Can L2 Pronunciation Be Evaluated without Reference to a Native Model? Pillai Scores for the Intrinsic Evaluation of L2 Vowels

Paolo Mairano, Fabián Santiago, Leonardo Contreras Roa

languages

MDPI

Article

Can L2 Pronunciation Be Evaluated without Reference to a Native Model? Pillai Scores for the Intrinsic Evaluation of L2 Vowels

Paolo Mairano [1,*], Fabián Santiago [2] and Leonardo Contreras Roa [3]

[1] UMR 8163 STL, Université de Lille, 59653 Villeneuve d'Ascq, France
[2] UMR 7023 SFL, Université Paris 8, 75017 Paris, France; fabian.santiago.ling@gmail.com
[3] EA 4295 Corpus, Université de Picardie Jules Verne, 80080 Amiens, France; l.contreras.roa@u-picardie.fr
* Correspondence: paolo.mairano@univ-lille.fr

Abstract: In this article, we explore the possibility of evaluating L2 pronunciation, and, more specifically, L2 vowels, without referring to a native model, i.e., intrinsically. Instead of comparing L2 vowel productions to native speakers' productions, we use Pillai scores to measure the overlap between target vowel categories in L2 English (/iː/ — /ɪ/, /ɑː/ — /æ/, /ɜː/ — /ʌ/, /uː/ — /ʊ/) for L1 French, L1 Spanish, and L1 Italian learners (n = 40); and in L2 French (/y/ — /u/, /ø/ — /o/, /ø/ — /e/, /ɛ̃/ — /e /, /ɑ̃/ — /a/, /ɔ̃/— /o/) for L1 English, L1 Spanish, and L1 Italian learners (n = 48). We assume that a greater amount of overlap within a contrast indicates assimilated categories in a learner's production, whereas a smaller amount of overlap indicates the establishment of phonological categories and distinct realisations for members of the contrast. Pillai scores were significant predictors of native ratings of comprehensibility and/or nativelikeness for many of the contrasts considered. Despite some limitations and caveats, we argue that Pillai scores and similar methods for the intrinsic evaluation of L2 pronunciation can be used, (i) to avoid direct comparisons of L2 users' performance with native monolinguals, following recent trends in SLA research; (ii) when comparable L1 data are not available; (iii) within longitudinal studies to track the progressive development of new phonological categories.

Keywords: L2 pronunciation; L2 phonology; acquisition; vowels; comprehensibility

Citation: Mairano, Paolo, Fabián Santiago, and Leonardo Contreras Roa. 2023. Can L2 Pronunciation Be Evaluated without Reference to a Native Model? Pillai Scores for the Intrinsic Evaluation of L2 Vowels. *Languages* 8: 280. https://doi.org/10.3390/languages8040280

Academic Editor: Elena Babatsouli

Received: 6 August 2023
Revised: 15 November 2023
Accepted: 21 November 2023
Published: 28 November 2023

1. Introduction

1.1. Evaluating L2 Pronunciation

The assessment of second language (henceforth, L2) pronunciation is reputedly a complex and problematic task, and has been the subject of debate in the literature (see, for instance, Isaacs and Trofimovich 2012). Traditionally, pronunciation in an L2 has been evaluated with reference to a native model, and native-like pronunciation was considered the ultimate goal. Yet, the growing use of English as an International Language has led some authors to reconsider these positions, and focus on constructs such as intelligibility and comprehensibility rather than native-likeness and foreign-accentedness, or to look at pronunciation in the context of communicative language abilities (e.g., Jenkins 2000, 2006). In parallel, global and influential trends in second language acquisition research propound that L2 learners (or L2 'users') should not be judged by native monolingual standards, and that the final goal of L2 classes should not be to bring learners to become as similar as possible to native monolingual speakers (Cook 2016).

The concepts of intelligibility and comprehensibility have been at the core of much literature in L2 pronunciation since the pivotal study by Munro and Derwing (1999). Both terms allude to how understandable L2 speech is, the former specifically referring to the proportion of L2 speech that is actually understood by listeners and usually operationalised

via transcription tasks (Kang et al. 2018), the latter referring to the amount of effort demanded by listeners for understanding L2 speech and usually operationalised via rating scales (Crowther et al. 2022; but alternative approaches use response times or keystroke latency in transcriptions—see, for instance, Gallant 2023). Transcription tasks and rating scales are behavioural, and as such, they are not entirely objective and reproducible. Unsurprisingly, some studies have found individual differences in pronunciation scoring, and various linguistic and extralinguistic factors have been revealed to affect or bias such assessments (Bent and Bradlow 2003; Isaacs and Thomson 2013).

In the effort to develop methods for the objective evaluation of L2 pronunciation, several studies have searched for phonetic correlates of comprehensibility and intelligibility, trying to derive objective measures for automatic scoring. Often, studies within this domain have proposed global temporal or prosodic measures, such as speech rate, articulation rate, number/length/frequency of pauses, and pitch range (e.g., Cucchiarini et al. 2000; Kang 2010; Kang and Pickering 2013). More rarely, authors have also considered local segmental or suprasegmental measures. For example, Isaacs and Trofimovich (2012) considered ratios of segmental errors, syllable structure errors, word stress errors, vowel reduction errors, and errors in pitch contours, and found that all of them correlated with native ratings of comprehensibility. Among the many variables considered (which also included many non-phonological features relating to fluency, linguistic resources, and discourse), the only one that was not significantly correlated with native ratings was pitch range. A following study by Saito et al. (2017) expanded on these findings, confirming that comprehensibility ratings correlated with a full set of variables at many linguistic levels, while ratings of nativelikeness mainly correlated with phonological variables (further confirmed by Saito 2021).

