadwaj, S. V. (2012). Developmental patterns in children's speech: Patterns of spectral change in vowels. In *Vowel Inherent Spectral Change*, (pp. 199–230). Springer Science + Business Media.
URL http://dx.doi.org/10.1007/978-3-642-14209-3_9

Baayen, H. R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database. Release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Bagwell, C. (2018). SOX 14.4.2 – SOund eXchange.
URL http://sox.sourceforge.net/

Ballier, N. (2014). La modélisation statistique du rythme et la dissolution de la structure syllabique. In C. Blanckaert, J. Léon, & D. Samain (Eds.) *Modèles et modélisations en sciences du langage, de l'homme et de la société. Perspectives historiques et épistémologiques*, (pp. 401–417). Paris: L'Harmattan.

Barreda, S. (2014). *phonTools: Functions for phonetics in R*. R package version 0.2-2.0.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Benesty, J., Sondhi, M. M., & Huang, Y. A. (2007). *Springer Handbook of Speech Processing*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.) *Speech perception and linguistic experience: Theoretical and methodological issues*, (pp. 171–204). Baltimore: York Press.

Bigi, B. (2012a). Sppas: A tool for the phonetic segmentations of speech. In LREC (Ed.) *Proc. of LREC 2012*, (pp. 1748–1755).

Bigi, B. (2012b). SPPAS: a tool for the phonetic segmentations of Speech. In *The eight international conference on Language Resources and Evaluation*, (pp. 1748–1755). Istanbul.

Bigi, B., & Hirst, D. (2012). SPeech Phonetization Alignment and Syllabification (SPPAS):a tool for the automatic analysis of speech prosody. In *Speech Prosody*, (pp. 19–22). Shanghai.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. version 5.3.62, retrieved 2 january 2014 from http://www.praat.org/.

Bradlow, A. (1995). A comparative acoustic study of english and spanish vowels. *Journal of the Acoustical Society of America*, *97*(3), 1916–1924.

Brand, C., & Götz, S. (2013). Fluency versus accuracy in advanced spoken learner language. In *Benjamins Current Topics*, (pp. 117–137). John Benjamins Publishing Company.
URL https://doi.org/10.1075/bct.52.05bra

Brand, C., & Kämmmerer, S. (2006). The Louvain International Database of Spoken English Interlanguage (LINDSEI): Compiling the German component. In S. Braun, K. Kohn, & J. Mukherjee (Eds.) *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt am Main: Peter Lang.

Burin, L., & Ballier, N. (2017). Accommodation in learner corpora: A case study in phonetic convergence. *Anglophonia*, (24).
URL https://doi.org/10.4000/anglophonia.1127

Bybee, J. (2007). *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive Science*, *34*, 685–697.

Clopper, C., Pison, D., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical society of America*, *118*(3), 1661–1676.

Davenport, M., & Hannahs, S. (2013). *Introducing Phonetics and Phonology*. Taylor & Francis.
URL https://books.google.fr/books?id=qV1ACQAAQBAJ

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.
URL http://dx.doi.org/10.3758/BRM.41.2.385

Dell, F. (1973). *Les règles et les sons*. Collection Savoir. Hermann.
URL https://books.google.fr/books?id=XrA0AQAAIAAJ

Dellwo, V., & Wagner, P. (2003). Relations between language rhythm and speech rate. In *Proceeding of the 15th international congress of phonetic sciences*, (pp. 471 - 474). Barcelona.

Dowle, M., & Srinivasan, A. (2017). *data.table: Extension of 'data.frame'*. R package version 1.10.4.
URL https://CRAN.R-project.org/package=data.table

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification (2Nd Edition)*. Wiley-Interscience.

Ferragne, E., & Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, *40*(1), 1–34.

Flege, J. (1995). Second-language Speech Learning: Theory, Findings, and Problems. In W. Strange (Ed.) *Speech Perception and Linguistic Experience: Issues in cross-language research*, (pp. 233–277). Timonium, MD: York Press.

Flege, J. (2005). *Origins and development of the Speech Learning Model*. 1st ASA Workshop on L2 Speech Learning: Simon Fraser Univ., Vancouver, BC.
URL http://jimflege.com/files/Vancouver_April_2005.pdf

Gendrot, C., & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In Eurospeech (Ed.) *Proceedings Eurospeech*, (pp. 2453–2456).

Gendrot, C., & Adda-Decker, M. (2007). Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages. In International Conference on Phonetics Sciences. In *International Conference on Phonetics Sciences*, (pp. 1417–1420). Saarbrücken, Germany.
URL https://halshs.archives-ouvertes.fr/halshs-00188113

Gnevsheva, K. (2015). Acoustic analysis in theAccents of non-native english(ANNE) corpus. *International Journal of Learner Corpus Research*, *1*(2), 256–267.
URL https://doi.org/10.1075/ijlcr.1.2.04gne

Goutéraux, P. (2013). Learners of English and Conversational Proficiency. In S. Granger, G. Gilquin, & F. Meunier (Eds.) *20 Years of Corpus Research: Looking back, Moving ahead (Corpora and Language in Use 1)*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Harrington, J. (2010). *Phonetic Analysis of Speech Corpora*. Wiley Publishing.

Hillenbrand, J., Getty, L., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Hillenbrand, J. M. (2012). Static and Dynamic Approaches to Vowel Perception. In *Vowel Inherent Spectral Change*, (pp. 9–30). Springer Science + Business Media.
URL http://dx.doi.org/10.1007/978-3-642-14209-3_2

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013–3022.

Hillenbrand, J. M., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, *109*(2), 748–763.
URL http://dx.doi.org/10.1121/1.1337959

Iverson, P., & Evans, B. G. (2009). Learning english vowels with different first-language vowel systems II: Auditory training for native spanish and german speakers. *The Journal of the Acoustical Society of America*, *126*(2), 866–877.
URL https://doi.org/10.1121/1.3148196

Jones, D., Roach, P., Setter, J., & Esling, J. (2011). *Cambridge English Pronouncing Dictionary*. Cambridge University Press.
URL https://books.google.fr/books?id=bfLXAAAAQBAJ

Jongman, A., Fourakis, M., & Sereno, J. A. (1989). The acoustic vowel space of modern greek and german. *Language and Speech*, *32*(3), 221–248.
URL https://doi.org/10.1177/002383098903200303

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2004). Domain-initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.) *Papers in Laboratory Phonology VI : Phonetic interpretation.*, (pp. 145–163.). Cambridge: CUP.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B*, *363*, 979–1000.

Labov, W., Ash, S., & Boberg, C. (2006). *Atlas of North American English: Phonetics, Phonology and Sound Change ; a Multimedia Reference Tool*. Mouton de Gruyter.
URL https://books.google.co.uk/books?id=qa4-dFqi6iMC

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Phonological Theory. Wiley.
URL https://books.google.co.uk/books?id=h1byJz_rWUcC

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*, 839–862.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Ohala, & J. Jaeger (Eds.) *Experimental Phonology*. New-York: Academic New York.

Lobanov, B. (1971). Classification of Russian vowels spoken by different speakers. *J. Acoust. Soc. Am.*, *49*, 606–608.

Long, J. (2012). *Longitudinal Data Analysis for the Behavioral Sciences Using R*. Sage Publications, Inc.

Mazerolle, M. J. (2017). *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*. R package version 2.1-1.
URL https://cran.r-project.org/package=AICcmodavg

Middleton, E., & Schwartz, M. (2010). Density pervades: an analysis of phonological neighbourhood den- sity effects in aphasic speakers with different types of naming impairment. *Cognitive Neuropsychology*, *27 (5)*, 401-427.

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational english. *Language and speech*, *21*(3), 221–241.

Morrison, G. (2008). Comment on "a geometric representation of spectral and temporal vowel features: Quantification of vowel overlap in three linguistic varieties". *Journal of the Acoustical Society of America*, *123*(1), 37–40.

Morrison, G., & Nearey, T. (2006). A cross-language vowel normalisation procedure. *Canadian Acoustics*, *34*(3), 94–95.

Morrison, G. S. (2012). Theories of Vowel Inherent Spectral Change. In *Vowel Inherent Spectral Change*, (pp. 31–47). Springer Science + Business Media.
URL http://dx.doi.org/10.1007/978-3-642-14209-3_3

Nearey, T., & Assmann, P. (1986). Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *125*, 2387-97.

Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels*. Ph.D. thesis, Indiana University Linguistics Club.

Nearey, T. M. (2012). Vowel Inherent Spectral Change in the Vowels of North American English. In *Vowel Inherent Spectral Change*, (pp. 49–85). Springer Science + Business Media.
URL http://dx.doi.org/10.1007/978-3-642-14209-3_4

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Stud. Sec. Lang. Acq.*, *29*(04).
URL http://dx.doi.org/10.1017/S027226310707043X

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. L. Bybee, & P. Hopper (Eds.) *Frequency and the Emergence of Linguistic Structure*, (pp. 137–157). John Benjamins Publishing Company.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/

Strik, H., Truong, K. P., de Wet, F., & Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, (pp. 1837–1840).
URL http://www.isca-speech.org/archive/interspeech_2007/i07_1837.html

Sundberg, J. (1977). *The Acoustics of the Singing Voice*. Scientific American offprints. W.H. Freeman.
URL https://books.google.fr/books?id=Q1-3tgAACAAJ

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of american english vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100.

Tabain, M., Breen, G., & Butcher, A. (2004). VC vs. CV syllables: a comparison of Aboriginal languages with English. *Journal of the International Phonetic Association*, *34*, 175–200.

Tambovtsev, Y., & Martindale, C. (2007). Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, *4*(2), 1–11.

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of french. *IRAL*, *40*(2), 117–150.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of french. *Applied Linguistics*, *17*(1), 84–119.
URL http://dx.doi.org/10.1093/applin/17.1.84

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100.

Trouvain, J., & Barry, W. (Eds.) (2007). *TRIAL*. Saarbrücken: Universität des Saarlandes.

Tubach, J.-P. (1989). *La parole et son traitement automatique*. Paris: Masson.

Vaissière, J. (2006). *La Phonétique*. Paris: Presses Universitaires de France.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, fourth ed. ISBN 0-387-95457-0.
URL http://www.stats.ox.ac.uk/pub/MASS4

Wand, D., & Fabricius, A. (2002). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1 ~ F2 plane. *Leeds Working Papers in Linguistics and Phonetics*, *9*, 159-173.

Weide, R. (1994). CMU Pronouncing Dictionary.
URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Wells, J. (2008). *Longman Pronunciation Dictionary*. London: Pearson Longman.

Wells, J. C. (1990). *Pronunciation Dictionary*. London: Longman.

White, K., Yee, E., Blumstein, S., & Morgan, J. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, *68*, 362-378.

Winkelmann, R., Jaensch, K., Cassidy, S., & Harrington, J. (2016). *emuR: Main Package of the EMU Speech Database Management System*. R package version 0.1.6.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America,*, *123(5)*, 5687-5690.

