Acoustic distances, Pillai scores and LDA classification scores as metrics of L2 comprehensibility and nativelikeness

Paolo Mairano[1], Caroline Bouzon[1], Marc Capliez[1], Valentina De Iacovo[2]

[1]STL UMR 8163 - University of Lille (France), [2]LFSAG - University of Turin (Italy)
(paolo.mairano, caroline.bouzon, marc.capliez)@univ-lille.fr, valentina.deiacovo@unito.it

ABSTRACT

Evaluating L2 pronunciation via acoustic measures is problematic. In the literature, this is sometimes accomplished via fluency metrics (speech rate, average length of IPUs, etc), which however do not capture pronunciation accuracy. Other studies use VOT measurements (only possible if L1 and L2 differ in this respect) or comparison of vowel formants with native values. This contribution uses vowel distances in an F1-F2 space, Pillai scores and classification scores for vowel pairs as acoustic metrics of L2 vowel pronunciation. We compute these metrics on speech produced by 25 learners of L2 English and we compare them with (a) fluency metrics computed on the same productions, (b) VOT measurements, (c) impressionistic judgments provided by native speakers. The advantage of this approach is that L2 pronunciation accuracy is not judged in reference to comparable native productions, but intrinsically: it measures the extent to which phonological vowel contrasts are kept apart in L2 speakers' realisations.

Keywords: vowels, L2 pronunciation assessment, vowel distances, Pillai score, LDA classification.

1. INTRODUCTION

1.1. Measuring L2 pronunciation accuracy

Measuring L2 pronunciation accuracy is reputedly a problematic task and has been a topic of debate [11]. Although sophisticated methods based on speech technologies have been made available in the last decades (e.g. [10]), they are seldom used in the literature, perhaps because of practical issues such as budget constraints or technical complexity. In fact, most studies resort to a few simple techniques: native judgments of comprehensibility or nativelikeness, fluency metrics, VOT. Unfortunately, all these methods have drawbacks. Native judgments are behavioural, and as such have the disadvantage of not being exactly reproducible. Additionally, it can sometimes be difficult (or impossible) to determine the linguistic events that affected native judgments, making them potentially hard to interpret. Fluency metrics (speech rate, length of pauses, number of pauses, etc.) by definition only provide a measure of how smoothly a speaker produces L2 speech, but do not give any insight into how vowels and consonants are pronounced. Finally, VOT (often used as a measure of nativelikeness, cf. [6]) is only viable when L1 and L2 differ in this respect (e.g., Italian/French learners of L2 English, but not German learners of L2 English).

1.2. Aim of this study

In this study, we test acoustic metrics derived from vowel formants as measures of nativelikeness and comprehensibility. In particular, we test:

a. ***Euclidean distances*** of teste-lax vowel pairs in an F1-F2 chart;
b. ***LDA (linear discriminant analysis) classification scores*** for tense-lax vowel pairs;
c. ***Pillai scores*** for tense-lax vowel pairs.

Such measures (cf. 3.1 - 3.3) are taken as indicative of the extent to which L1 phonological categories (e.g. /iː/ - /ɪ/) are kept distinct in L2 speakers' realisations: Euclidean distances indicate how far apart in the acoustic space the two vowels are realised; Pillai scores and LDA classification scores indicate the amount of overlap between realisations of the two vowel phonemes.

Euclidean distances in the acoustic F1-F2 space have been used in the literature with various purposes (comparing vowel systems of languages in [7], measuring prosodic effects on segments in [8], measuring L1 vowel drift in [4], etc.), but seldom as a cue of L2 pronunciation ([15]). The Pillai score has been used to assess the status of vowel mergers and splits in dialectal varieties ([9]), but not as a cue of L2 pronunciation. LDA has been used extensively in the literature for the classification of L2 vowels and consonants ([18]), but here we propose a different approach. The standard method consists in training an LDA model on native vowel realisations, and then use it to classify L2 vowel realisations: so, L2 vowels are evaluated in reference to native vowels. Instead, we train LDA models directly on L2 vowel realisations, and simply use the accuracy score of the model as an indication of the amount of overlap between realisations of different phonological categories. We assume that less overlap corresponds to higher pronunciation accuracy and therefore potentially to higher comprehensibility and nativelikeness.