The search for objective measures for L2 pronunciation assessment is obviously also relevant within the field of speech technologies. Recent advances in this domain, and particularly in automatic speech recognition (ASR), have driven the development of L2 pronunciation automatic assessment systems, as well as computer-assisted pronunciation training systems (CAPT) (see Xi 2010; Farrús 2023). Automatic assessment of L2 pronunciation is usually performed by statistical models built on top of exclusively or preponderantly native speech. Modern ASR-based CAPT systems typically compare L2 speech productions with matching L1 instances, often isolated words or short sentences. The ASR component can be used to compute a variety of scores, the most widespread being word error rate (see, for example, Liakin et al. 2015). This measure indicates how many words have been misrecognised by the ASR within a sentence produced by L2 learners, on the assumption that misrecognised words correspond to, and are caused by, pronunciation errors. Other approaches have been developed (see Witt 2012 for a review), such as log-likelihood-based pronunciation scores (e.g., the goodness of pronunciation score used by Witt and Young 2000; and Neri et al. 2008), or the use of classifiers to detect expected pronunciation errors for given L1-L2 pairs (see, for instance, Strik et al. 2009; Yoon et al. 2010). Although some recent CAPT systems use industrial closed-source ASR models (among others, Tejedor-Garcia et al. (2020) use Google's ASR; Liakin et al. (2015) use Nuance Dragon ASR), it is reasonable to assume that such models are mostly or exclusively trained with native speech data. More generally, as claimed by Witt and Young, since the goal is to "*to assess pronunciation quality with respect to native speaker performance, it is reasonable to use native speakers to train the acoustic models*" (Witt and Young 2000, pp. 97–98). For a recent and thorough literature review on the use of ASR for computer-assisted pronunciation training, see Farrús (2023).

We shall note that studies looking for phonetic and phonological correlations of comprehensibility and nativelikeness, as well as systems for automatic pronunciation assessment or feedback, tend to refer to a native model. As for automatic assessment, the native model is inscribed in the architecture of the system (L2 productions are compared to matching native instances to check for mispronunciations, e.g., deviations from the native model). As for studies looking for phonetic correlates of comprehensibility, the

phonological variables included in the analysis by Isaacs and Trofimovich (2012) and Saito et al. (2017) (among others) are computed with reference to what a native speaker would do (whether at the segmental or suprasegmental level): for instance, the ratio of segmental errors in L2 speech essentially indicates segmental deviations with respect to the pronunciation by a native speaker. Assessing L2 pronunciation in reference to a native model is, of course, the most obvious approach that one can adopt: we shall call this type of evaluation 'extrinsic', because it relies on external (native) data. Instead, the aim of this article is to explore the possibility of implementing objective measures of what we shall define as the 'intrinsic' evaluation of L2 pronunciation, i.e., without relying on external native data.

1.2. Reasons and Methods for the Intrinsic Evaluation of L2 Pronunciation

We believe that the attempt to evaluate L2 pronunciation intrinsically is interesting and potentially advantageous from various perspectives. From a theoretical point of view, the concept of intrinsic evaluation may respond to recent SLA trends avoiding direct comparisons of L2 learners with native monolingual speakers (see previous section), as well as to assessment approaches trying to abandon the focus on nativelikeness. From a practical perspective, evaluating L2 pronunciations without a reference to a native model has the notable advantage of eluding issues related to different native varieties; this may be particularly convenient for languages with multiple standardised varieties, such as English (General American, Southern British English, etc.), French, Spanish, and Portuguese. Moreover, intrinsic evaluation can come in handy when researchers wish to evaluate the pronunciation of L2 learners and do not dispose of comparable L1 data.

Having established that intrinsically evaluating L2 pronunciation implies desisting from referring to a native model, we shall look for acoustic measures that allow us to objectively assess L2 oral productions by exclusively examining L2 speech itself. In this initial exploration, we will focus on vowels, whose acoustic properties are widely known and relatively easily analysed. A possible approach for assessing L2 vowel pronunciation intrinsically has been inaugurated by Mairano et al. (2019) and consists of quantifying the development of new phonological categories in the L2. This approach is rooted in L2 phonology acquisition theories, and assumes that the production of new L2 vowel categories develops progressively during acquisition and is linked to perceptual discrimination. Models of L2 phonology postulate that instances of L2 phonological categories which are perceived as similar to existing categories in a learner's L1 may be perceptually assimilated to the L1 category. More specifically, the Perceptual Assimilation Model (PAM, Best and Tyler 2007) predicts that a non-native contrast in which both sounds have the same closest L1 equivalent will be mapped to the same L1 category, with L2 sounds being either equally good instances of the L1 category ('single category assimilation', e.g., English /i:/ — /ɪ/ mapped to /i/ by French$_{L1}$ listeners), or with one L2 sound being a better instance than the other ('goodness of fit assimilation', e.g., French /y/ — /u/ mapped to /u/ by Spanish$_{L1}$ listeners). L2 phonology models postulate a connection between L2 perception and L2 production, although the exact nature of this connection is much debated: for instance, the original Speech Learning Model claimed that a correct perception of L2 sounds needed to precede their production (Flege 1995), while the revised version of this model states that perception and production 'co-evolve without precedence' (Flege and Bohn 2021). Therefore, our assumption is that contrasts affected by single category or goodness-of-fit assimilation by an L2 learner will mostly result in assimilated realisations in production, other than in perception. For instance, Georgiou (2022) found that Greek$_{L1}$ learners of English$_{L2}$ could not discriminate perceptually English /i:/ — /ɪ/, and realisations for these vowel categories did not differ in their productions. Similarly, Spanish$_{L1}$ learners of French$_{L2}$ who perceptually assimilate /y/ and /u/ are also likely to produce the same vowel /u/ for both categories (see Racine and Detey 2018). For illustration purposes, Figure 1 shows simulated /y/ and /u/ realisations by three fictional L2 learners who systematically distinguish, only partially distinguish, and do not distinguish these vowel categories.