Appendix A

Extra tasks for Session 4

A.1 Map task

(from Anderson et al. (1991))

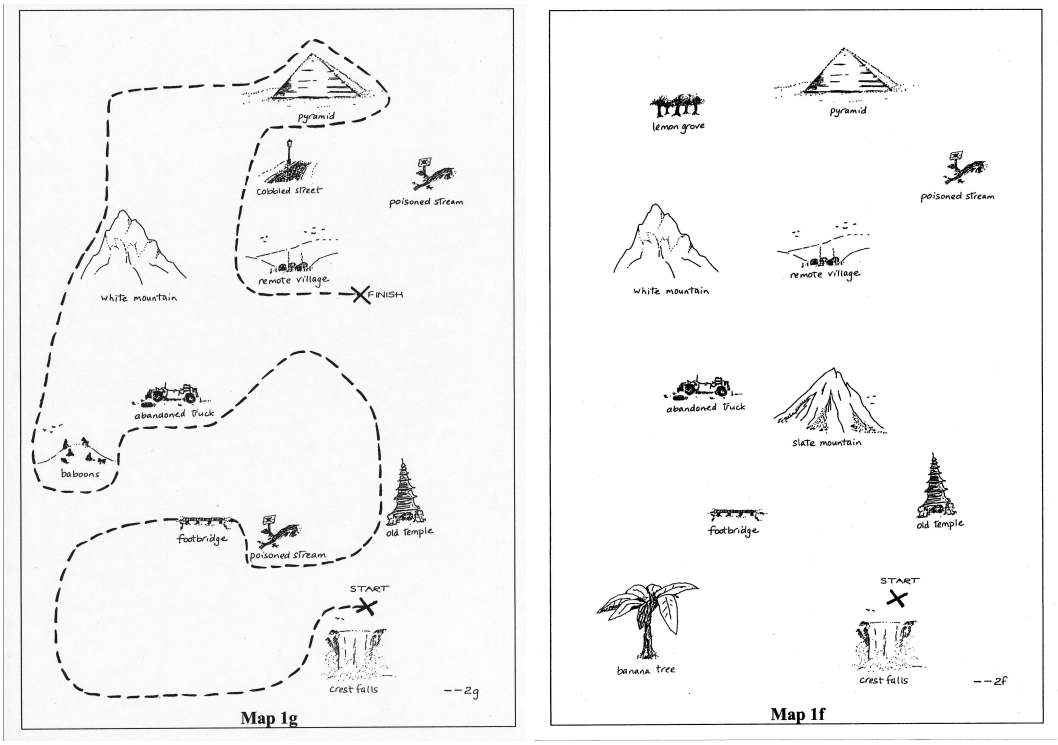


Fig. A.1: Map task for Session 4. On the left, the map the learner was given, with an instruction to guide the native speaker to the finish. On the right, the map the native speaker was given.

A.2 Reading lists

A.2.1 Reading task n°3: list of words

A: Read the following words:

1. ship, sick, milk, myth
2. step, shelf, friend, ready
3. bad, cab, hand, cancel
4. stop, rob, possible, quality
5. cub, rub, trunk, blood
6. full, put, look, good
7. staff, clasp, ask, dance
8. cross, long, off, origin
9. hurt, term, work, firm
10. seem, key, feel, people
11. weight, tape, great, day
12. ask, calm, spa, father
13. door, caught, law, broad
14. soap, soul, home, know
15. who, group, few, tune
16. ripe, night, buy, high
17. boy, noise, coin, royal
18. house, noun, crowd, now
19. beer, pier, fear, serious
20. wear, care, air, where
21. far, sharp, farm, heart
22. war, storm, for, born
23. floor, coarse, wore, oral
24. poor, tourist, pure, fury

B. Pronounce the following words, which have the same vowel as in:

1. KIT
   him, big, village, women, it
2. SEAT
   sea, feet, field, see
3. PUT
   put, wolf, good, look, pull
4. SHOE
   soon, do, soup, shoe, too, pool

A.2.2 Reading task n°4: *Le géant égoïste*

Lisez le texte suivant:

Le géant égoïste, d'Oscar WILDE

Chaque après-midi, en revenant de l'école, les enfants allaient jouer dans le jardin du géant. C'était un grand et beau jardin au doux gazon vert. Çà et là, sur le gazon, de belles fleurs brillaient comme des étoiles et il y avait douze pêchers qui, au printemps, se couvraient d'une délicate floraison rose et blanche et à l'automne portaient de beaux fruits. Les oiseaux perchés sur les arbres chantaient si bien que les enfants avaient coutume d'arrêter leurs jeux pour les écouter. « Comme nous sommes heureux ici ! » s'écriaient-ils souvent. [...] Un jour, le géant revint. [...] « Que faites-vous là ? » cria-t-il d'une voix très bourrue. Et les enfants s'enfuirent. [...] C'était un géant très égoïste. [...]

Un matin, le géant se prélassait dans son lit, lorsqu'il entendit une musique délicieuse. Elle était si douce à ses oreilles qu'il crut que les musiciens du roi passaient par là. [...] Il vit une scène stupéfiante.

Appendix B

LONGDALE transcription guidelines

http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/transnew.htm

1. Interview identification

DID03chiffres –S002

2. Speaker turns

Speaker turns are displayed in vertical format, i.e. one below the other. Whilst the letter 'A' enclosed between angle brackets always signifies the interviewer's turn, the letter 'B' between angle brackets indicates the interviewee's (learner's) turn. The end of each turn is indicated by either </A> or </B>.

e.g.: <A> okay so which topic have you chosen </A>

<B> the film or play that I thought was particularly good or bad really </B>

3. Overlapping speech

The tag <overlap /> (with a space between "overlap" and the slash) is used to indicate the beginning of overlapping speech. It should be indicated in both turns.

e.g.: <B> yeah I went on a bus to London once and I'll never <overlap /> do it again </B>

<A> <overlap /> that's even worse </A>

4. Punctuation

No punctuation marks are used to indicate sentence or clause boundaries.

5. Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing. The following three tier system is used: one dot for a 'short' pause (< 1 second), two dots for a 'medium' pause (1-3 seconds) and three dots for 'long' pauses (> 3 seconds).

e.g.: <B> erm .. it's a British film there aren't many of those these days </B>

6. Filled pauses and backchannelling

Filled pauses and backchannelling are marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm). No other fillers should be used.

e.g.: <B> yeah . well Namur was warmer (er) it was (eh) a really little town </B>

7. Unclear passages

A three tier system is used to indicate the length of unclear passages: <X> represents an unclear syllable or sound up to one word, <XX> represents two unclear words, and <XXX> represents more than two words.

e.g.: <B> <X> they're just begging <XX> there's there's honestly he did a course .. for a few weeks </B>

If transcribers are not entirely sure of a word or word ending, they should indicate this by having the word directly followed by the symbol '<?>'.

e.g.: <B>: I went to see a<?> friend at university there and stayed </B>

Unclear names of towns or titles of plays for example may be indicated as '<name of city>' or '<title of play>'.

e.g.: <B>: where else did we go er <name of city> it's in Bolivia </B>

8. Truncated words

Truncated words are immediately followed by an equals sign.

e.g.: <B> it still resem= resembled the theatre </B>

9. Contracted forms

All standard contracted forms are retained as they are typical features of speech.

10. Non-standard forms

Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way: *cos, dunno, gonna, gotta, wanna* and *yeah*.

11. Foreign words and pronunciation

Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word).

e.g.: <B> we couldn't go with er knives and so on <foreign> enfin </foreign> we were er </B>

As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag.

e.g.: <B> I didn't have the erm . <foreign> distinction </foreign> </B>

12. Acronyms

If acronyms are pronounced as sequences of letters, they are transcribed as a series of upper-case letters separated by spaces.

e.g.: <B> yes not really I did sort of basic G C S E French and German </B>

If, on the other hand, acronyms are pronounced as words, they are transcribed as a series of upper-case letters not separated by spaces.

e.g.: <A> mhm er you're doing a MAELT </A>

13. Dates and numbers

Figures have to be written out in words. This avoids the ambiguity of, for example, "1901", which could be spoken in a number of different ways.

e.g.: <B> an awful lot of people complain and say well the grants were two thousand two hundred </B>

14. Nonverbal vocal sounds

Nonverbal vocal sounds are enclosed between angle brackets.

e.g.: <B> I hope so I've I've got some <coughs> friends out there </B>

e.g.: <B> so I went back into Breda . and sat down again <imitates the sound of a guitar> </B>

15. Contextual comments

Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example).

e.g.: <A> no it's true it's nice to have your own bathroom </A>

<somebody enters the room>

<B> hi </B>

16. Prosodic information: voice quality

If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <begin laughter> or <begin whisper> immediately before the specific stretch of speech and <end laughter> or <end whisper> at the end of it.

e.g.: <B> <begin laughter> I don't have to assess it I only have to write it <end laughter> </B>

17. Phonetic features

(a) Syllable lengthening

A colon is used to indicate that the preceding syllable is lengthened. Colons should not be inserted inside words.

e.g.: <B> that's something I'll I'll plan to: to learn </B>

(b) Articles

-when pronounced as [ei], the article 'a' is transcribed as 'a[ei]';

e.g.: <B> and it's about erm . life in a[ei] eh public school in America I think </B>

-when pronounced as [i:] the article 'the' is transcribed as 'the[i:]'.

e.g.: <B> and the[i:] villa we were staying in was in one of the valleys </B>

18. Tasks

The three tasks making up the interview (set topic, free discussion and picture description) should be separated from each other. This is done using the following tags: <S> (before the set topic), </S> (after the set topic), <F> (before the free discussion), </F> (after the free discussion), <P> (before the picture description), </P> (after the picture description). These tags should occupy a separate line and should not interrupt a turn.

e.g.: <S>

<A> did you . manage to choose a topic </A>

19. End

All interviews should end with the following tag (on a separate line): </h>

20. Questions?

If you have any questions regarding these transcription guidelines, don't hesitate to get in touch with us!

Appendix C

Code snippets

C.1 P2FA Bash script

This is the `bash` script used to downsample the original audio files to 16,000Hz using `Sox`, convert the transcriptions to upper case (as required by P2FA), and execute the python script on a set of files in the folder.

```
#!/bin/bash
for txtf in *.txt
do
   EXT='.txt'
   RAD=${txtf%$EXT}
   P2F="$RAD-pp2fa"
   dd if=$RAD.txt of=$P2F.txt conv=ucase
   sox $RAD.wav -r 16000 -c 1 $P2F.wav
   python2 /home/adrien/softs/p2fa/align.py $RAD.wav $P2F.txt $P2F.TextGrid
done
```

C.2 Calculation of EPENTHETIC

This is the experimental section of `PRAAT03` that aims at detecting potential vocalic fillers after syllables at the end of words featuring a consonantal coda.

```
startOfIntep1= startOfIntep + 0.05
startOfIntep2= startOfIntep + 0.1
startOfIntep3= startOfIntep + 0.15
startOfIntep4= startOfIntep + 0.2
select pitch
vF0e1 = Get value at time.. 'startOfIntep1' Hertz Linear
vF0e2 = Get value at time.. 'startOfIntep2' Hertz Linear
```

```
vF0e3 = Get value at time.. 'startOfIntep3' Hertz Linear
vF0e4 = Get value at time.. 'startOfIntep4' Hertz Linear
select intensity
meanIntensity_ep = Get mean\ldots startOfIntep1 startOfIntep4 dB
if (min_int_ep>40 and vF0e1<>undefined and vF0e2<>undefined and
    ↪ vF0e3<>undefined and vF0e3<>undefined)
epenthetic$="YES"
endif
```

C.3 Modified list of English phonemes for SPPAS syllabification algorithm

The table below presents the English phonemic categories to be used by SPPAS algorithm-based syllabification. It is converted from the built-in French categories in SPAAS. The symbol after `PHONCLASS` lists an English phoneme as represented in either the SAMPA version of the CUMPD or the LPD (*c.f.* Table 1.4). The symbol next to it indicates the manner of articulation of the phoneme, which the native system of rules in SPPAS uses to create syllabic boundaries.

```
# vowels (P)
PHONCLASS i V
PHONCLASS e V
PHONCLASS E V
PHONCLASS a V
PHONCLASS A V
PHONCLASS O V
PHONCLASS o V
PHONCLASS u V
PHONCLASS y V
PHONCLASS 2 V
PHONCLASS 9 V
PHONCLASS @ V
PHONCLASS EU V
PHONCLASS I V
PHONCLASS i:  V
PHONCLASS u:  V
PHONCLASS U V
PHONCLASS  V
PHONCLASS aI V
PHONCLASS eI V
PHONCLASS OI V
PHONCLASS aU V
PHONCLASS @U V
PHONCLASS eU V
PHONCLASS O: V
PHONCLASS 3:r V
PHONCLASS V V
# glides (G)
PHONCLASS j G
PHONCLASS H G
PHONCLASS w G
# liquids (L)
PHONCLASS l L
PHONCLASS R L
PHONCLASS r L
# occlusives (O) PHONCLASS p O
PHONCLASS t O
PHONCLASS k O
PHONCLASS b O
PHONCLASS d O
PHONCLASS g O
PHONCLASS 4 O
# affricates (A)
PHONCLASS tS A
PHONCLASS dZ A
# nasals (N)
PHONCLASS N N
PHONCLASS n N
PHONCLASS m N
PHONCLASS N N
PHONCLASS J N
# fricatives (F)
PHONCLASS s F
PHONCLASS S F
PHONCLASS z F
PHONCLASS Z F
PHONCLASS v F
PHONCLASS f F
PHONCLASS D F
PHONCLASS T F
PHONCLASS h F
```

