ASSESSING L2 PHONEMIC ACQUISITION: A NORMALIZATION-INDEPENDENT METHOD?

Adrien MÉLI, NICOLAS BALLIER

Univ Paris Diderot – Sorbonne Paris Cité
CLILLAC-ARP – EA 3967
adrienmeli@gmail.com | nicolas.ballier@univ-paris-diderot.fr

ABSTRACT

This paper aims at addressing the issues that emerge when attempting to analyse the acquisition of L2 vowels: studies usually embed vowels in the same /hVd/ template (e.g. Hillenbrand [14], Ferragne [10], Clopper [9]) in order to reduce coarticulatory effects, and the number of occurrences of each vowel is carefully controlled in order to comply with normalization constraints (e.g. Lobanov [17]). However, such methods make it difficult to test the predictions of SLA models that base phonemic acquisition on phonemic parameters only (Flege [11], Best [2],[3]). This study investigates the development of advanced French learners for the acquisition of phonemic contrasts (/ɪ-iː/ and /ʊ-uː/) in the longitudinal DIDEROT LONGDALE [13] corpus. 15 speakers (12 female & 3 male) were recorded in spontaneous interviews over a period of two years. To test whether the acquisitions of the /ɪ-iː/ and /ʊ-uː/ constrasts are similar, a metric, the Ratio of the (contrast) DIstance to the vowel space Convex Hull (RaDiCHull, /ˌræd.ɪk.ˈhʌl/), is explored with different normalizing procedures to measure learner input as compared to native speakers of English (the reference points are from Hillenbrand [14] for values in Hz, and from Clopper [9] for values in Bark).

Keywords: Second Language Acquisition, interphonology, phonemes, contrasts, learner data modelling.

1. THEORETICAL BACKGROUND

Models in Second Language Acquisition (SLA) 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's [16] Native Language Magnet Theory expanded, or Flege's [11] Speech Learning Model, Major's [18] Ontogeny and Philogeny Model or Best's [2] Perceptual Assimilation Model.

However, outside the field of SLA, formalizations of inter-level interactions exist: for instance, exemplar theories (Pierrehumbert [19], Bybee [6], Bybee [7]) relate phonemic pronunciation to frequency of use; prosodic positions have been shown to influence the realization of phonemes (Fougeron [15]); syllabic structure and places of articulation have been shown to be connected (Tabain [20]); phonemic processing and speech-errors likewise depend upon phonological neighbourhood density and clustering coefficients (the similarities between phonological neighbours, Chan [8]).

The present study aims at exploring a way to process the data which would make it possible to assess the potential influences of extra-phonemic parameters on phonological acquisition. The evolution of the pronunciation of the /ɪ-iː/ and /ʊ-uː/ English contrasts in French learners is investigated: if a difference in the evolution of these two contrasts is observed, then extra-phonemic parameters are bound to play a role, since French only has one sound – /i/ and /u/ respectively – for each of these contrasts.

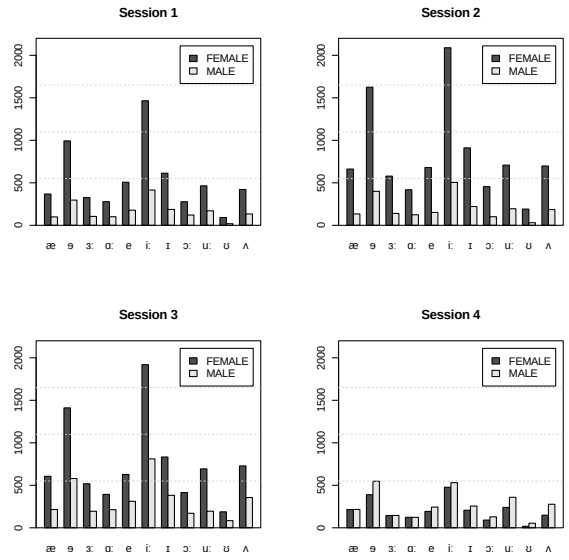
However, reliably observing whether such a difference exists does not go without challenges. One first obstacle lies in the constraints that accurate phonetic analyses impose: most studies (e.g. Hillenbrand [14], Ferragne [10], Clopper [9]) embed vowels in the /hVd/ template, to control and reduce coarticulatory effects – but this precludes the exploration of syllable effects on phonemic realizations. Resorting to normalization methods such as Lobanov's (Lobanov [17], also called the 'z-score' method) also requires the collection of a homogeneous number of occurrences for each phoneme – a feature that may overlook frequency effects. Finally, with dispersion arguably being a defining feature of both learners' phonetic realizations and automatic extraction, a reliable method to measure and represent actual pronunciation must be devised.