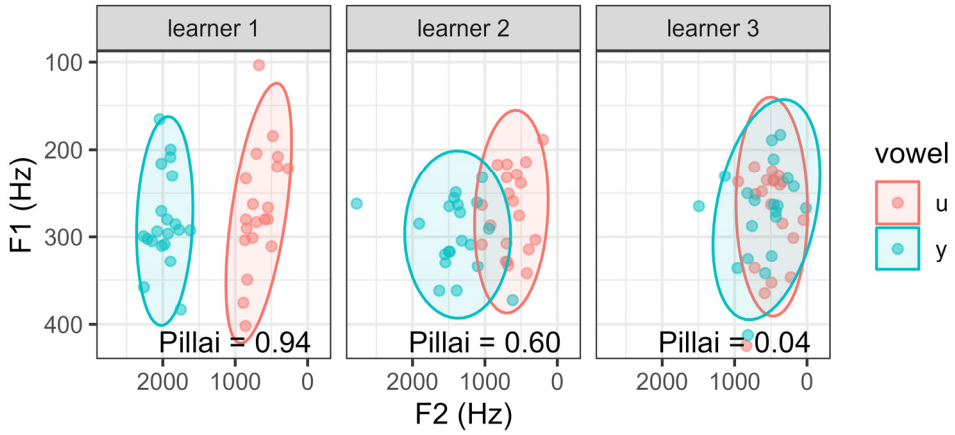

Figure 1. Simulated instances of /y/ and /u/ for three fictional learners of French$_{L2}$.

Based on these considerations, the intrinsic evaluation of L2 vowels can be implemented by quantifying the amount of distance and/or overlap between realisations of the two target phonological categories: a greater amount of overlap should indicate that the L2 learner has not (fully) developed distinct phonological categories; vice versa, a smaller amount of overlap should indicate that the L2 learner is able to produce distinct realisations for the two phonological categories. Vowel distance and overlap can be quantified by a certain number of measures, such as Euclidean distance on the F1/F2 vowel space, Mahalanobis distance, Bhattacharyya's Affinity measure, Pillai score, among others. While Pillai scores have seldom been used for L2 speech, they have been extensively studied in the sociophonetic literature for investigating vowel mergers and splits, and there seems to be growing consensus that Pillai scores perform best in quantifying vowel distance and overlap (Nycz and Hall-Lew 2013; Kelley and Tucker 2020). They are a statistical measure output from a Manova, and are easily computed on mainstream software for statistical analysis (see Nycz and Hall-Lew 2013, for details); the score ranges from 0 (complete overlap) to 1 (complete separation). One of their advantages is that there is no limit to the number of dimensions over which overlap is computed, making it possible to include not only F1 and F2, but, for example, duration and/or F3.

Finally, it should be noted that measures of vowel distance or overlap do not constitute an intrinsic evaluation per se. In fact, Euclidean distances, Mahalanobis distances, and Pillai scores have sometimes been used in the L2 literature to measure the distance or overlap of L2 realisations with comparable native realisations (e.g., Flege et al. 1997; Kartushina and Frauenfelder 2014; Perry and Tucker 2019); in this case, they are used to compute the distance from (or overlap with) a native model, and therefore qualify as an extrinsic (rather than intrinsic) evaluation. Studies having used these measures for intrinsic evaluations are few and very recent (Mairano et al. 2019; Kabakoff et al. 2020; Bailles et al. 2022; De Jonge et al. 2022; Valenzuela Farias 2022; Leppik et al. 2023).

1.3. Aim of This Contribution

The goal of this article is to propound Pillai scores as a method for the intrinsic evaluation of L2 vowels. We expand on a previous exploratory study (Mairano et al. 2019), and apply Pillai scores to a larger cohort of participants and languages. While the original study only analysed English$_{L2}$ data by Italian$_{L1}$ and French$_{L1}$ participants, we now add English$_{L2}$ data by Spanish$_{L1}$ learners, as well as French$_{L2}$ data by English$_{L1}$, Spanish$_{L1}$ and Italian$_{L1}$ learners. Intrinsic evaluation is compared to traditional native ratings of comprehensibility and nativelikeness obtained by English$_{L1}$ and French$_{L1}$ speakers, respectively, for the two types of data analysed. Furthermore, intrinsic vs. extrinsic evaluations are compared for the French$_{L2}$ data, using Pillai scores computed intrinsically (i.e., measuring the overlap between vowel pairs for target contrasts; see above) and extrinsically (i.e., measuring the overlap between native and non-native realisations for target phonemes).

The speech data are taken from learner corpora designed for phonetic analysis, and are presented in detail in Section 2, alongside the methods used for annotation, normalisation, and acoustic parameter extraction. Sections 3.1 and 3.2 present the results of intrinsic evaluation with Pillai scores for data of English$_{L2}$ and French$_{L2}$ respectively, while Section 3.3 presents the results of an extrinsic evaluation using Pillai scores. Section 4 discusses the results, outlines other possible methods for the intrinsic evaluation of L2 pronunciation, envisages the extension of this paradigm to consonants and other pronunciation features, and alerts users to a number of pitfalls and caveats. Section 5 provides a concise conclusion promoting the use of intrinsic evaluation under certain circumstances.

2. Materials and Methods

2.1. English$_{L2}$ Data

The English$_{L2}$ data analysed in this article comes from the IPCE (InterPhonology of Contemporary English) learner corpus and includes 40 speakers: 15 Italian$_{L1}$ learners, 15 French$_{L1}$ learners, and 10 Spanish$_{L1}$ learners. While the protocol for the IPCE learner corpus includes numerous production tasks (word lists, readings, and dialogues; see Herry-Bénit et al. 2021, for a complete description), we only used data from the read-aloud task of a newspaper article (506 words).