C.4 Pitch in PRAAT03

The following piece of code can be found from line 228 onwards of PRAAT03. As per the PRAAT manual, the time step was set to 0; the pitch-floor, to 75 Hz for men, 100 for women; the pitch ceiling was set to 300 Hz for men, 500 Hz for women.

```
if (sex$=="MALE")
maxfreq = 5000
minpitch = 75
maxpitch = 300
else
maxfreq = 5500
minpitch = 100
maxpitch = 500
endif
To Formant (burg)... 0 5 maxfreq 0.025 50
select Sound 'soundfile$'
To Pitch... 0 minpitch maxpitch
pitch = selected("Pitch")
select Sound 'soundfile$'
To Intensity... 75 0.001
```

C.5 Syllable check

This *R* script

```
{
library(data.table)
rm(list=ls(all=TRUE))

   {
      sppas <- as.data.frame(fread('sppas-global.csv', stringsAsFactor =
         ↪ TRUE))
      p2f  <- as.data.frame(fread('p2f-global.csv', stringsAsFactor = TRUE))
      #        global$UKPHONEME[global$UKPHONEME==''']<-"i"
      #vv <- c(1:4,6:9,13:25,31:41)
      #for (i in 1:length(vv)) {
      #  global[,vv[i]] <- factor(global[,vv[i]])
      #}
      #        global <- na.omit(global)
   }

}
{
   # sppas
   # cleaning data
   # checking syllables
```

```
   table(sppas$ESKELS)
   sppascompilesyllerrorsdf <- data.frame()
   tl=length(dimnames(table(sppas$ESKELS))[[1]]);tl
   for(i in 1:tl) {
      if (dimnames(table(sppas$ESKELS))[[1]][i]!="CV" &
          ↪ dimnames(table(sppas$ESKELS))[[1]][i]!="CVC" &
          ↪ dimnames(table(sppas$ESKELS))[[1]][i]!="V" &
          ↪ dimnames(table(sppas$ESKELS))[[1]][i]!="VC" ) {
         tempdf <- sppas[sppas$ESKELS==dimnames(table(sppas$ESKELS))[[1]][i]
            ↪ ,c(1,3,4,6,7,12,17,23)];tempdf
         sppascompilesyllerrorsdf <- rbind(sppascompilesyllerrorsdf,tempdf);
      }
   }
   # compilesyllerrorsdf
   write.table(sppascompilesyllerrorsdf,file='sppascompilesyllerrors.txt',sep='\t',row.names=
}
{
   # p2f
   # cleaning data
   # checking syllables
   table(p2f$ESKELS)
   p2fcompilesyllerrorsdf <- data.frame()
   tl=length(dimnames(table(p2f$ESKELS))[[1]]);tl
   for(i in 1:tl) {
      if (dimnames(table(p2f$ESKELS))[[1]][i]!="CV" &
          ↪ dimnames(table(p2f$ESKELS))[[1]][i]!="CVC" &
          ↪ dimnames(table(p2f$ESKELS))[[1]][i]!="V" &
          ↪ dimnames(table(p2f$ESKELS))[[1]][i]!="VC" ) {
         tempdf <- p2f[p2f$ESKELS==dimnames(table(p2f$ESKELS))[[1]][i]
            ↪ ,c(1,3,4,6,7,12,17,23)];tempdf
         p2fcompilesyllerrorsdf <- rbind(p2fcompilesyllerrorsdf,tempdf);
      }
   }
   # compilesyllerrorsdf
   write.table(p2fcompilesyllerrorsdf,file='p2fcompilesyllerrors.txt',sep='\t',row.names=F,qu
}

npscheck<-sppas[sppas$NPW>4,c(1:4,17)]
write.table(npscheck,file='npscheck.csv',sep='\t',row.names=F,quote=F)
sppasp2fsc<-sppas[sppas$SPPASSC!=sppas$P2FSC,c(1:4)]
write.table(sppasp2fsc,file='sppasp2fsc.csv',sep='\t',row.names=F,quote=F)
```

<++>

C.6 Common R code

This is the introductory snippet used in R scripts from chapter 2 onwards to extract monophthongs with durations longer than 0.03s. `fread` (Dowle & Srinivasan (2017)) enables much faster loading times. Only monophthongs, as defined by the LPD, are selected (see the reasons why in section 2.1). The levels of the obtained dataframes were reorganized to a more intuitive, alphabetical order. Colour codes were also added, in order to ensure that monopthongs were assigned the same colours across all graphs.

```
{
   rm(list=ls(all=TRUE))
   # ordering levels by alpha order
   alpha <- c("&", "A:", "e", ":", "@", "I", "i:", "i", "Q", "O:", "V", "U",
       ↪ "u:", "u")
   # for LaTeX conversion
   tipa <- c("\textipa{\ae}", "\textipa{a:}", "\textipa{e}", "\textipa{3:}",
       ↪ "\textipa{@}", "\textipa{I}", "\textipa{i:}", "\textipa{i}",
       ↪ "\textipa{6}", "\textipa{O:}", "\textipa{2}", "\textipa{U}",
       ↪ "\textipa{u:}", "\textipa{u}")
   # for scatterplot3d use this:
   tipa2 <- c("\\textipa{\\\\ae}", "\\textipa{a:}", "\\textipa{e}",
       ↪ "\\textipa{3:}", "\\textipa{@}", "\\textipa{I}", "\\textipa{i:}",
       ↪ "\\textipa{i}", "\\textipa{6}", "\\textipa{O:}", "\\textipa{2}",
       ↪ "\\textipa{U}", "\\textipa{u:}", "\\textipa{u}")

   # colour codes
   colourstyles <- c("dodgerblue3", "dodgerblue4", "steelblue3",
       ↪ "steelblue4", "black", "seagreen4", "seagreen3", "seagreen2",
       ↪ "plum3", "plum4", "grey30", "firebrick4", "firebrick3",
       ↪ "firebrick2")
   # line styles
   linestyles=rep(c(1:6),2)
   linestyles <- append(linestyles,c(1:3))

   #sppas
   sppasglobal <- as.data.frame(fread('$MYPATH',stringsAsFactor = TRUE))
   # let's exclude too short durations
   # clean sppas global
   sppas <- sppasglobal[sppasglobal$PHONDUR>0.03,]

   #p2f
   p2fglobal <- as.data.frame(fread('$MYPATH',stringsAsFactor = TRUE))
   # let's exclude too short durations
   # clean p2f global
   p2f <- p2fglobal[p2fglobal$PHONDUR>0.03,]

   # natives
   sppasnatives <- as.data.frame(fread('$MYPATH',stringsAsFactor = TRUE))
   sn <- sppasnatives[sppasnatives$PHONDUR>0.03,]
   p2fnatives <- as.data.frame(fread('$MYPATH',stringsAsFactor = TRUE))
```

```
pn <- p2fnatives[p2fnatives$PHONDUR>0.03,]

# lists of words
sppaslist <- as.data.frame(fread('$MYPATH',stringsAsFactor = TRUE))
sl <- sppaslist[sppaslist$PHONDUR>0.03,]
p2flist <- as.data.frame(fread('$MYPATH',stringsAsFactor = TRUE))
pl <- p2flist[p2flist$PHONDUR>0.03,]

# monopthongs
# sppas
sm <- sppas[sppas$PHONDUR>0.03,]
sm<-sm[!(sm$LPDPHONEME %in% c("", "aI@", "e@", "eI@", "i:@", "U@", "u@",
    ↪ "eI", "aU", "@U", "aI", "OI", "I@", "O:R", ":", ":@", ":aI", "::",
    ↪ "A", "QU")),]
sm$WORD <- gsub("it'll", "it",sm$WORD)
sm$WORD <- gsub("it's", "it",sm$WORD)
sm$WORD <- gsub("it'd", "it",sm$WORD)
sm$WORD <- factor(sm$WORD)
sm$LPDPHONEME <- factor(sm$LPDPHONEME)
sm$LPDPHONEME <- factor(sm$LPDPHONEME,levels=alpha)
# p2f
pm <- p2f[p2f$PHONDUR>0.03,]
pm<-pm[!(pm$LPDPHONEME %in% c("", "aI@", "e@", "eI@", "i:@", "U@", "u@",
    ↪ "eI", "aU", "@U", "aI", "OI", "I@", "O:R", ":", ":@", ":aI", "::",
    ↪ "A", "QU")),]
pm$LPDPHONEME <- factor(pm$LPDPHONEME)
pm$LPDPHONEME = factor(pm$LPDPHONEME,levels=alpha)
# natives
snm <-sn[sn$PHONDUR>0.03,]
snm<-snm[!(snm$LPDPHONEME %in% c("", "aI@", "e@", "eI@", "i:@", "U@",
    ↪ "u@", "eI", "aU", "@U", "aI", "OI", "I@", "O:R", ":", ":@", ":aI",
    ↪ "::", "A", "QU")),]
snm$LPDPHONEME <- factor(snm$LPDPHONEME)
snm$LPDPHONEME <- factor(snm$LPDPHONEME,levels=alpha)
pnm <-pn[pn$PHONDUR>0.03,]
pnm<-pnm[!(pnm$LPDPHONEME %in% c("", "aI@", "e@", "eI@", "i:@", "U@",
    ↪ "u@", "eI", "aU", "@U", "aI", "OI", "I@", "O:R", ":", ":@", ":aI",
    ↪ "::", "A", "QU")),]
pnm$LPDPHONEME <- factor(pnm$LPDPHONEME)
pnm$LPDPHONEME <- factor(pnm$LPDPHONEME,levels=alpha)
#list of words
slm <-sl[sl$PHONDUR>0.03,]
slm<-slm[!(slm$LPDPHONEME %in% c("", "aI@", "e@", "eI@", "i:@", "U@",
    ↪ "u@", "eI", "aU", "@U", "aI", "OI", "I@", "O:R", ":", ":@", ":aI",
    ↪ "::", "A", "QU")),]
slm$LPDPHONEME <- factor(slm$LPDPHONEME)
slm$LPDPHONEME <- factor(slm$LPDPHONEME,levels=alpha)
plm <-pl[pl$PHONDUR>0.03,]
```

```
   plm<-plm[!(plm$LPDPHONEME %in% c("", "aI@", "e@", "eI@", "i:@", "U@",
       ↪ "u@", "eI", "aU", "@U", "aI", "OI", "I@", "O:R", ":", ":@", ":aI",
       ↪ "::", "A", "QU")),]
   plm$LPDPHONEME <- factor(plm$LPDPHONEME)
   plm$LPDPHONEME <- factor(plm$LPDPHONEME,levels=alpha)
   # vector for F1, F2 & F3
   # F1 column numbers
   F1 <- seq(39,534,5);F1
   # F2 column numbers
   F2 <- seq(40,535,5);F2
   # F3 column numbers
   F3 <- seq(41,536,5);F3

   FF <- c(rbind(F1,F2,F3))
}
```