Such metrics have been computed on tense-lax vowel pairs /iː - ɪ/, /uː - ʊ/, /ɑː - æ/, /ɔː - ɒ/ on speech of 25 French and Italian learners of L2 English. We chose tense-lax vowel pairs because they tend to be assimilated to the same phonological category by L1 Italian ([3]) and L1 French ([12, 15]) learners of L2 English. The values of these metrics are then compared to more common metrics of L2 pronunciation (VOT, fluency measures) and native judgments.

2. METHODOLOGY

2.1. Data

In order to test vowel metrics as described above, we analysed recordings of 25 learners of L2 English from the ICE-IPAC corpus ([1]). 15 learners (F = 11, M = 4, age = 22.3 ±2.46) were native speakers of Italian and were recorded in a sound-proof booth at the University of Turin. 10 learners (F = 8, M = 2, age = 22.5 ±3.44) were native speakers of French and were recorded in a quiet room at the University of Lille. All speakers were attending courses of English at levels spanning B1 to C1. The recording protocol included various tasks (word lists, text reading and dialogues). For this study, we consider only the read-aloud task of a newspaper article (506 words).

2.2. Data annotation

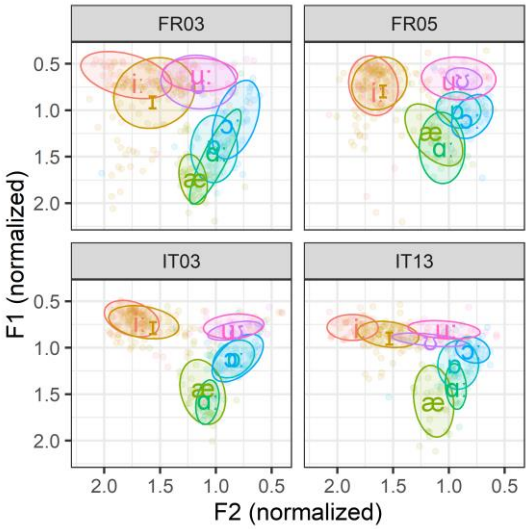
For all recordings, the canonical SBE English transcription was generated and forced-aligned to the signal with *WebMAUS* [13] at word and phoneme levels. A thorough manual verification of the transcription and alignment was then performed on *Praat* [2]. Transcription errors were fixed; misread words, false starts, hesitations and other disruptions were marked for exclusion. Crucially, during the manual step, the phonetic transcription was edited as little as possible in order to reflect the target sounds, not the actual realisations. For instance, the /iː/ in *Peter* was transcribed as [iː] (target sound), irrespective of the actual realisation by learners. Neutralised vowels /i/ and /u/ were transcribed following [21] (e.g. *many* as [ˈmeni]) and were not included in the analysis of /iː - ɪ/ or /uː - ʊ/ pairs.

2.3. Extraction of acoustic parameters

In total, we analysed 95.36 minutes of L2 speech. Duration in ms and formant values in Hz (F1 and F2) were extracted for all realisations of the 8 target vowels /iː - ɪ/, /uː - ʊ/, /ɑː - æ/, /ɔː - ɒ/ via an *ad hoc Praat* script. Formants were extracted from the midpoint of each vowel (cf. [17]) to minimise coarticulation effects, using the Burg method in a band lower than 5.5 kHz for women and 5 kHz for men. Although /iː/ can be slightly diphthongised in some dialects, it was treated as a monophthong (cf. [5]). After the elimination of vowels marked as hesitations, false starts and misreadings, we were left with 6794 realisations. Raw values were then normalised with the Watt & Fabricius method [20].

Fig. 1 shows the F1-F2 plot of the eight target vowels for four learners. The ellipses suggest that speakers FR3 and IT13 distinguish tense-lax vowel pairs to some extent; conversely, tense-lax vowel pairs mostly overlap for speakers FR05 and IT03.

Figure 1: Normalised F1-F2 plots of target vowels for 4 speakers. Ellipses include 1 stdev. of the mean.



3. ANALYSIS AND RESULTS

3.1. Euclidean distances of L2 vowel pairs

Euclidean distances were computed on the normalised F1-F2 chart between tense-lax vowel pairs for each speaker. For instance, we computed the distance between the mean realisation of /iː/ and the mean realisation of /ɪ/ for each speaker. For the four speakers represented in Fig. 1, Euclidean distances are shown in Table 1. The underlying assumption is that distances within each pair will be larger for learners who have developed phonological categories for tense vs lax vowels (e.g. FR03, IT13), while they will be close to zero for learners who have not yet developed such categories (e.g. FR05, IT03).