The 15 French$_{L1}$ learners were native speakers of metropolitan French (12 F, 3 M, age = 24, SD = 6.59). Ten of them were students at the University of Lille, and five of them were students at the University of Paris 13 Nanterre. Their L2 English level ranged from B1 to C1 of the CEFR, and their age of first contact with English was 10.1 years on average (SD = 3.36). Some of them reported speaking other languages, namely Spanish (n = 4), German (n = 2), Arabic (n = 1), Berber (n = 1), Italian (n = 1), and Portuguese (n = 1).

The 15 Italian$_{L1}$ learners (11 F, 4 M; age = 22.5, SD = 2.38) were students at the University of Turin at the time of recording. Ten of them were from the Turin area, two of them had grown up in other areas of northern Italy, two of them in southern Italy, and one in Sardinia. Their L2 English level ranged from B1 to C1 of the CEFR, and their age of first contact with English was 10.6 years, on average (SD = 3.36). All of them reported speaking one or more other foreign languages, namely French (n = 5), Spanish (n = 5), German (n = 5), Arabic (n = 2), Russian (n = 2), Catalan (n = 1), Japanese (n = 1), and Portuguese (n = 1).

The 10 Spanish$_{L1}$ learners (3 F, 7 M; age = 30.2, SD = 6.98) were from South America, more precisely Peru (n = 5), Colombia (n = 3) and Chile (n = 2). Their L2 English level ranged from B1 to C1 of the CEFR, and their age of first contact with English was 7.4 years on average (SD = 2.84). Most of them lived in Paris or Lille at the time of recording and therefore also spoke French (n = 9). One of them also spoke Italian, and another spoke Portuguese and German. One participant lived in Chile and did not speak any foreign language other than English. The fact that most L1 Spanish participants lived in France at the time of recording may be considered a confound. Clearly, we do not exclude some influence of French on their English$_{L2}$ performance, but these learners can clearly be auditorily classified as Spanish accented; additionally, previous studies on /s/ voicing for these participants (Mairano et al. 2021) did not reveal any strong influence of French.

Recording conditions and equipment varied for the three groups, but all participants were recorded in a sound-attenuated or quiet room, in order to ensure good audio quality suitable for acoustic analysis. Since Italian$_{L1}$ and Spanish$_{L1}$ learners had origins from various regions (or countries, in the case of Spanish$_{L1}$ learners), their L1 vowel inventories may have been affected by L1 regional variety, and we cannot exclude that this is reflected in their L2 English production patterns. In effect, some studies have revealed differences in L2 perception and L2 production depending on L1 regional variety (see Escudero et al. 2012; Escudero and Williams 2012; Marinescu 2013), although evidence for L2 production is more controversial (see Simon et al. 2015). However, we do not consider potential L1 dialect effects on L2 vowels to be problematic within our study, given that our goal is to measure the degree to which learners manage to distinguish L2 vowel categories. The fact

that learners of a certain L1 variety may potentially find it more or less easy to distinguish specific vowel contrasts, or that the phonetic patterns of their realisations may be different across L1 dialects, only yields a richer and more heterogeneous testbed.

2.2. French$_{L2}$ Data

The French$_{L2}$ data analysed in this article includes Italian$_{L1}$, Spanish$_{L1}$, and English$_{L1}$ learners and were taken from three different corpora, respectively—*ProSeg* (Delais-Roussarie et al. 2018), *Coreil* (Delais-Roussarie and Yoo 2010), and *AixOx* (Herment et al. 2014). As a consequence, the L2 French data are more heterogeneous than the L2 English data. Although the protocols of the various corpora differ, all of them include read speech, which was therefore used for our analysis.

The 25 Italian$_{L1}$ learners (21 F, 4 F; age = 25.28, SD = 3.7) from the *ProSeg* corpus were learners of L2 French recruited at the University of Turin. Their self-reported proficiency levels ranged from B1 to C1. While the *ProSeg* corpus includes various tasks, for the present study we only considered the reading aloud of eight short passages.

The 13 Spanish$_{L1}$ learners (6 F, 7 M; age = 25.4, SD = 8.8) from the *Coreil* corpus were students at the Autonomous University of Mexico (UNAM). Their self-reported proficiency level ranges from A2 to B2. We analysed read speech of short passages.

The 10 English$_{L1}$ learners (5 F, 5 M; age = 22, SD = 2) from the *AixOx* corpus were students at the University of Oxford, and their proficiency level is reported to span from B1 to B2. They were recorded while reading the same short passages used in the *Coreil* corpus.

Since the *AixOx* corpus was designed to study English$_{L2}$ and French$_{L2}$ phonology, it includes recordings of French$_{L1}$ and English$_{L1}$ learners speaking both French$_{L2}$ and English$_{L2}$. So, apart from the recordings of 10 English$_{L1}$ learners, we analysed the recordings of 10 French$_{L1}$ speakers (5 F, 5 M; age = 35, SD = 14) in order to compute extrinsic L2 pronunciation models for comparison with intrinsic evaluation (see Section 3.3).

Recording conditions and equipment differ for the three groups of learners analysed, but all corpora were designed for the study of L2 phonetics/phonology and provide a sufficient audio quality for acoustic analysis.

2.3. Selection of Vowel Contrasts

We established that our method for the intrinsic evaluation of L2 vowels should quantify the amount of overlap for difficult L2 contrasts. In particular, we need contrasts for which the two items may be assimilated perceptually to the same category (in terms of the PAM model, they may fall into single-category or category goodness assimilation; see Section 1.2), on the assumption that their production will likewise be assimilated to a single articulatory pattern.