C.7 Code for Optimal Centiles

This code creates a function `opticent`, which returns the optimal centile of each monophthong, *i.e.* the centile with the lowest product of F_1, F_2 and F_3 standard deviations.

```
# this function returns the "optimal centile"
# for each vowel, i.e. the centile where the
# product of F1, F2 and F3 standard deviations
# is the lowest
# it takes a dataframe with at least a column
# for vowels, and all formant values for each centile
# on the other columns (301 columns minimum in total)
# input arguments are of the form:
# df is the dataframe with the 301 columns
# vowel is an integer specifying the column number
# where the vowels in the df are stored
# f1 is a vector specifying the 100 F1 values column
# f2 is a vector specifying the 100 F2 values column
# f3 is a vector specifying the 100 F3 values column


opticent <- function(df,vowel,f1,f2,f3){

   ssdf1 <- data.frame()
   ssdf2 <- data.frame()
   ssdf3 <- data.frame()
   for (i in 1:length(levels(df[,vowel]))) {
      # sd f1
      sdftempf1 <- df[df[,vowel]==levels(df[,vowel])[i],f1]
      svctempf1 <- apply(sdftempf1,2,sd)
      ssdf1 <- rbind(ssdf1,svctempf1)
```

```
      # sd f2
      sdftempf2 <- df[df[,vowel]==levels(df[,vowel])[i],f2]
      svctempf2 <- apply(sdftempf2,2,sd)
      ssdf2 <- rbind(ssdf2,svctempf2)

      # sd f2
      sdftempf3 <- df[df[,vowel]==levels(df[,vowel])[i],f3]
      svctempf3 <- apply(sdftempf3,2,sd)
      ssdf3 <- rbind(ssdf3,svctempf3)

   }
   # rearranging the dfs
   ssdf1$LPDPHONEME <- levels(df[,vowel])
   ssdf1$LPDPHONEME <- factor(ssdf1$LPDPHONEME)
   ssdf1 <- ssdf1[,c(101,1:100)]
   ssdf2$LPDPHONEME <- levels(df[,vowel])
   ssdf2$LPDPHONEME <- factor(ssdf2$LPDPHONEME)
   ssdf2 <- ssdf2[,c(101,1:100)]
   ssdf3$LPDPHONEME <- levels(df[,vowel])
   ssdf3$LPDPHONEME <- factor(ssdf3$LPDPHONEME)
   ssdf3 <- ssdf3[,c(101,1:100)]
   # F1 x F2 x F3
   tempoc <- as.matrix(ssdf1[,c(2:101)]) *
   as.matrix(ssdf2[,c(2:101)]) * as.matrix(ssdf3[,c(2:101)])
   # minimal values for each phoneme
   minv <- apply(tempoc,1,min)
   # on which centile is that minimum value?
   oc <- c()
   for (i in 1:length(levels(df[,vowel]))) {
      oc[i] <- which(tempoc[i,]==minv[i])
   }

   opticentdf <- data.frame(levels(df[,vowel]),oc)
   return(opticentdf)
}
```

C.8 Multimodel Comparisons

The following code snippet was used in section 3.4.3 to compute LMER models and compare them.

```
   {

        m1 <- lmer(RV ~ 1 + ( 1 | SPEAKER),DATAFRAME,REML = FALSE)
        m2 <- lmer(RV ~ SESSION + (1 | SPEAKER),DATAFRAME,REML = FALSE)
        m3 <- lmer(RV ~ SESSION + I(SESSION^2) + (1 |
            ↪ SPEAKER),DATAFRAME,REML = FALSE)
```

```
        m4 <- lmer(RV ~ SESSION + (SESSION | SPEAKER),DATAFRAME,REML = FALSE)
        #non linear
        m5 <- lmer(RV ~ SESSION + I(SESSION^2) + (SESSION |
            ↪ SPEAKER),DATAFRAME,REML = FALSE)

}
```

Appendix D

Dataframes: column names and lists of words

D.1 English dataframe

1. SPEAKER
2. SEX
3. SESSION
4. WORD
5. CLXFREQ
6. LPDPRON
7. PRON
8. LPDSC
9. SC
10. LPDPHONEME
11. PHONEME
12. LPDSYLL
13. ESYLLSTRUC
14. FSYLLSTRUC
15. ECVSTRUC
16. FCVSTRUC
17. ESKELS
18. FSKELS
19. STRESS
20. PHONDUR
21. ESDUR
22. FSDUR
23. LOCINFILE
24. INTNB
25. INTENSITY
26. PHONBEFORE
27. PRECOART
28. BEFVOICE
29. BEFMOA
30. BEFPOA
31. PHONAFTER
32. POSTCOART
33. AFTVOICE
34. AFTMOA
35. AFTPOA
36. EPENTHETIC
37. TOTALDUR
38. F01
39. F11
40. F21
41. F31
42. F41
43. F02
44. F12
45. F22
46. F32
47. F42
48. F03
49. F13
50. F23
51. F33
52. F43
53. F04
54. F14
55. F24
56. F34
57. F44
58. F05
59. F15
60. F25
61. F35
62. F45
63. F06
64. F16
65. F26
66. F36
67. F46
68. F07
69. F17
70. F27
71. F37
72. F47
73. F08
74. F18
75. F28
76. F38
77. F48
78. F09
79. F19
80. F29
81. F39
82. F49
83. F010
84. F110
85. F210
86. F310
87. F410
88. F011
89. F111
90. F211
91. F311
92. F411
93. F012
94. F112
95. F212
96. F312

```
 97. F412    148. F023    199. F133    250. F243
 98. F013    149. F123    200. F233    251. F343
 99. F113    150. F223    201. F333    252. F443
100. F213    151. F323    202. F433    253. F044
101. F313    152. F423    203. F034    254. F144
102. F413    153. F024    204. F134    255. F244
103. F014    154. F124    205. F234    256. F344
104. F114    155. F224    206. F334    257. F444
105. F214    156. F324    207. F434    258. F045
106. F314    157. F424    208. F035    259. F145
107. F414    158. F025    209. F135    260. F245
108. F015    159. F125    210. F235    261. F345
109. F115    160. F225    211. F335    262. F445
110. F215    161. F325    212. F435    263. F046
111. F315    162. F425    213. F036    264. F146
112. F415    163. F026    214. F136    265. F246
113. F016    164. F126    215. F236    266. F346
114. F116    165. F226    216. F336    267. F446
115. F216    166. F326    217. F436    268. F047
116. F316    167. F426    218. F037    269. F147
117. F416    168. F027    219. F137    270. F247
118. F017    169. F127    220. F237    271. F347
119. F117    170. F227    221. F337    272. F447
120. F217    171. F327    222. F437    273. F048
121. F317    172. F427    223. F038    274. F148
122. F417    173. F028    224. F138    275. F248
123. F018    174. F128    225. F238    276. F348
124. F118    175. F228    226. F338    277. F448
125. F218    176. F328    227. F438    278. F049
126. F318    177. F428    228. F039    279. F149
127. F418    178. F029    229. F139    280. F249
128. F019    179. F129    230. F239    281. F349
129. F119    180. F229    231. F339    282. F449
130. F219    181. F329    232. F439    283. F050
131. F319    182. F429    233. F040    284. F150
132. F419    183. F030    234. F140    285. F250
133. F020    184. F130    235. F240    286. F350
134. F120    185. F230    236. F340    287. F450
135. F220    186. F330    237. F440    288. F051
136. F320    187. F430    238. F041    289. F151
137. F420    188. F031    239. F141    290. F251
138. F021    189. F131    240. F241    291. F351
139. F121    190. F231    241. F341    292. F451
140. F221    191. F331    242. F441    293. F052
141. F321    192. F431    243. F042    294. F152
142. F421    193. F032    244. F142    295. F252
143. F022    194. F132    245. F242    296. F352
144. F122    195. F232    246. F342    297. F452
145. F222    196. F332    247. F442    298. F053
146. F322    197. F432    248. F043    299. F153
147. F422    198. F033    249. F143    300. F253
```

301. F353
302. F453
303. F054
304. F154
305. F254
306. F354
307. F454
308. F055
309. F155
310. F255
311. F355
312. F455
313. F056
314. F156
315. F256
316. F356
317. F456
318. F057
319. F157
320. F257
321. F357
322. F457
323. F058
324. F158
325. F258
326. F358
327. F458
328. F059
329. F159
330. F259
331. F359
332. F459
333. F060
334. F160
335. F260
336. F360
337. F460
338. F061
339. F161
340. F261
341. F361
342. F461
343. F062
344. F162
345. F262
346. F362
347. F462
348. F063
349. F163
350. F263
351. F363
352. F463
353. F064
354. F164
355. F264
356. F364
357. F464
358. F065
359. F165
360. F265
361. F365
362. F465
363. F066
364. F166
365. F266
366. F366
367. F466
368. F067
369. F167
370. F267
371. F367
372. F467
373. F068
374. F168
375. F268
376. F368
377. F468
378. F069
379. F169
380. F269
381. F369
382. F469
383. F070
384. F170
385. F270
386. F370
387. F470
388. F071
389. F171
390. F271
391. F371
392. F471
393. F072
394. F172
395. F272
396. F372
397. F472
398. F073
399. F173
400. F273
401. F373
402. F473
403. F074
404. F174
405. F274
406. F374
407. F474
408. F075
409. F175
410. F275
411. F375
412. F475
413. F076
414. F176
415. F276
416. F376
417. F476
418. F077
419. F177
420. F277
421. F377
422. F477
423. F078
424. F178
425. F278
426. F378
427. F478
428. F079
429. F179
430. F279
431. F379
432. F479
433. F080
434. F180
435. F280
436. F380
437. F480
438. F081
439. F181
440. F281
441. F381
442. F481
443. F082
444. F182
445. F282
446. F382
447. F482
448. F083
449. F183
450. F283
451. F383
452. F483
453. F084
454. F184
455. F284
456. F384
457. F484
458. F085
459. F185
460. F285
461. F385
462. F485
463. F086
464. F186
465. F286
466. F386
467. F486
468. F087
469. F187
470. F287
471. F387
472. F487
473. F088
474. F188
475. F288
476. F388
477. F488
478. F089
479. F189
480. F289
481. F389
482. F489
483. F090
484. F190
485. F290
486. F390
487. F490
488. F091
489. F191
490. F291
491. F391
492. F491
493. F092
494. F192
495. F292
496. F392
497. F492
498. F093
499. F193
500. F293
501. F393
502. F493
503. F094
504. F194

505. F294
506. F394
507. F494
508. F095
509. F195
510. F295
511. F395
512. F495
513. F096
514. F196
515. F296
516. F396
517. F496
518. F097
519. F197
520. F297
521. F397
522. F497
523. F098
524. F198
525. F298
526. F398
527. F498
528. F099
529. F199
530. F299
531. F399
532. F499
533. F0100
534. F1100
535. F2100
536. F3100
537. F4100
538. BIRTHYEAR
539. ESCDAYS
540. WD
541. NPW
542. REFINT

D.2 French dataframe

1. SPEAKER
2. SEX
3. SESSION
4. WORD
5. PHONEME
6. DURATION
7. LOCINFILE
8. INTENSITY
9. PHONBEFORE
10. PRECOART
11. BEFVOICE
12. BEFMOA
13. BEFPOA
14. PHONAFTER
15. POSTCOART
16. AFTVOICE
17. AFTMOA
18. AFTPOA
19. EPENTHETIC
20. TOTALDUR
21. SPPASMEANF010
22. SPPASMEANF110
23. SPPASMEANF210
24. SPPASMEANF310
25. SPPASMEANF410
26. SPPASMEANF020
27. SPPASMEANF120
28. SPPASMEANF220
29. SPPASMEANF320
30. SPPASMEANF420
31. SPPASMEANF030
32. SPPASMEANF130
33. SPPASMEANF230
34. SPPASMEANF330
35. SPPASMEANF430
36. SPPASMEANF040
37. SPPASMEANF140
38. SPPASMEANF240
39. SPPASMEANF340
40. SPPASMEANF440
41. SPPASF010
42. SPPASF110
43. SPPASF210
44. SPPASF310
45. SPPASF410
46. SPPASF020
47. SPPASF120
48. SPPASF220
49. SPPASF320
50. SPPASF420
51. SPPASF030
52. SPPASF130
53. SPPASF230
54. SPPASF330
55. SPPASF430
56. SPPASF040
57. SPPASF140
58. SPPASF240
59. SPPASF340
60. SPPASF440
61. SPPASF050
62. SPPASF150
63. SPPASF250
64. SPPASF350
65. SPPASF450
66. SPPASF060
67. SPPASF160
68. SPPASF260
69. SPPASF360
70. SPPASF460
71. SPPASF070
72. SPPASF170
73. SPPASF270
74. SPPASF370
75. SPPASF470
76. SPPASF080
77. SPPASF180
78. SPPASF280
79. SPPASF380
80. SPPASF480
81. SPPASF090
82. SPPASF190
83. SPPASF290
84. SPPASF390
85. SPPASF490
86. BIRTHYEAR
87. ESCDAYS

D.3 Subfiles

D.3.1 Duration file

Header of the file containing the duration of all the manually aligned TextGrid intervals:

1. SPEAKER
2. SESSION
3. LOCINFILE
4. SMALLDURATION
5. LABEL
6. INTERVAL

D.3.2 Word file

Header of the file containing all the words pronounced by the learners:

1. SPEAKER
2. SESSION
3. WORD
4. LOCINFILE

D.3.3 PVI file

Header of the file containing the durations of all the vocalic and consonantal intervals as aligned by SPPAS or P2FA (NBPHON is the number of phonemes each interval contains):

1. SPEAKER
2. SESSION
3. ALIGNER
4. LABEL
5. DURATION
6. NBPHON
7. LOCINFILE

D.3.4 Phoneme duration file

Header of the file containing the duration of each phoneme:

1. SPEAKER
2. SESSION
3. ALIGNER
4. PHONEME
5. DURATION
6. MOA

D.4 Syllable mismatches

This section provides the list of words that featured a mismatch between the number of syllables in the LPD (listed in the LPDSC column) and that established by the aligner in the SC column. Section D.4.1 lists the mismatches in the SPPAS-aligned dataset, section D.4.2 the mismatches in the P2FA-aligned dataset.