Table 1: Euclidean distances of vowel pairs on a F1-F2 chart for fours speakers (values are in st. deviations, given that formants are normalised).

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

3.2. LDA classification scores of L2 vowel pairs

For each speaker, we trained four LDA models for classifying vowels in each of the four target pairs: /iː - ɪ/, /uː - ʊ/, /ɑː - æ/, /ɔː - ɒ/. Models were built on *R* 3.5.1 [15] with the *MASS* 7.3 library [18] on all realisations of each speaker taking F1, F2 and duration as predictor variables, with 0.5 prior probability for each vowel. We then ran each model on the same data used for training, in order to obtain a classification accuracy score (i.e. the percentage of realisations within each vowel pair that is correctly classified by the model). We take this score as indicative of the amount of overlap between realisations of the two phonological categories within each vowel pair: a score of 50% (chance level) indicates a thorough overlap of the two phonological categories in a pair, while a score of 100% indicates no overlap. Therefore, we expect higher scores for speakers who have developed phonological categories for tense vs lax vowels. For example, confusion matrices in Table 2 show that /uː - ʊ/ realisations by speakers FR05 and IT03 are more often misclassified (discrimination is close to chance level) than realisations by speakers FR03 and IT13: this suggests a higher amount of overlap for these phonological categories in speakers FR05 and IT03, as correspondingly observed in Fig. 1.

Table 2: LDA confusion matrices for /uː/ - /ʊ/ for four speakers.

		FR03		FR05	
		actual		*actual*	
		/uː/	/ʊ/	/uː/	/ʊ/
predicted	/uː/	70%	43%	50%	45%
	/ʊ/	30%	57%	50%	55%
		IT03		IT13	
		actual		*actual*	
		/uː/	/ʊ/	/uː/	/ʊ/
predicted	/uː/	53%	50%	75%	11%
	/ʊ/	47%	50%	25%	89%

3.3. Pillai scores

Another way of measuring the overlap between vowel categories is the Pillai score, given in the summary of MANOVA. It has been used in sociophonetics to account for the status of vowel mergers and splits in groups of speakers ([8]). We computed it on realisation of target vowel pairs of each L2 speaker, in a MANOVA specified as *F1 + F2 + duration ~ vowel*. Similar to LDA scores and vowel distances, we expect Pillai scores to be higher for participants who have developed phonological categories for tense and lax vowels (FR03, IT13) than for speakers who have not yet developed such categories (FR05, IT03). This is confirmed by the data in Table 3.

Table 3: Pillai scores for vowel pairs produced by 4 speakers as computed within a F1+F2+duration MANOVA.

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

3.4. Traditional L2 pronunciation metrics

As mentioned above, VOT is a standard measure of nativelikeness ([6]). We extracted VOT for all realisations of voiceless plosives occurring immediately before a primary or secondary lexically stressed vowel and not preceded by /s/. Target /p, t, k/ which were not realised as plosives or which were labelled as hesitations or misreadings were excluded from the analysis, leaving 749 observations (n = 204 for /p/, n = 313 for /t/, n = 232 for /k/). VOT was measured on *Praat* from the burst of the plosive to the start of periodic signal.

We also extracted fluency metrics, as these are often used in the literature to evaluate the pronunciation of L2 learners. We computed speech rate (SR, in phone/sec. incl. pauses), articulation rate (AR, in phon/sec. excl. pauses), pause/speech ratio (PSR), and average pause length (APL, in sec.).

3.5. Native judgments

Finally, our recordings of 25 learners of L2 English were evaluated by native speakers. This has been done not only because native judgments are often used in the literature for evaluating the pronunciation of L2 leaners, but also because we wanted to validate the reliability of the vowel metrics under scrutiny.

Native judgments were obtained via *LimeSurvey* [14] from 5 native speakers (3 Southern British English speakers and 2 General American English speakers) who were blind to the aim of our research. Each participant rated the same 100 audio stimuli extracted from the recordings (4 sentences x 25 speakers) and presented in random order. The participants could listen to the audio samples as many times as they wished, and provided ratings of (i) nativelikeness and (ii) comprehensibility on separate ten-point Likert scales. The intra-class correlation coefficients calculated on ratings averaged over the 4 sentences of each speaker were 0.92 (*CI* = 0.86-0.96) for nativelikeness and 0.94 (*CI* = 0.89-0.97) for comprehensibility. For the final analysis, ratings provided by different participants for the same learner were averaged in order to obtain single datapoints.