For English$_{L2}$, we selected tense vs. lax vowel pairs, namely /iː/ — /ɪ/, /ɑː/ — /æ/, /ɜː/ — /ʌ/, and /uː/ — /ʊ/. We shall note that occurrences of neutralised vowels /i/ and /u/ (also known as *happy vowel* and *thank you vowel*, respectively), occurring in an unstressed word-final position as well as in unstressed syllable-final position, if immediately followed by another vowel, were of course not included in the analysis of /iː/ — /ɪ/ or /uː/ — /ʊ/, given that this contrast is neutralised in this context. We also chose not to consider the /ɔː/ — /ɒ/ contrast, due to its neutralisation in many varieties of English. Our English$_{L2}$ learners do not have a tense vs. lax contrast in their L1, and unsurprisingly the production of these L2 sounds has been described as difficult in the literature for French$_{L1}$ (Méli and Ballier 2019), Spanish$_{L1}$ (Fullana Rivera and MacKay 2003), Italian$_{L1}$ (Flege et al. 1999) learners, as well as for learners of many other L1s such as Greek$_{L1}$ (Georgiou 2022) and Catalan$_{L1}$ (Cebrian 2006). We therefore consider tense and lax vowels as a potential case of category assimilation in L2 perception and production.

The main difficulties in French$_{L2}$ are front rounded vowels (/y/, /ø/, /œ/) and nasal vowels (/ɛ̃/, /ɑ̃/, /ɔ̃/, /œ̃/), which are marked sounds and relatively infrequent in the world's languages. It seems reasonable to assume that front rounded vowels can be assimilated to the corresponding primary cardinal vowels, i.e., either unrounded front

vowels /i/, /e/, /ɛ/, or rounded back vowels /u/, /o/, /ɔ/. There is evidence in the literature that /y/ tends to be assimilated to /u/ by English$_{L1}$ (Ruellot 2011; Darcy et al. 2012; Liakin et al. 2015; Melnik-Leroy et al. 2022), Spanish$_{L1}$ (Racine and Detey 2018), and Italian$_{L1}$ learners (Pillot-Loiseau and Grando 2020); however, it seems that English$_{L1}$ learners tend to assimilate /ø/ — /o/ (Darcy et al. 2012), while Spanish$_{L1}$ and Italian$_{L1}$ learners tend to assimilate /ø/ — /e/ (Kartushina and Frauenfelder 2014; Mairano and Santiago 2020). Therefore, we tested mid front rounded vowels against the corresponding back rounded vowels for English$_{L1}$ learners, and against front unrounded vowels for Spanish$_{L1}$ and Italian$_{L1}$ learners. As for nasal vowels, since the nasalisation feature is absent in English, Spanish, and Italian, we can reasonably assume that such vowels will be assimilated to the corresponding oral vowels. With the aim of simplifying the analysis, we merged realisations of /ø/ and /œ/, as well as /ɛ̃/ and /œ̃/ (see Santiago and Mairano 2021, for a similar approach). This is justified because (i) mid-close and mid-open vowels are neutralised in many native varieties of French, and (ii) /ø, œ/ are not distinguished by spelling and /e, ɛ/ only partially so, and they are often not taught in L2 classes. In summary, we considered the following vowel contrasts in French$_{L2}$: /y/ — /u/, /ɛ̃/—- /e, ɛ/, /ɑ̃/ — /a/, /ɔ̃/ — /o, ɔ/ for all learners; /ø, œ/ — /e, ɛ/ for Italian$_{L1}$ and Spanish$_{L1}$ learners; /ø, œ/ — /o, ɔ/ for English$_{L1}$ learners.

2.4. Data Preparation, Extraction, and Analysis

For all recordings, the canonical transcription was generated and forced-aligned to the signal at word and phoneme level. For the English$_{L2}$ data, we used Southern British English transcriptions and acoustic model on *WebMAUS* (Kisler et al. 2017); for the French$_{L2}$ data, we used Parisian French transcriptions and acoustic model on *Easyalign* (Goldman 2011). A thorough manual verification of the transcription and alignment was then performed on *Praat* (Boersma and Weenink 2023) by the authors. Transcription errors were fixed; misread words, false starts, hesitations, and other disruptions were marked for exclusion. Other than that, the phonetic transcription was edited as little as possible in order to reflect target sounds, rather than actual realisations. For instance, the /i:/ in *Peter* was transcribed as [i:] (target sound), irrespective of the actual realisations produced by learners.

For all target vowels, we extracted acoustic parameters via an ad hoc *Praat* script written by the first author: for English$_{L2}$ vowels, we extracted F1, F2, and duration (the latter being one of the primary cues for the tense-lax distinction); for French$_{L2}$ vowels, we extracted F1, F2 and F3 (the latter being an acoustic cue of labialisation). Additionally, the following fluency metrics were calculated: AR (articulation rate, i.e., number of phonemes per second excluding pauses), SR (speech rate, i.e., number of phonemes per second including pauses), PSR (pause/speech ratio, i.e., total pause duration / total speech duration), APL (average pause length). Formants were extracted from the midpoint of each vowel (Rathcke et al. 2017) to minimise coarticulation effects, using the Burg method in a band lower than 5.5 kHz for women and 5 kHz for men. Although /i:/ can be slightly diphthongised in some dialects of English, it was treated as a monophthong (Ferragne and Pellegrino 2010).

The data were saved in .csv format and imported into *R* (R Core Team 2023) for analysis. We eliminated vowels marked as hesitations, false starts, and misreadings. In order to address potential formant detection errors, we eliminated realisations for which F1 or F2 was beyond 2.5 *SDs* from the mean of each vowel. It should be noted that some authors use filters to eliminate detection errors, (e.g., Gendrot et al. 2016), but this approach is not viable for L2 data, where some vowels may be deviant from a given L1 norm. After this, we were left with 10,294 vowel realisations in English$_{L2}$ and 27,612 vowel realisations in French$_{L2}$ (see Table 1 for details). Raw values were then normalised with Nearey1's (formant intrinsic) method with the *PhonTools* library (version 0.2.2.1, Barreda 2015). Pillai scores were computed for every participant and for every target contrast on *R*, running Manovas with F1, F2 and duration for English$_{L2}$ data, and F1, F2, F3 for French$_{L2}$ data. The relation between Pillai scores and native ratings of comprehensibility and nativelikeness

was estimated with linear mixed-effects models fitted with the *lme4* package (version 1.1.33, Bates et al. 2015). *p* values were obtained with the *lmerTest* library (version 3.1.3, Kuznetsova et al. 2017), and R^2 values with the *MuMin* library (version 1.47.5, Bartoń 2023).