D.4.1 SPPAS syllable mismatches

1. every
2. several
3. family
4. physically
5. globalized
6. angeles
7. us
8. fjords
9. references
10. partnership
11. passed
12. especially
13. mystery
14. interests
15. literally
16. discovery
17. julius
18. tolkien
19. eventually
20. korean
21. seventeenth
22. homeless
23. there'll
24. ladies
25. machu
26. parents'
27. gelato
28. stuntman
29. actually
30. ideas
31. toward
32. personally
33. tv
34. minivan
35. fairytale
36. travelling
37. halloween
38. interested
39. usually
40. fire
41. practically
42. inspire
43. touched
44. ira

45. monotheist
46. plosive
47. gru
48. idealized
49. radically
50. satisfied
51. tiring
52. inspiring
53. picchu
54. ireland
55. australian
56. australians
57. different
58. conference
59. devilish
60. interesting
61. probably
62. las
63. violent
64. simplest
65. bodies
66. separate
67. pyjamas
68. normally
69. medieval
70. plannning
71. national
72. actual
73. raphaelite
74. happier
75. scifi
76. theatre
77. compense
78. ok
79. familiar
80. dancing
81. britney
82. restaurant
83. eiffel
84. anti
85. applied
86. australia
87. uncomfortable
88. literature
89. junior
90. our
91. realize
92. california
93. happiest
94. listening
95. etcetera
96. gruffalo
97. families
98. phd
99. general
100. orpheus
101. realistic
102. philosophical
103. haywire
104. trying
105. generally
106. teachings
107. restaurants
108. uk
109. inspires
110. vertically
111. mcdonald's
112. pretentions
113. theater
114. favorite
115. comfortable
116. history
117. usa
118. idea
119. favourite
120. scandinavian
121. towards
122. roll
123. frightening
124. hour
125. worried
126. interest
127. vowels
128. shakespeare
129. typically
130. hours
131. korea
132. lyrical
133. picadilly
134. violet
135. trainings
136. am
137. niagara
138. opera
139. parisian
140. blurry

D.4.2 P2FA syllable mismatches

1. states
2. actually
3. restaurant
4. sales
5. feel
6. applied
7. conference
8. natural
9. while
10. history
11. trying
12. angeles
13. idea
14. fjords
15. halloween
16. happiest
17. partnership
18. passed
19. practically
20. medieval
21. inspire
22. touched
23. vowels
24. jail
25. shakespeare
26. scale
27. steal
28. smile
29. hour
30. am
31. teachings
32. restaurants
33. physically
34. parents'
35. opera
36. australians
37. theater
38. australia
39. devilish
40. i'll
41. mail
42. tv
43. minivan
44. us
45. simplest
46. style
47. towards
48. privilege

49. every
50. frightening
51. general
52. favorite
53. national
54. ireland
55. ira
56. julius
57. happier
58. wild
59. child's
60. seventeenth
61. satisfied
62. violet
63. families
64. male
65. it'll
66. australian
67. anti
68. we'll
69. toward
70. uncomfortable
71. personally
72. average
73. probably
74. las
75. fairytale
76. california
77. pale
78. family
79. interested
80. geographically
81. deals
82. foreigners
83. literally
84. desperate
85. died
86. actual
87. pre
88. philosophical
89. haywire
90. theatre
91. radically
92. listening
93. there'll
94. ladies
95. sail
96. beverage
97. eiffel
98. they'll
99. several
100. feels
101. comfortable
102. field
103. dale
104. junior
105. different
106. realize
107. scandinavian
108. violent
109. deal
110. separate
111. usually
112. tale
113. worried
114. difference
115. s
116. learned
117. fields
118. realistic
119. especially
120. korea
121. ourselves
122. traveling
123. ok
124. tiring
125. dancing
126. inspires
127. fails
128. mcdonald's
129. blurry
130. ideas
131. pure
132. child
133. literature
134. globalized
135. usa
136. interesting
137. our
138. references
139. bodies
140. normally
141. roll
142. tales
143. mystery
144. interests
145. etcetera
146. mails
147. hours
148. orpheus
149. typically
150. eventually
151. idealized
152. lyrical
153. homeless
154. rationally
155. familiar
156. fail
157. niagara
158. vertically
159. parisian

Appendix E

Extra graphs and tables

This appendix contains figures that were optional to understanding the main body of text, but whose observation may give further insight into the arguments it developed.

E.1 Extra-graphs for the French and English reading lists

E.2 Extra graphs: Onset-to-Offset Distances

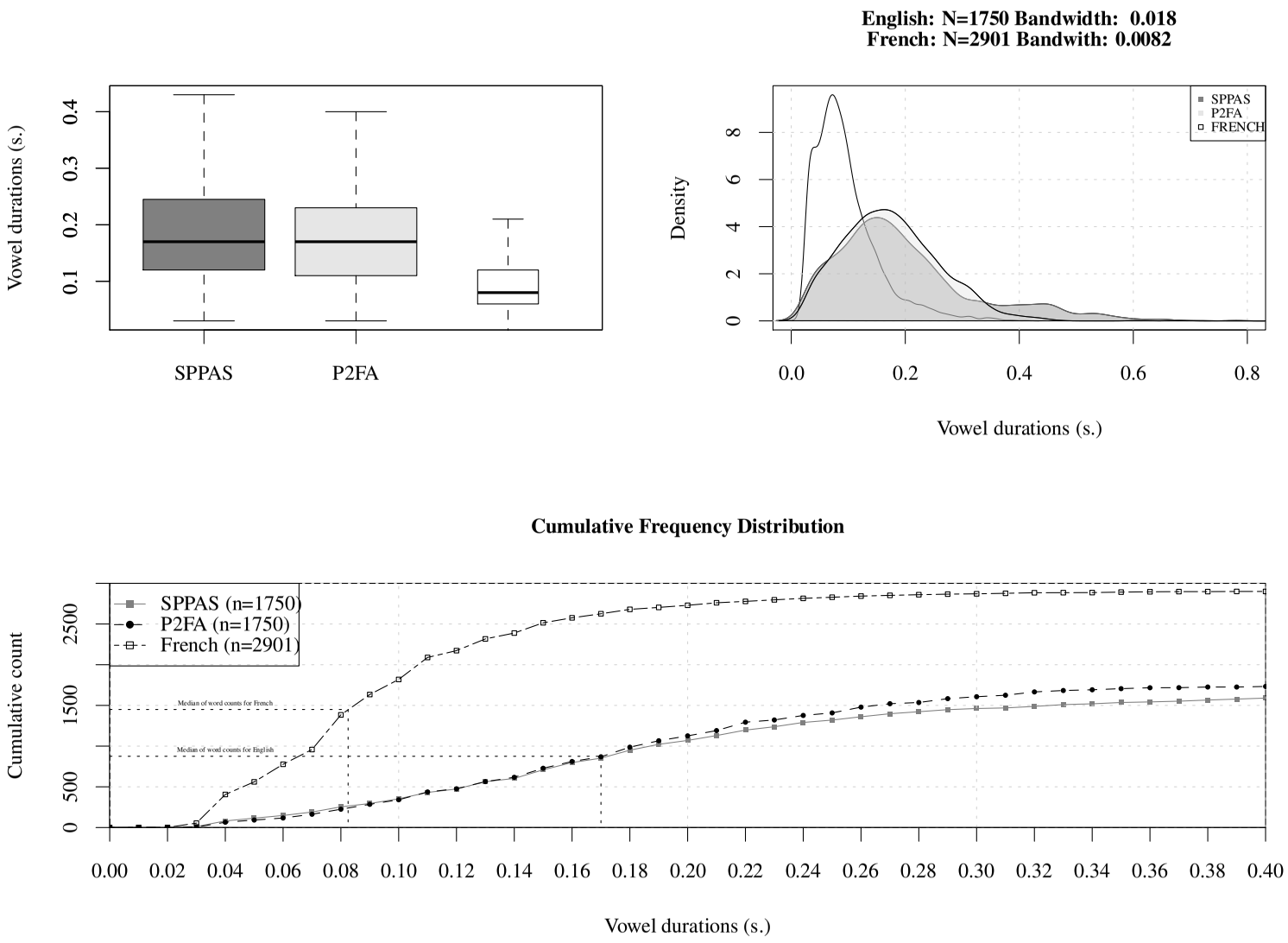


Fig. E.1: Per-aligner distribution of vowel durations in the reading tasks; medians and quartiles (left); kernel density plot (right); cumulative frequency (bottom).

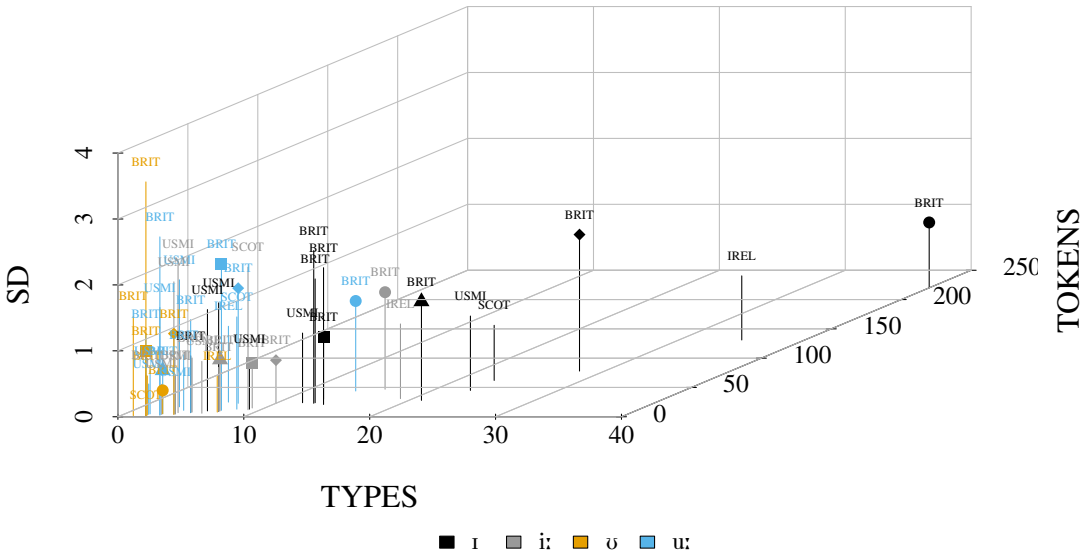


Fig. E.2: Standard deviations of the native OODs for /ɪ/, /iː/, /ʊ/ and /uː/ against syllable tokens and types.

E.3 Mean differences of OOD standard deviations

E.4 KNN: results based on the NSS

	æ	aː	e	ɜː	ə	ɪ	iː	ɒ	ɔː	ʌ	ʊ	uː
æ	43	1	12	6	16	9	1	9	2	9	0	2
aː	0	0	0	0	0	0	0	0	0	0	0	0
e	4	0	1	1	8	4	1	1	0	1	0	0
ɜː	0	0	0	0	0	0	0	0	0	0	0	0
ə	30	1	12	4	96	31	14	24	10	6	3	17
ɪ	11	0	11	0	32	115	28	4	2	2	0	9
iː	0	0	1	0	3	25	52	2	0	0	1	16
ɒ	0	6	0	0	4	2	1	18	2	2	0	0
ɔː	1	4	0	0	7	2	2	11	24	4	0	3
ʌ	0	6	3	2	3	2	0	2	1	7	0	0
ʊ	0	0	0	0	0	0	0	0	0	0	0	0
uː	0	0	0	0	0	0	2	0	0	0	0	0

Table E.1: Confusion matrix of the last pass of the KNN algorithm on the British female natives (NSS).