3.6. Correlation analysis

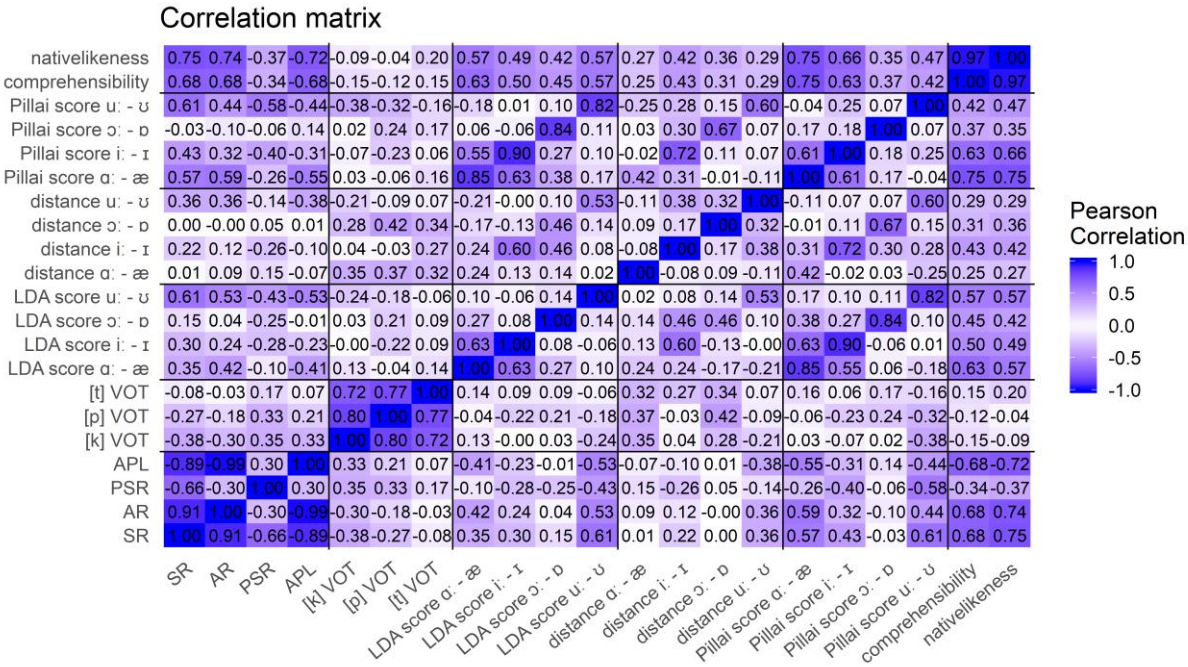
In order to compare all pronunciation metrics and to validate them against native judgments, we computed the correlation matrix of all variables considered, as shown below. Firstly, ratings of nativelikeness and comprehensibility are very strongly correlated with each other (R = .97), but this is no surprise. Secondly, fluency measures (SR, AR, APL, but not PSR) correlate strongly among them, and with native judgments. In fact, among all the metrics considered, they are the ones that best correlate with ratings of both nativelikeness (R > .7) and comprehensibility (R = .68) (all *p-values* < .001). On the other hand, VOT values of /p, t, k/ surprisingly do not seem to correlate significantly with native judgments, the highest correlation value being between /t/'s VOT and ratings of nativelikeness (R = .2, p = .14).

Coming to vowel metrics, LDA classification scores correlate significantly with native judgments of comprehensibility (R = .45 ~ .63, all *p-values* < .05) and, to a slightly lesser extent, nativelikeness (R = 0.42 ~ 0.56, all *p-values* < .05). Interestingly, LDA scores for /uː - ʊ/ and /ɑː - æ/ pairs seem to correlate most with native judgments, with R ~ 0.6 (all *p-values* < .001). Pillai scores correlate strongly with native judgments of comprehensibility and nativelikeness (R = 0.42 ~ 0.75, all *p-values* < .05), except for the /ɔː - ɒ/ pair (R = 0.35 ~ 0.37, all *p-values* < .1). Euclidean distances show mild correlations with native judgments, reaching statistical significance only for /iː - ɪ/ (R = 0.42 and 0.43, p < .05 for both). The better performance of Pillai and LDA scores over Euclidean distances is probably due to the fact that the former (but not the latter) take duration in consideration.