Table 1. Total number of tokens analysed per vowel for English$_{L2}$ and French$_{L2}$. Figures in parentheses give the tokens analysed for control French$_{L1}$ speakers.

English								
/iː/	/ɪ/	/ɑː/	/æ/	/uː/	/ʊ/	/ɜː/	/ʌ/	
1177	5078	418	1433	707	419	221	841	
French								
/y/	/u/	/ø, œ/	/e, ɛ/	/o, ɔ/	/a/	/ɛ̃/	/ɑ̃/	/ɔ̃/
1721 (235)	1271 (272)	1107 (238)	8739 (1818)	2957 (485)	6447 (1263)	1427 (321)	2342 (483)	1601 (289)

2.5. Native Ratings

We collected native ratings of comprehensibility and nativelikeness for all learners with the aim of verifying how intrinsic the pronunciation assessment correlates with human judgments. Native ratings were obtained via an online experiment on the *LimeSurvey* platform (Limesurvey Project Team 2012) from 5 native speakers of English (3 Southern British English speakers and 2 General American English speakers), and 5 native speakers of French (four speakers of Parisian French and 1 speaker of Southern French from Toulouse), all of whom were blind to the aim of our research. All raters had L2 teaching experience of English and French in various countries, and had therefore been exposed to foreign accents. Four sentences were extracted from the recordings of each learner (the same four sentences were used for all English$_{L2}$ learners; meanwhile this was not possible for French$_{L2}$, since the *ProSeg* corpus did not have the same texts as *Coreil* and *AixOx*).

Every native listener rated all the audio stimuli (4 sentences × 40 speakers of English$_{L2}$ = 160 stimuli, 4 sentences × 43 speakers of French$_{L2}$ = 172 stimuli) extracted from the recordings and presented in random order. Participants could listen to the audio samples as many times as they wished, and provided ratings of nativelikeness and comprehensibility on separate ten-point Likert scales. The intra-class correlation (ICC) coefficients calculated on ratings averaged over the 4 sentences produced by each learner were high for nativelikeness as well as comprehensibility, and for English as well as French (see Table 2). Unsurprisingly, ratings for nativelikeness were globally lower than for comprehensibility (see Table 2), and both were correlated with each other (R = 0.73 for the English data, R = 0.76 for the French data). Ratings of comprehensibility were comparable across English$_{L2}$ and French$_{L2}$, whereas ratings for nativelikeness were lower for French$_{L2}$ than for English$_{L2}$: this may be because French$_{L2}$ learners were on average less nativelike than English$_{L2}$ learners, or because French listeners were stricter in their evaluations. For the final analysis, ratings provided by different listeners for the same learner were averaged in order to obtain single datapoints.

Table 2. Information about native ratings performed on a 10-point Likert scale.

	English$_{L2}$		**French$_{L2}$**	
	Comprehensibility	**Nativelikeness**	**Comprehensibility**	**Nativelikeness**
Mean	7.2 (*SD* = 1.84)	5.28 (*SD* = 1.69)	7.29 (*SD* = 1.58)	3.46 (*SD* = 2.04)
ICC	0.90 (*CI* = 0.84, 0.94)	0.95 (*CI* = 0.92, 0.97)	0.91 (*CI* = 0.86, 0.94)	0.88 (*CI* = 0.81, 0.92)

3. Results

3.1. Intrinsic Evaluation of English$_{L2}$ Vowels with Pillai Scores

The intrinsic evaluation of the English$_{L2}$ vowel contrasts obtained with Pillai scores was compared with native judgments in order to analyse their relation. Figure 2 shows

Pillai scores obtained for every English$_{L2}$ learner and target vowel contrast, plotted with native judgments of comprehensibility and nativelikeness for the same learner.

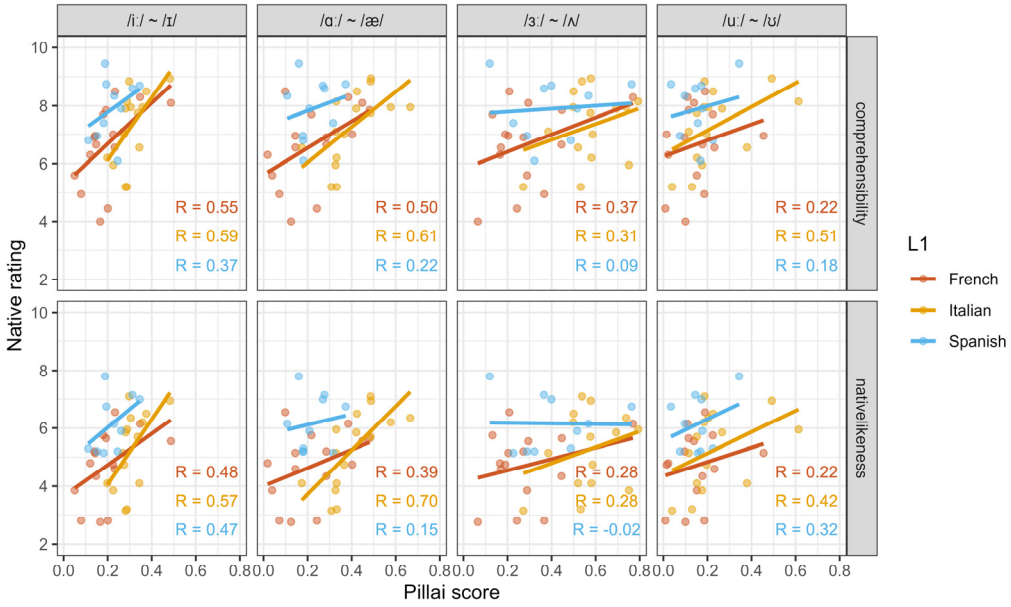

Figure 2. By−participant Pillai scores for each English$_{L2}$ target vowel contrast, plotted with native ratings of comprehensibility (**above**) and nativelikeness (**below**).