This paper presents a method attempting to solve the above-mentioned issues and tentatively suggests a difference in the acquisition of the two English contrasts.

2. CORPUS AND METHODOLOGY

2.1. Corpus

Figure 1: Per-sex number of US monophthongs across the four sessions.



The data used for this investigation was collected for the DIDEROT LONGDALE [13], and are distributed in the following fashion:

- 10 female students were recorded three times over six-month intervals;
- 3 male students and 2 other female students were recorded four times over six-month intervals.

50 recordings, lasting 6'45" on average, were thus obtained. 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. In session 1 and 2, the student had to describe a country, or a film/play or 'an experience which has taught you an important lesson', replicating the LINDSEI protocol (Brand et al. 2006). In session 3, four paintings were to be described. In session 4, a map task as designed by Anderson [1] aimed at eliciting questions from the learner. All 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.

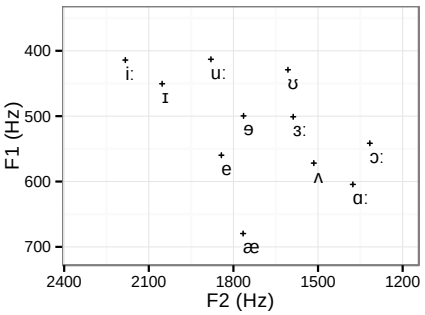
2.2. Methodology

The recordings were analyzed in the following fashion: the transcriptions of short, consistent sentences were aligned on a PRAAT TextGrid (Boersma [5]), which were then extracted and automatically aligned at the segmental level with SPPAS (Bigi [4]) using an American transcription of the CMU dictionary. Some of the labels produced by SPPAS for targets of the expected realisations were altered to accommodate for reduced vowels (e.g. "and", transcribed /ænd/ in the CMU dictionary, was changed to /ənd/). For each vowel, a PRAAT script then collected information such as the structure and stress of the syllable, the values of F0, F1, F2, F3 & F4, the preceding and succeeding phonemes, their places and manners of articulations.

45,260 phonemes, 35,504 monophthongs and 10,206 diphthongs were thus collected. The formant analyses below were all carried out from the mid-temporal values of the vowels. Fig. 1 shows the distribution of these monophthongs across the sessions.

2.3. Accuracy of extraction

Figure 2: Unnormalized mean formant values for US-labeled monophthongs across all four sessions



One central issue of automatic alignment and extraction is its accuracy. Two major obstacles were to be overcome in this study: *(i)* the length of the recordings; *(ii)* the non-native character of the speech to be aligned. The first issue was solved by manual fine-grained alignments in TextGrids: short sequences of consistent speech were aligned with their transcription on intervals spanning a maximum duration of 7 seconds. The second issue can only be tentatively worked around. Assumed acoustic targets were identified from the transcriptions, with SPPAS providing their pronunciations from the built-in Carnegie Melon University American dictionary. However, pronunciation errors being part of the study, *de facto* excluding outliers was here not

deemed acceptable. This came at the cost of teasing apart errors due to pronunciation from those caused by the automatic extraction of the data.

Figure 3: Standard Deviations for each monophthong across the four sessions.