4. CONCLUSION

The results of this study have shown that LDA classification scores and, to a lesser extent, Euclidean distances of tense-lax vowel pairs correlate significantly with native judgments of nativelikeness and comprehensibility for 25 speakers of L2 English. Although the number of speakers and vowels analysed is relatively small, we believe that these results are encouraging. If replicated on more data, they may provide a viable alternative to traditional methods for evaluating L2 speech. This approach evaluates the ability of L2 speakers to produce distinct realisations for different phonological categories. In other words, L2 pronunciation is measured intrinsically, without referring to L1 speech data. We believe that this advantage makes it a viable solution for studies where no control data produced by L1 speakers is available.

Finally, we would like to point out two limitations of these metrics. Firstly, they can of course be affected by well-known issues in formant detection and normalisation. Secondly, they do not directly measure pronunciation *accuracy*, i.e. the similarity of L2 realisations to L1 realisations. Rather, they evaluate L2 pronunciation intrinsically by measuring the extent to which phonological categories are kept apart. A learner producing the /iː/ - /ɪ/ contrast as [iː] - [e] would be likely to score erroneously high. Finally, these metrics are clearly less useful for L2s with simple vowel systems, where no cases of single-category assimilations are likely in L2 speech. Nevertheless, the results of our study are encouraging, and we think these metrics deserve further investigation.



7. REFERENCES

[1] Andreassen, H. N., Herry-Bénit, N., Kamiyama, T., & Lacoste, V. (2015). The ICE-IPAC project: testing the protocol on Norwegian and French learners of English. Proc. of the 18 th ICPhS, Glasgow UK.

[2] Boersma, P. & Weenink, D. 2018. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.43, retrieved 8.9.2018 from http://www.praat.org/

[3] Busà, M. G. 1995. *L'inglese degli italiani: l'acquisizione delle vocali*. Padova: Unipress.

[4] Chang, C. B. 2011. Systemic drift of L1 vowels in novice L2 learners. *Proc. 17th ICPhS* Hong-Kong, 428-443.

[5] 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.

[6] Flege, J. E. 1991. Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *JASA*, 89(1), 395-411.

[7] 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. *Proc. 16th ICPhS* Saarbrücken, 1417-1420.

[8] Gendrot, C., & Gerdes, K. 2009. Prosodic hierarchy and spectral realization of vowels in French. *Interface Discours & Prosodie*, 191-205.

[9] Hall-Lew, L. 2010. Improved representation of variance in measures of vowel merger. *Proc. of Meetings on Acoustics* (Vol. 9, No. 1).

[10] Hincks, R. 2003. Speech technologies for pronunciation feedback and evaluation. *ReCALL*, 15(1), 3-20.

[11] Isaacs, T., & Thomson, R. I. 2013. Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135-159.

[12] Iverson, P., Pinet, M., & Evans, B. G. 2012. Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, 33(1), 145-160.

[13] Kisler, T. and Reichel U. D. and Schiel, F. 2017. Multilingual processing of speech via web services, *Computer Speech & Language*, 45, 326–347.

[14] LimeSurvey Project Team. 2012. *LimeSurvey: An Open Source survey tool*. http://www.limesurvey.org

[15] Méli, A & Ballier, N. 2015. Assessing L2 phonemic acquisition: a normalisation-independent method? *Proc. 18th ICPhS* Glasgow, 805-810.

[16] R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

[17] Rathke, T., Stuart-Smith, J. Torsney, B. & Harrington, J. 2016. The beauty in a beast: Minimising the effects of diverse recording quality on vowel formant measurements in sociophonetics real-time studies. *Speech Communication*, 86, 24-41.

[18] Strik, H., Truong, K. P., Wet, F. D., & Cucchiarini, C. 2007. Comparing classifiers for pronunciation error detection. *Proc. of INTERSPEECH*, 1837-1840.

[19] Venables, W. N. & Ripley, B. D. 2002. *Modern Applied Statistics with S*. 4th Ed. Springer, New York.

[20] Watt, 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(9), 159-173.

[21] Wells, J. C. 2008. *Longman pronunciation dictionary*. Longman.