The results show a connection between Pillai scores and native ratings: in order to evaluate this relation statistically, we fitted linear mixed-effects models predicting comprehensibility and nativelikeness with Pillai scores for each vowel contrast, with a random intercept for L1. The results of the models are reported in Table 3 and indicate that the best predictors of comprehensibility were Pillai scores for /iː/ — /ɪ/, /ɑː/ — /æ/, and to a lesser extent /uː/ — /ʊ/.

Table 3. Output of the models predicting comprehensibility (above) and nativelikeness (below) on the basis of Pillai scores. Models were fitted separately, with one single fixed effect and with a random intercept for L1, e.g., *Comprehensibility ~ Pillai score for /iː/ — /ɪ/ + (1 | L1)*. Marginal R_2 values estimated the amount of variance accounted for by fixed effects, while conditional R_2 values refer to the amount of variance accounted for by fixed and random effects combined. Asterisks indicate statistical significance.

	/iː/ — /ɪ/	**/ɑː/ — /æ/**	**/ɜː/ — /ʌ/**	**/uː/ — /ʊ/**
		Models predicting comprehensibility		
p value for fixed effect	<0.001 ***	0.002 **	0.071	0.036 *
marginal R^2	0.252	0.22	0.097	0.1
Conditional R^2	0.423	0.452	0.203	0.236
		Models predicting nativelikeness		
p value for fixed effect	0.002 **	0.003 **	0.188	0.041 *
Marginal R^2	0.201	0.185	0.05	0.086
Conditional R^2	0.458	0.503	0.248	0.312

The same scenario applies to models of nativelikeness, where Pillai scores for /iː/ — /ɪ/ and /ɑː/ — /æ/ are the best predictors, followed by /uː/ — /ʊ/. Instead, Pillai scores for /ɜː/ — /ʌ/ were not significant predictors of comprehensibility or nativelikeness.

Combined models including all three significant Pillai scores were able to fit the data satisfactorily for comprehensibility (marginal R^2 = 0.394; conditional R^2 = 0.701) as well as nativelikeness (marginal R^2 = 0.335; conditional R^2 = 0.709). We also verified the relation of Pillai scores with fluency metrics. Pillai scores for /ɑ:/ — /æ/ and /u:/ — /ʊ/ were significant predictors of fluency metrics (all p values < 0.009 for predicting SR and AR).

3.2. Intrinsic Evaluation of French$_{L2}$ Vowels with Pillai Scores

A similar analysis was carried out for French$_{L2}$ data. Figure 3 shows Pillai scores obtained for every French$_{L2}$ learner and for target vowel contrasts involving front rounded vowels /y/ and /ø, œ/, plotted with native judgments of comprehensibility and nativelikeness for the same learners. Figure 4 illustrates vowel contrasts involving nasal vowels.

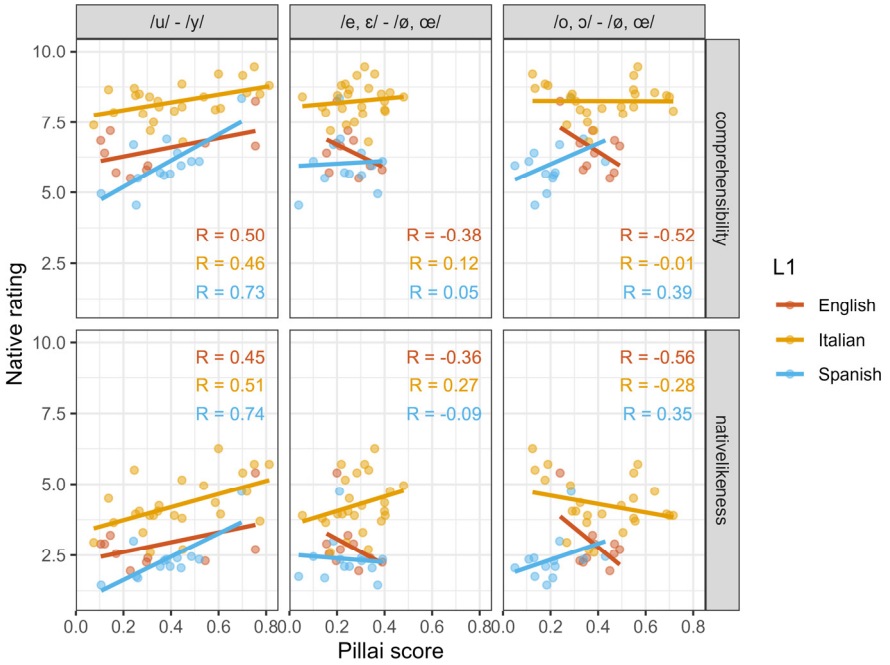

Figure 3. By−participant Pillai scores for French$_{L2}$ target vowel contrasts involving front rounded vowels, plotted with native ratings of comprehensibility (**above**) and nativelikeness (**below**).