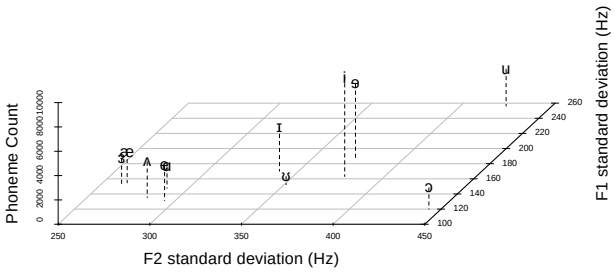


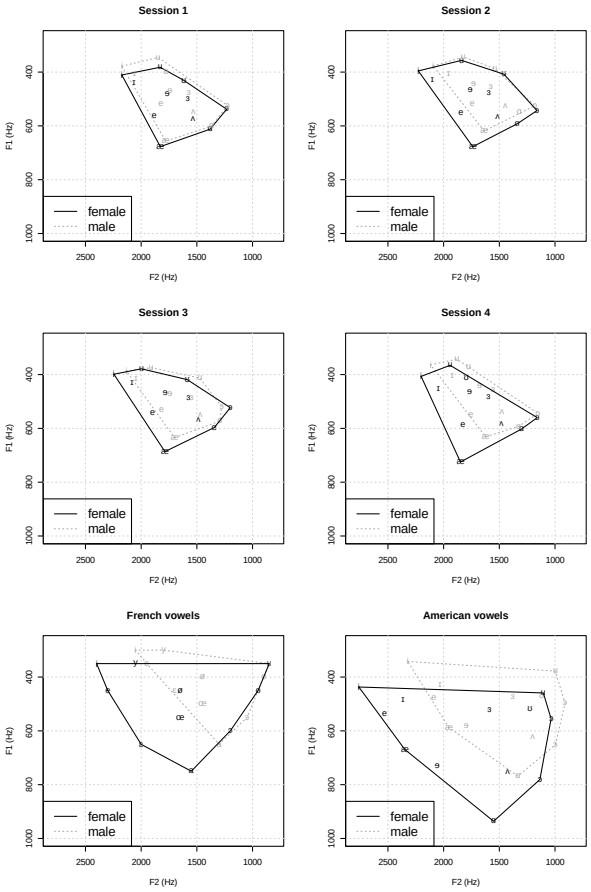
Figure 2 shows the raw data for monophthongs. The consistency of the obtained vowel space may give an indication of the accuracy of the extraction procedure. The corresponding standard deviations are represented in figure 3. No correlation was found between phoneme count and the standard deviations of either formant frequency (For F1: $R^2 = 0.0649$ with $F = 0.625, p = 0.4496$. For F2: $R^2 = 0.0622$ with $F = 0.596, p = 0.4597$.), and the four phonemes under scrutiny /ɪ/, /iː/, /ʊ/ and /uː/, all feature comparatively high standard deviations – an indication of the instability of their acquisition.

3. ANALYSIS

In order to assess the acquisition of the two contrasts, a method independent from the dictionary-based labeling and from normalization procedures was needed. In this study, the Euclidian distances from /ɪ/ to /iː/ and from /ʊ/ to /uː/ were divided by the area of the entire vowel space. The ratio, here called the 'RaDiCHull', was then compared to that of native values. What we call the 'area of the vowel space' is the area of the convex hull formed by the subset of vowels lying on the convex hull of all monophthongs. This area was calculated by the triangle method. The calculations were carried out with unnormalized interquartile values (Section 3.1) and with Lobanov-normalized values (Section 3.2). For unnormalized formants, reference values for native speakers were taken from Hillenbrand [14] for American and Gendrot [12] for French. Z-score values were taken from Clopper [9].

3.1. Unnormalized interquartile values

Figure 4: Unnormalized mean interquartile formant values for US-labeled monophthongs in each session – the convex hull visualizes the maximally used vocal tract space for vowels.



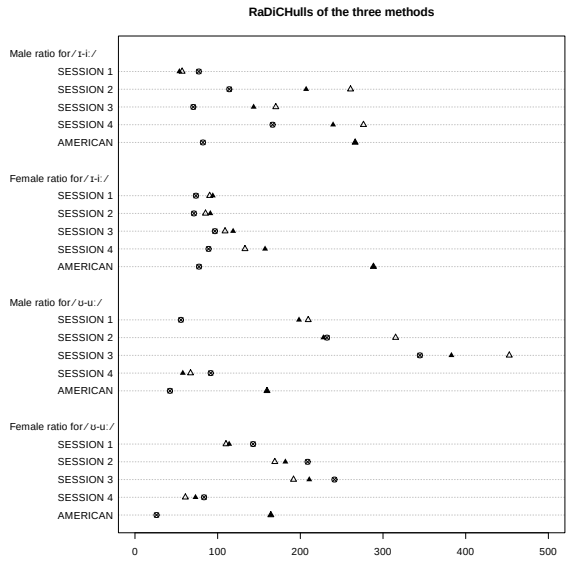
The procedure to obtain these values consisted in selecting the vowels whose F1 *and* F2 values were both superior to the first quantile of their category, and inferior to the third quantile, for each gender and for each session. Representations of the vowel spaces thus obtained are shown in figure 4 and comparisons of the obtained ratios can be found in figure 5[1]. Assuming these interquartile representations are accurate, there seems to be a clear difference in the acquisition of the two contrasts. Learners' ratios are very close to native vaues in the case of /ɪ/-/iː/, which could indicate better acquisition, whereas /ʊ/-/uː/ ratios are not only greater than native values (for a comparatively smaller vowel space), but also more dispersed.

3.2. Z-score values

Lobanov's [17] normalization method requires the same number of tokens for each phoneme. In the

case of unevenly distributed corpora such as ours, a simple random sampling of the same number of each vowel would yield robust results. However, in practice, that number must correspond to the lowest number of occurrences, which in this study turns out to be very low: /ʊ/ occurred 18 times in male speakers's fourth session. Such a low number precludes statistical robustness: no convergence of the RaDiCHulls was observed after normalizing our data set 100 times with the Lobanov method from the random sampling of 18 vowels in each category : the sample is too low to capture any statistical trend.