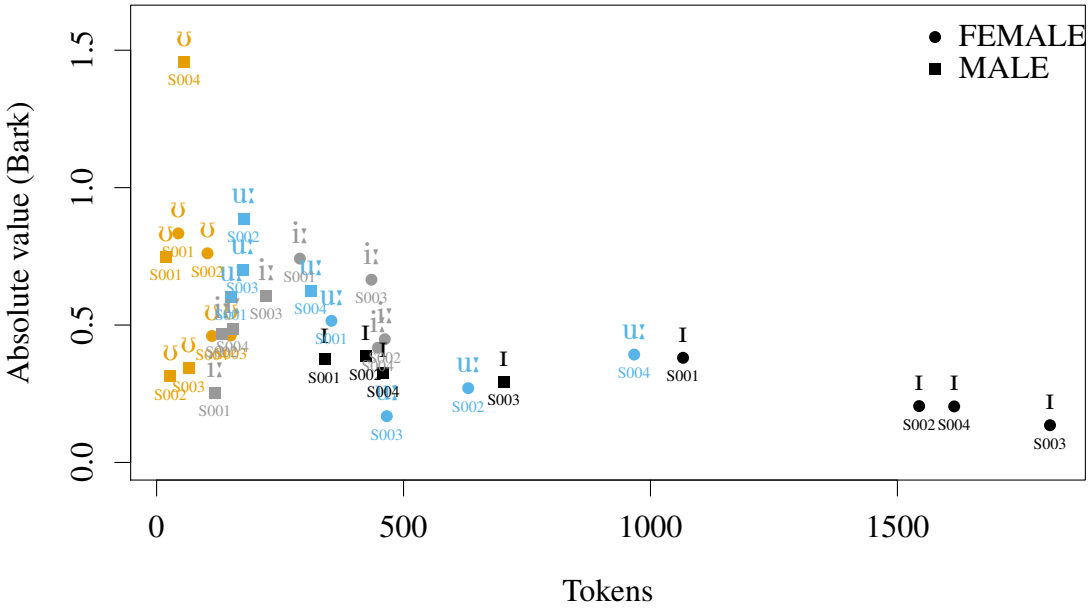


Fig. E.3: Per-session, per-gender mean absolute values of the differences between learners' and native speakers' OOD standard deviations against the number of tokens.

E.5 Corrected response variables

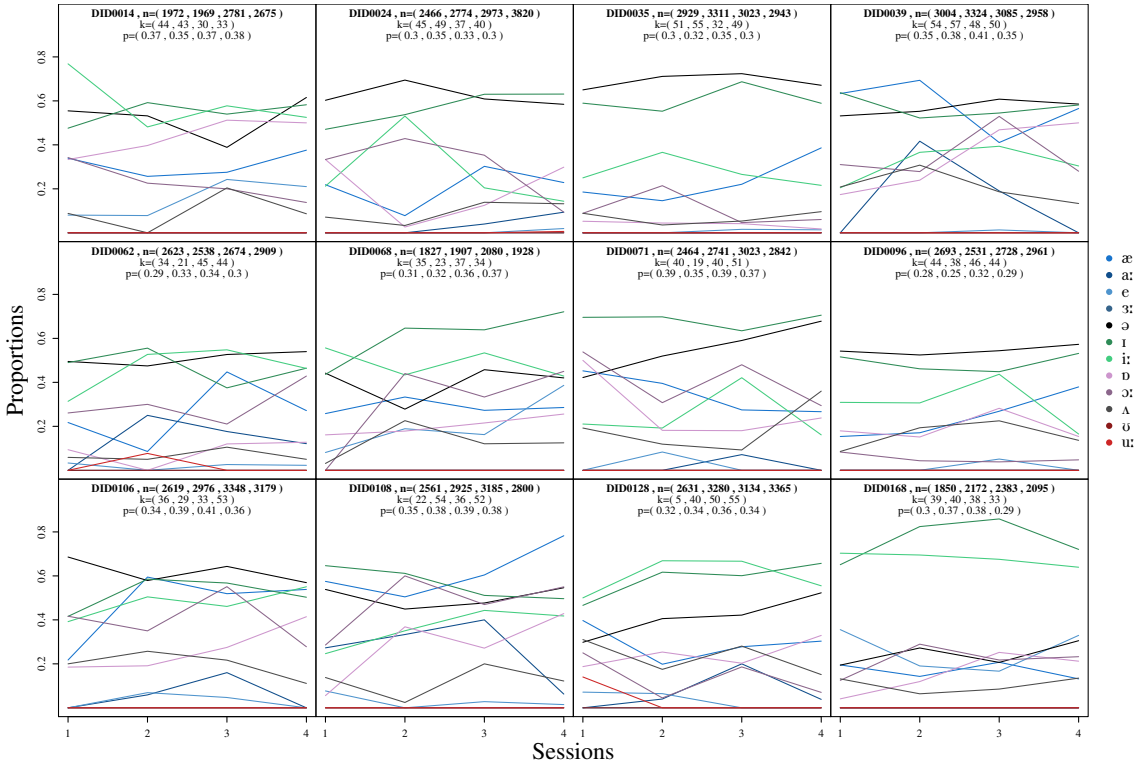


Fig. E.4: Proportion of correctly labeled phonemes in the BDM-normalized F_1 / F_2 space using the KNN classification method with the NSS as a training set. In each panel, the total number of tokens n for each session is indicated, along with the optimal k-values and the global proportion of accurately labeled phonemes.

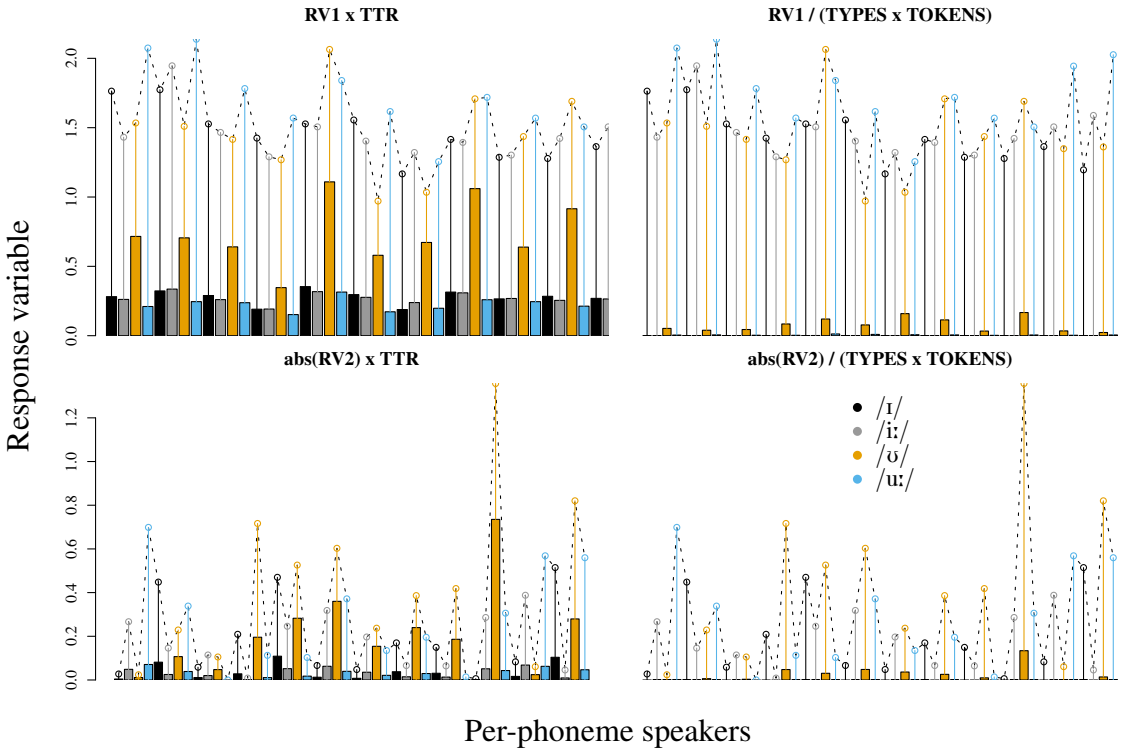


Fig. E.5: Per-speaker corrected response variables by formulas. *Top row:* RV1; *bottom row:* RV2; *left column:* the response variable is multiplied by the corresponding TTRs; *Right column:* the response variable is divided by the product of the numbers of types and tokens. The dotted line in each panel show the original response variable, the continuous vertical line materializes the distance between the original response variable and the corrected one.

E.6 Multimodel comparisons with log-transformed response variables

Table E.2: Per-phoneme, per- log-transformed response variable results of the multi-model comparisons. *AICcWt:* weight of evidence; *p-value:* p-value of the Shapiro-Wilk test carried out on the residuals of the fitted models.

Phoneme	Response Variable	Model Number	AICcWt	p-value
ɪ	RV1	M2	0.69	0.83
ɪ	RV1	M4	0.71	0.18
ɪ	RV1C	M3	0.96	0.78
ɪ	RV1C	M5	0.97	0.50
ɪ	RV2	M1	0.53	0.02
ɪ	RV2	M4	0.57	0.03
ɪ	RV2C	M3	0.54	0.00
ɪ	RV2C	M5	0.75	0.02
iː	RV1	M2	0.52	0.61
iː	RV1	M4	0.61	0.80
iː	RV1C	M3	0.88	0.06
iː	RV1C	M5	0.96	0.13
iː	RV2	M1	0.55	0.00
iː	RV2	M4	0.73	0.00
iː	RV2C	M1	0.45	0.00
iː	RV2C	M4	0.61	0.01
ʊ	RV1	M1	0.65	0.11
ʊ	RV1	M4	0.78	0.09
ʊ	RV1C	M1	0.49	0.03
ʊ	RV1C	M4	0.64	0.06
ʊ	RV2	M1	0.54	0.00
ʊ	RV2	M4	0.80	0.00
ʊ	RV2C	M2	0.44	0.00
ʊ	RV2C	M4	0.77	0.00
uː	RV1	M1	0.41	0.91
uː	RV1	M5	0.53	0.40
uː	RV1C	M3	0.74	0.24
uː	RV1C	M5	0.83	0.95
uː	RV2	M1	0.71	0.00
uː	RV2	M4	0.80	0.00
uː	RV2C	M1	0.45	0.02
uː	RV2C	M4	0.77	0.00

E.7 DCT: extra-graphs

E.7.1 Procedure to calculate intra- and inter- phoneme proportions

Figure E.6 explains the procedure in section 3.5.2, and especially how the results presented in figure 3.19 were obtained. The proportions are calculated from the points within the circle whose radius is a proportion of the maximal distance.

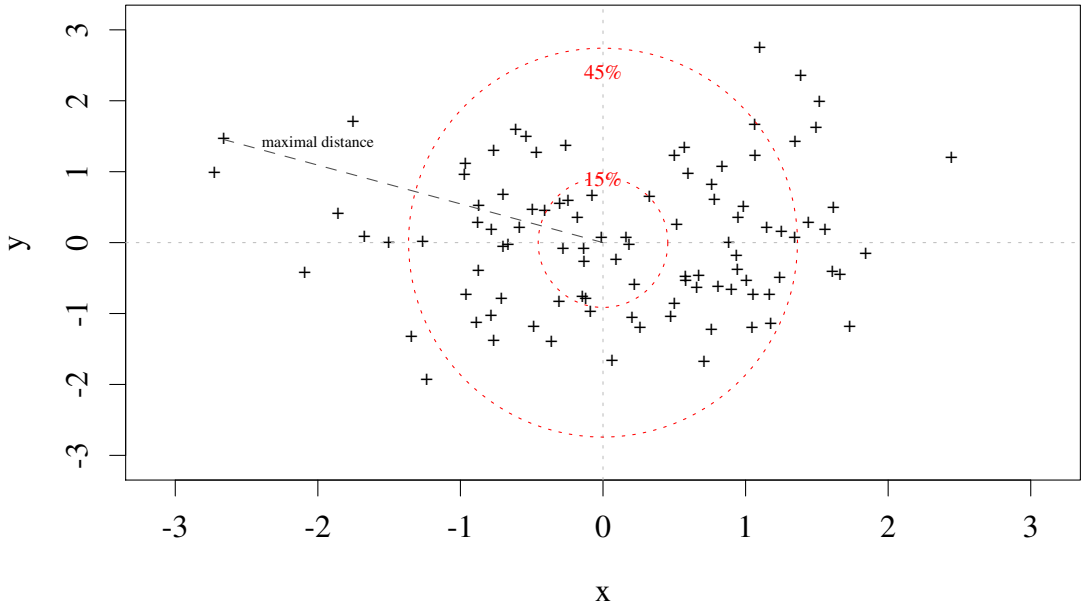


Fig. E.6: Explanations of the procedure to find the intra- and inter- phoneme proportions in section 3.5.2. The radius of the circle is a proportion of the maximal distance from the origin. Two circles have been drawn here: one with a radius amounting to 15% of the maximal distance, the other 45%.