Figure 5: Per-sex, per-session ratio of the contrast distance to the vowel space convex hull (RaDiCHull) for /ɪ/-/iː/ and /ʊ/-/uː/. - Empty triangles : z-score normalizing method over the entire data set (Bark^{-1}). Reference values for American English are taken from Clopper [9]. - Black triangles: z-score by sex and session with unevenly distributed numbers of occurrences (Bark^{-1}). Reference values for American English are taken from Clopper [9]. - Crossed circles: RaDiCHulls for interquartile values (Hz^{-1}). Reference values for American English are taken from Hillenbrand [14].



Considering that the z-score method embeds dispersion in its formula, and considering the impossibility to obtain the same number of occurrences in each phoneme in spontaneous speech, we tried normalizing the formants in the two following manners: *(i)* we first normalized them over the entire data, and then proceeded to classify by sex and session. *(ii)* we normalized the formants by sex and session, but using all the available phonemes and therefore disregarding the requirement to have the same number. The obtained RaDiCHulls are shown in Figure 5.

4. DISCUSSION

The computation of RaDiCHulls aims at providing a method of assessing the acquisition of contrasts regardless of the normalization methods. The results obtained are summarized in Table 1, where Lobanov 1 refers to normalization over the entire dataset, and Lobanov 2, per-sex, per-gender normalization. The figures give the sum of the distances from the RaDiCHulls to the reference values. At first sight, it looks as if the computation fails to achieve its goal: while the interquartile method seems to indicate a better acquisition of the /ɪ/-/iː/ contrast, Lobanov 2 seems to indicate the contrary. However, some consistent patterns, independent from the method used, appear too: values are more dispersed with /ʊ/-/uː/ than with /ɪ/-/iː/; normalized female productions seem off the mark for /ɪ/-/iː/, but they evolve little from one session to another – unlike /ʊ/-/uː/ productions; and, perhaps more interestingly, RaDiCHulls appear to be the translation of one another along the x-axis across methods: whether that remains the case with other normalization methods has yet to be investigated. Normalization methods entail extra distributional constraints for spontaneous speech. For instance, the Lobanov normalisation procedure requires an even number of vowels to be adequately performed (Lobanov [17]), which goes against the distribution of vowels in native or learner connected speech. The phone inventory in English is skewed to /ɪ/ and /iː/ in the lexicon, but /ʊ/ and /uː/ appear in high-frequency words (e.g. "good" and "too"). It was beyond the scope of this study to account for these parameters, but the need exists for the reliable phonetic assessment of phonemic features in skewed corpora, if only to explore extra-phonemic effects on segmental realizations. These RaDiCHull computations aimed at exploring how to fulfill this need.

Table 1: Sum of the absolute distance from the learners' RaDiCHulls to the American reference values for each method.

	Unnormalized interquartiles	Lobanov 1	Lobanov 2
/ɪ-iː/ (Male)	132.74	320.93	421.31
/ɪ-iː/ (Female)	40.65	736.03	692.38
/ʊ-uː/ (Male)	554.65	590.95	431.84
/ʊ-uː/ (Female)	572.20	189.83	205.60

5. REFERENCES

[1] 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.

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

[3] Best, C. T., McRoberts, G., Goodell, E. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listeners native phonological system. *Journal of the Acoustical Society of America* 109, 775–794.

[4] Bigi, B. 2012. Sppas: A tool for the phonetic segmentations of speech. LREC, , (ed), *Proc. of LREC 2012* 1748–1755.

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

[6] Bybee, J. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

[7] Bybee, J. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

[8] Chan, K. Y., Vitevitch, M. S. 2010. Network structure influences speech production. *Cognitive Science* 34, 685–697.

[9] 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.

[10] 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.

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

[12] Gendrot, C., Adda-Decker, M. 2005. Impact of duration on f1/f2 formant values of oral vowels: an automatic a nalysis of large broadcast news corpora in french and german. Eurospeech, , (ed), *Proceedings Eurospeech* 2453–2456.

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

[14] 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.

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

[16] 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.

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

[18] Major, R. 2001. *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*. Second Language Acquisition Research Series. Taylor & Francis.

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

[20] Tabain, M., Breen, G., Butcher, A. 12 2004. VC vs. CV syllables: a comparison of Aboriginal languages with English. *Journal of the International Phonetic Association* 34, 175–200.

[1] The unit of RaDiCHulls in the interquartile method is Hz^{-1}. For both Lobanov normalization methods, it is $Bark^{-1}$. For the combined purposes of readability and cross-comparison, RaDiCHulls in Hz^{-1} have been multiplied by 10^5, and by 10^3 for $Bark^{-1}$.