E.7.2 Per-session, per-speaker evolution of k_2

E.7.3 QDA model results

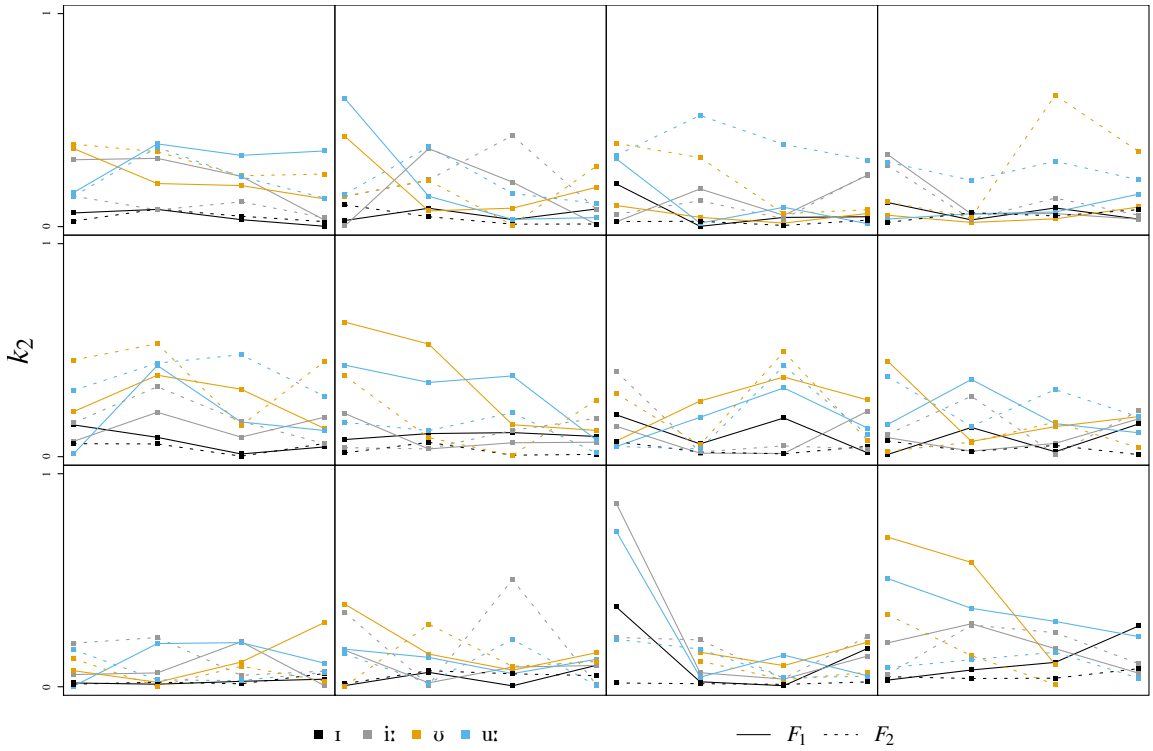


Fig. E.7: Per-session, per-speaker evolution of the absolute values of the differences from native values for k_2.

Model	ɪ	iː	ʊ	uː	Global Proportion
m_1	0.95	0.09	0.09	0.10	0.59
m_2	0.96	0.16	0.10	0.20	0.63
m_3	0.94	0.19	0.12	0.07	0.59
m_4	0.96	0.15	0.12	0.19	0.63
m_5	0.95	0.02	0.00	0.12	0.58
m_6	0.95	0.12	0.00	0.14	0.60
m_7	0.94	0.09	0.00	0.02	0.56
m_8	0.94	0.16	0.00	0.09	0.59
m_9	0.88	0.27	0.10	0.17	0.59
m_{10}	0.93	0.20	0.11	0.25	0.63
m_{11}	0.88	0.24	0.10	0.19	0.60
m_{12}	0.92	0.22	0.11	0.23	0.62

Table E.3: Mean of the per-phoneme, per-model results for each sex of the QDAs

Appendix F

List of papers and conferences

Publications

- Méli, A. & Ballier, N. (2015). Assessing L2 phonemic acquisition: a normalization-independent method? In *Proceedings of the 18th International Congress of Phonetic Sciences.* (pp. 805-810) Glasgow, August 10 - 14 2015. Link: ipa.org
- Ballier, N. Méli, A. (2015). CV-patterned transfers among French speakers of English. In *Proceedings of EPIP4, 4th International Conference on English Pronunciation: Issues Practices.* (pp. 14–17) Prague, May 21 - 23, 2015. Link: cuni.cz
- Méli, A. (2013). Phonological acquisition in the French-English interlanguage: rising above the phoneme. In A. Ballier, N. & Díaz-Negrillo, & P. Thompson (Eds.) *Automatic Treatment and Analysis of Learner Corpus Data*, (pp. 207–226). Amsterdam: John Benjamins.

Communications

- *Analysing the emergence of vowel categorisation in a longitudinal learner corpus: the kernel estimate method.* Ballier, N. Méli, A. EPIP5, Caen, France, 17 - 19 May 2017.

- *Challenging the lexical set approach with classifiers for the investigation of the interphonology of /i/ vs. /ɪ/.* Méli, A. N. Ballier. Phonologie de l'anglais contemporain, Aix-en-Provence, France, 29 Septembre 2016.
- *Learner phonetic variability and the lexicon: a pilot study for two phonemic contrasts.* N. Ballier Méli, A. Third Learner Corpus Research Conference, Nijmegen, The Netherlands, September 11 - 13, 2015.
- *Assessing phonemic acquisition - Phone-gating: a normalization-dependent procedure?* Méli, A. Workshop on Phonetic Learner Corpora Satellite workshop of ICPhS, Glasgow, UK, August 12, 2015.
- *CV-patterned transfers among French speakers of English.* N. Ballier Méli, A. 4th International Conference on English Pronunciation: Issues & Practices, Prague, Czech Republic, May 21-23, 2015.
- *Vowel acquisition in the French-English interphonology.* Méli, A. Phonologie de l'Anglais Contemporain, Toulouse, April 13, 2015.
- *Investigating interlanguage stages: Vowel phonemic distinctions among French speakers of English* Méli, A. N. Ballier. 47th Annual Meeting of the Societas Linguistica Europaea, Adam Mickiewicz University, Poznań, Poland - 11 – 14 September 2014.
- *Designing an EFL learner corpus to analyse phonetic and phonological variation* Ballier, N. Méli, A. ICAME, Giessen, Germany, 26 - 30 May 2010

Appendix G

Résumé en français

Ce travail entreprend d'évaluer l'évolution de l'acquisition phonologique par des étudiants français des contrastes anglais /I/-/i:/ et /U/-/u:/. Le corpus étudié provient d'enregistrements de conversations menées avec des étudiants natifs autour de tâches préalablement définies dans le cadre du projet LONGDALE (Goutereaux 2013) entre l'université Paris Diderot et l'université catholique de Louvain. 12 étudiants, 9 femmes et 3 hommes, ont été suivis lors de 4 sessions espacées chacune d'un intervalle de six mois. L'approche adoptée est résolument quantitative, et agnostique quant aux théories d'acquisition d'une deuxième langue (par exemple Flege 2005, Best 1995, Kuhl 2008). Celles-ci prédisent toutes des difficultés identiques pour apprendre à prononcer ces deux contrastes, en raison de la symétrie entre langue source (le français) et langue cible (l'anglais) : au son français /i/ correspondent les sons similaires anglais /I/ et /i:/, de la même manière qu'au son français /u/ correspondent les sons anglais /U/ et /u:/.

Le premier chapitre de la thèse s'attache à décrire les méthodes déployées afin de collecter les données. Des analyses préliminaires indépendantes des qualités vocaliques sont aussi effectuées. Afin d'estimer les éventuels changements de prononciation des voyelles de ces deux contrastes par les étudiants français, une procédure automatique d'alignement et d'extraction des données acoustiques a été conçue à partir du logiciel PRAAT (Boersma

2001). Dans un premier temps, deux autres logiciels, le SPeech Phonetization Alignment and Syllabification (SPPAS, Bigi (2012b), Bigi & Hirst (2012)), et le Penn Phonetics Lab Forced Aligner Toolkit (P2FA, Yuan & Liberman (2008)) avaient aligné les transcriptions des enregistrements au phonème près. Le script PRAAT écrit pour cette étude fit ensuite les choses suivantes :

- récupérer la prononciation de chaque mot dans le dictionnaire Longman Pronunciation Dictionary (Wells 2008) ;
- créer pour chaque aligneur des niveaux dans les fichiers d'alignement correspondant aux découpages syllabiques de chaque mot; ce découpage syllabique a été effectué selon les prononcitions établies dans le *Longman Pronunciation Dictionary* ;
- recueillir pour chaque voyelle, dans les intervalles alignés par chacun des deux aligneurs, un ensemble qui se voulait exhaustif de données permettant de procéder aux analyses acoustiques.

Ces données sont constituées d'informations telles que le nombre de syllabes du mot, de la transcription acoustique du dictionnaires, des phonèmes suivant et précédant la voyelle, de leur lieu et manière d'articulation, de leur appartenance ou non au même mot, mais surtout des relevés formantiques de F0, F1, F2, F3 et F4. Ces relevés formantiques ont été effectués à chaque pourcentage de la durée de la voyelle afin de pouvoir tenir compte des influences des environnements consonantiques sur ces formants. Par ailleurs, des théories telles que le changement spectral inhérent aux voyelles (Nearey & Assmann (1986), Morrison & Nearey (2006), Hillenbrand (2012), Morrison (2012)), ou des méthodes de modélisation du signal telles que la transformation cosinoïdale discrète (Harrington 2010) requièrent que soient relevées les valeurs formantiques des voyelles tout au long de leur durée. À partir des alignement générés par les deux aligneurs SPPAS et P2FA, des fichiers d'alignement `PRAAT` de type "`TextGrid`" comportant des intervalles ajustés à chaque phonème, chaque mot, chaque syllabe française, chaque syllabe anglaise et chaque groupe consonantique ou

vocalique ont été créés pour chacun des 81 enregistrements disponibles dans le corpus (*c.f.* figure 1.8 dans le chapitre 1). Les mêmes `TextGrids` ont été générés pour trois corpus de taille inférieure : deux groupes d'enregistrements, l'un de listes de mots, l'autre d'un texte lu en français, ont ainsi été traités. Un corpus de conversations spontanées de locuteurs natifs a aussi été constitué indépendamment, et a suivi le même traitement. Au total, ce sont donc 120 fichiers d'alignement qui ont été générés. Une ambition certaine d'exhaustivité a dominé la collecte d'information pour chaque voyelle : outre sa catégorie phonémique, 541 informations supplémentaires (86 pour le corpus du texte lu en français) ont été récupérées. Ce total doit qui plus est être multiplié par deux, puisqu'un tableau de données par aligneur a été généré (sauf encore une fois pour le corpus du texte lu en français, celui-ci n'ayant été aligné automatiquement que par SPPAS). Les informations extraites pour chaque voyelle sont aussi bien extra-linguistiques, portant sur le locuteur, la session ou le nombre de jours passés dans un pays de lanque anglaise, que linguistiques : ont ainsi été collectées des informations telles que le mot et la syllabe dans lesquels la voyelle apparaît, les différentes transcriptions (celles du dictionnaire propre à chaque aligneur d'un côté et celle du *Longman Pronunciation Dictionary* de l'autre), l'accentuation de la syllabe, la structure syllabique, les phonèmes précédant et suivant la voyelle, leur lieu et manière d'articulation. À ces informations doivent s'ajouter les données purement acoustiques, telles que les valeurs formantiques, récupérées à chaque pourcentage de la durée de la voyelle, sa durée, son intensité... Au total, si l'on additionne les voyelles dont les données ont été récupérées à partir des intervalles créés dans les fichiers d'alignement par les deux aligneurs (c'est-à-dire en prenant en compte des voyelles dont l'alignement diffère d'un aligneur à l'autre), 199 950 voyelles ont été extraites sur l'ensemble des corpus. Le nombre de cellules disponibles dans les tableaux de données générés s'élève à 107 052 945. La deuxième partie du chapitre 1 s'attache à étudier les erreurs d'extractions pour chaque formant et chaque centile, indépendamment de la catégorie phonologique de chaque voyelle ; les variations, explicables ou non, dans

l'étiquetage phonologique opéré par les aligneurs pour les voyelles de mots fréquents ; les erreurs de syllabification dues aux décalages entre les transcriptions des dictionnaires, ainsi que des analyses détaillées de la durée des voyelles et des débits d'élocution.

Après ces analyses préliminaires, qui ne tenaient pas compte des catégories phonologiques, le chapitre 2 entreprend de décrire les spécificités de chaque catégorie phonologique, tous locuteurs confondus. Seules les monophtongues sont étudiées. Avant de procéder aux analyses acoustiques, une évaluation de la qualité des extractions acoustiques est effectuée. Il est démontré qu'une majorité des valeurs formantiques de chaque monophtongue sur chaque pourcentage de leur durée est comprise entre des intervalles (en Hertz) raisonnables et réalistes. De cette étude le conclusion est tirée que l'extraction automatique opérée selon la méthode décrite dans le chapitre précédent est bien fiable. La répartition des voyelles dans le trapèze vocalique est ensuite étudiée, et comparée aux valeurs natives, ainsi que la variété lexicale atenante à chaque monophtongue. Les proportions écrasantes de mots grammaticaux pour certaines catégories a été notée, une caractéristique que la recherche ultérieure devra prendre en compte d'une manière plus subtile que dans cette étude, l'approche adoptée n'ayant guère inclus cette varitété lexicale. La dispersion des valeurs formantiques F_1, F_2 et F_3 pour chaque pourcentage de la durée de la voyelle est ensuite analysée, et mène à la découverte de dispersions supérieures pour les phonèmes /ʊ/ et /uː/. Une procédure exploratoire est mise en place afin de récupérer pour chaque formant et chaque voyelle le centile dont les valeurs sont les moins dispersées. Ces décalages de dispersion entre les différentes catégories, combinées à ceux entre les nombres d'occurrences respectifs, mènent à se poser la question de la façon de traiter les données acoustiques. Les méthodes de normalisation des valeurs formantiques habituellement préconisées, telles que la méthode de Lobanov (Lobanov (1971)), sont généralement utilisées dans des corpus comportant des effectifs égaux d'occurrences de chaque phonème. Dans le cadre d'analyses de conversations spontanées, présentant par définition des effectifs inégaux, voire déséquilibrés, ce genre de

méthode est inadapté. Ces déséquilibres étant similaires dans le corpus d'apprenants et dans le corpus de natifs, il est préconisé d'utiliser une méthode de normalisation intrinsèque aux phonèmes, plutôt qu'extrinsèque. La validité de cette suggestion est démontrée en comparant les différentes méthodes et en les appliquant à un corpus aux effectifs rigoureusement égaux, le corpus de Peterson & Barney (Peterson & Barney (1952)). Deux méthodes intrinsèques sont ensuite comparées, la méthode Bark (Traunmüller (1990)) et la métrique de différence en Bark (Syrdal & Gopal (1986)). Cette dernière est finalement recommandée, parce qu'elle permet d'intégrer davantage d'information, notamment la F_3, et réduit à deux dimensions des données normalement tridimensionnelles. Le chapitre s'achève sur une étude des relations entre la longueur, dans l'espace vocalique normalisé, des contrastes, à savoir /ɪ/-/iː/ et /ʊ/-/uː/, et de la surface du polygone vocalique reliant les monophthongues entre elles. Cette relation, mesurée par le quotient entre la distance du contraste et la surface du polygone vocalique, a été établie pour tous les corpus (d'apprenants, de natifs, de listes de mots, et celui de Peterson & Barney). La plus grande cohérence des mesures dans le cas de /ɪ/-/iː/ que dans celui de /ʊ/-/uː/ semble indiquer une meilleure conception des cibles articulatoires à atteindre. Il reste toutefois à établir dans quelle mesure cette meilleure conception des cibles articulatoires constitue une preuve d'une meilleure acquisition phonologique des contrastes.

Le chapitre 2 ayant mis à jour des similitudes, indéepndantes des locuteurs, dans les taux de dispersion de /ʊ/-/uː/, il était temps de prendre en compte les spécificités de l'évolution de l'acquisition phonologique des deux contrastes chez chaque apprenant. La première préoccupation du chapitre 3 est de comparer les changements spectraux inhérents à chaque occurrence vocalique, pour chaque apprenant dans chaque session, aux changements spectraux des locuteurs natifs. L'analyse de ces changements spectraux, normalisés en conformité avec les recommendations du chapitre précédent, permet de prendre en compte des valeurs formantiques autres que celles à la moitié de la durée de la voyelle : les points de départ et d'arrivée des changements spectraux correspondent aux valeurs prises à 20% et 80% de la

durée. Les longueurs de chaque vecteur ainsi obtenu sont ensuite comparées aux longueurs chez les locuteurs natifs, une plus grande attention étant portée aux quatre phonèmes /ɪ/, /iː/, /ʊ/ et /uː/. Il apparaît à ce stade qu'une plus grande cohérence dans les longueurs existe dans le cas de /ɪ/-/iː/ que dans celui de /ʊ/-/uː/, indépendamment de leur localisation dans l'espace vocalique normalisé. De tels résultats constituent une preuve notable de l'absence de similarité entre les acquisitions phonologiques des deux contrastes. Cette conclusion provisoire est corroborée par l'étude des dispersions des changements spectraux, où il est établi que que les dispersions de /ʊ/-/uː/ sont encore une fois supérieures à leurs homologues /ɪ/-/iː/, en dépit d'un nombre d'occurrences bien inférieur. Ces résultats peuvent-ils être confirmés par des algorithmes de classification ? Si certaines cibles phonémiques sont mieux acquises que d'autres, il semble raisonnable de supposer que ces catégories seront mieux reconnues par de tels algorithmes.

La méthode de classification choisie est celle des k plus proches voisins. Un dispositif expérimental spécifique, utilisé plus tard pour les analyses quadratiques discriminantes, a été conçu de la façon suivante : au lieu de découper les données en échantillons aléatoires servant tour à tour d'ensembles d'entraînement et de test, ce sont les valeurs des locuteurs natifs qui servent d'ensemble d'entraînement. Les variables étudiées sont les valeurs formantiques normalisées F_1 et F_2 prises à la moitié de la durée de chaque voyelle. Afin de contrôler les influences potentielles des environnements consonantiques, seules les occurrences de monophtongues apparaissant dans des structures syllabiques existant dans le corpus natif furent retenues. En raison du mode calculatoire de la méthode des k pus proches voisins, la sélection du k optimal s'est effectuée en appliquant l'algorithme 1 000 fois à tous les phones de chaque locuteur à chaque session. Chaque passe faisait varier k de 1 à $\sqrt{n}$, où n représente le nombre total de phones de la session. En raison de la répartition inégale des occurrences de chaque catégorie phonologique, les valeurs natives choisies pour constituer l'ensemble d'entraînement sont celles du corpus de Peterson & Barney. Au premier abord,

rien dans les résultats n'indique de véritables différences dans les taux de classification des quatre phonèmes étudiés. L'étude des meilleures solutions alternatives, c'est-à-dire des prédictions phonémiques arrivant au deuxième rang des prédictions, révèle toutefois de grandes incohérences phonologiques dans le cas de /ʊ/ et /uː/, ces dernières étant souvent prédites comme étant des voyelles frontales. Plus crucialement peut-être, ces prédictions alternatives présentent des pourcentages d'identification supérieurs à ceux d'une identification correcte. De tels résultats semblent renforcer l'idée que le contraste /ɪ/-/iː/ est mieux acquis que /ʊ/-/uː/, mais à ce stade, aucune analyse véritablement longitudinale n'a été effectuée.

Afin d'établir l'existence d'un tel effet, une expérience est alors menée avec les régressions linéaires à effets mixtes. Plusieurs modèles sont comparés, les effets temporels étant tour à tour soit inexistants (l'évolution étant alors modélisée par une droite de pente 0), soit augmentant ou décroissant de façon linéaire, soit évoluant à la manière d'une parabole. Plusieurs variables de réponse furent étudiées, prenant en compte la plupart du temps la distance séparant les valeurs normalisées de F_1 et F_2 chez les apprenants de celles des natifs, ou bien les écarts-types de ces distances. Bien que ces résulats doivent être interpétés avec la plus grande prudence, il en ressort qu'une évolution des valeurs vers les valeurs natives est plus cohérente dans le cas de /ɪ/ et /iː/ que dans /ʊ/ et /uː/. Finalement, une analyse est conduite qui s'efforce de prendre en compte l'intégralité du signal, c'est-à-dire les cent relevés formantiques effectués tout au long de la durée de la voyelle. Pour ce faire, une modélisation du signal devait être effectuée, afin de réduire le nombre de variables à inclure dans les calculs. La méthode retenue, qui suit Harrington (2010), est celle des transformations cosinusoïdales discrètes, qui, appliqués aux relevés formantiques normalisés, permet de réduire le nombre de paramètres pour une occurrence d'un phonème de 300 à 6. Les comparisons, ici aussi, ont été menées avec les valeurs natives, après avoir restreint les occurrences à celles de voyelles apparaissant dans des environnements consonantiques communs aux deux corpus. À l'aide d'une procédure permettant de préserver les répartitions inégales des occurrences de

chaque catégorie, /uː/ se révèle être systématiquement sous-représenté dans les proportions attendues de phones similaires aux valeurs natives. Une conclusion provisoire est alors encore que le contraste /ɪ/-/iː/ est mieux acquis que /ʊ/-/uː/. Ces résultats auraient-ils été différents si les valeurs formantiques à la moitié de la durée de la voèyelle avaient été choisies ? Quels avantages les transformations cosinusoïdales, beaucoup plus exigentes en termes de programmation et de calcul, présent-elles face aux valeurs formantiques classiques ? Afin de répondre à ces questions cruciales, une comparaison est effectuée entre des modèles utilisant tour à tour le signal transformé, les formants à mi-durée, ainsi que la durée de la voyelle. Les similarités avec les valeurs natives furent établies en recourant à des analyses quadratiques discriminantes. La nécessité d'inclure, afin d'obtenir de plus hauts taux d'identification, la durée des voyelles fut établie de façon très robuste sur l'ensemble des modèles étudiés. Bien que l'efficacité et la simplicité des modèles utilisant des valeurs formantiques prises à la moitié de la durée de la voyelle, les modèles fondés sur les signaux modélisés permettent une meilleure reconnaissance des catégories présentant un nombre d'occurrences peu élevé.

Cette étude recommande finalement vivement d'étudier davantage les corpus de conversations spontanées, en dépit de leur complexité. Il est aussi préconisé de maximiser la quantité d'information traitée, et une méthode de normalisation intrinsèque telle que la métrique de différence en Bark, combinée à des transformations cosinusoïdales discrètes, permet de réduire considérablement le nombre de paramètres à prendre en compte tout en préservant autant que possible les données originellement présentes. L'application de ces procédures semble révéler des différences d'acquisition phonologique des contrastes /ɪ/-/iː/ et /ʊ/-/uː/.