The results are less clear-cut than for English$_{L2}$, and only the contrast /y/ — /u/ seems tightly related to native ratings. As in the previous analysis, we fitted linear mixed-effects models predicting comprehensibility and nativelikeness with each Pillai score, with a random intercept for L1. The results of the models are reported in Table 4 and confirm that the only relevant predictor of comprehensibility is the Pillai score for /y/ — /u/. Contrasts involving front mid vowels were treated in separate models for English$_{L1}$ vs. Spanish$_{L1}$ and Italian$_{L1}$ learners due to different predictions for these L1 groups; however, none of them seemed to reflect ratings of nativelikeness. We shall remind the reader that mid-close and mid-open vowels were conflated in the same category (see Section 2.3); however, we also ran the analysis with separate categories, and the results did not change. Among the contrasts involving nasal vowels, Pillai scores for /ɑ̃/ — /a/ and to a lesser extent /ɛ̃/ — /e, ɛ/ tend to significance as predictors of comprehensibility, and the chart shows that this trend is particularly strong for Spanish$_{L1}$ learners.

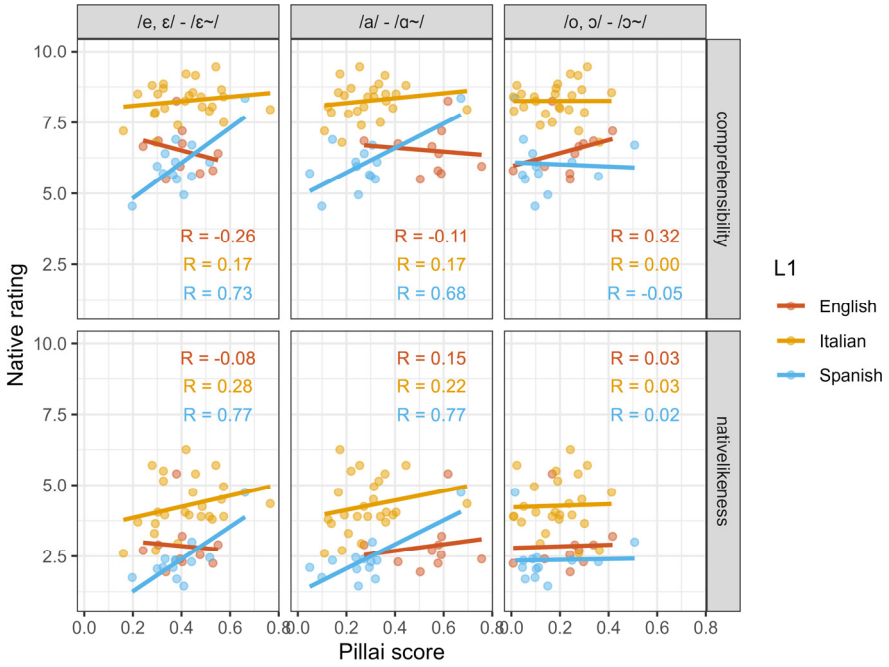

Figure 4. By−participant Pillai scores for French$_{L2}$ target vowel contrasts involving nasal vowels, plotted with native ratings of comprehensibility (**above**) and nativelikeness (**below**).

Table 4. Output of the models predicting comprehensibility (above) and nativelikeness (below) on the basis of Pillai scores. Models were fitted separately, with one single fixed effect and with a random intercept for L1, e.g.: *Comprehensibility ~ Pillai score for /y/ — /u/ + (1|L1)*. The model for /ø, œ/ — /o, ɔ/ only included English$_{L1}$ learners, so this was fitted as a simple linear model (adj. R^2 = 0.176 for comprehensibility; 0.229 for nativelikeness). Asterisks indicate statistical significance.

	/y/ - /u/	**/ø, œ/ - /o, ɔ/ (*Eng$_{L1}$*)**	**/ø, œ/ - /e, ɛ/ (*Sp$_{L1}$* & *It$_{L1}$*)**	**/ɛ̃/ - /e, ɛ/**	**/ɑ̃/ - /a/**	**/ɔ̃/ - /o, ɔ/**
			Models predicting comprehensibility			
p value for fixed effect	<0.001 ***	0.126	0.078	0.084	0.052	0.765
Marginal R^2	0.087	/	0.002	0.020	0.035	<0.001
Conditional R^2	0.749	/	0.803	0.703	0.75	0.69
			Models predicting nativelikeness			
p value for fixed effect	<0.001 ***	0.092	0.325	0.027 *	0.021 *	0.836
Marginal R^2	0.146	/	0.009	0.05	0.07	<0.001
Conditional R^2	0.624	/	0.668	0.56	0.649	0.523

A similar scenario applies for predictions of nativelikeness, where the Pillai score for /y/ — /u/ is the best predictor. In this case, Pillai scores for /ɛ̃/ — /e, ɛ/ and /ɑ̃/ — /a/ are also significant, although their contribution seems to be limited, with marginal R^2 values of 0.05 and 0.07, respectively.

3.3. Extrinsic Evaluation of French$_{L2}$ Vowels with Pillai Scores

Since French$_{L1}$ control data are available in the *AixOx* corpus, we used it to perform an extrinsic evaluation of French$_{L2}$ vowel realisations. Although extrinsic evaluation is usually achieved via measures other than Pillai scores, it is entirely possible to use Pillai scores to measure the amount of overlap between realisations for a given vowel by one (or more) L2 learner(s) and a cohort of L1 control speakers. In this case, a smaller amount of overlap should indicate a closer-to-target (i.e., more native-like) pronunciation. Therefore, we computed Pillai scores for realisations of the five marked French$_{L2}$ vowels (/y/, /ø/, /ɛ̃/,

/ɑ̃/, /ɔ̃/) by each L2 learner, and the corresponding vowels produced by the 10 French$_{L1}$ control speakers. Figure 5 illustrates the values obtained, plotted with native judgments of comprehensibility and nativelikeness for the same learners.

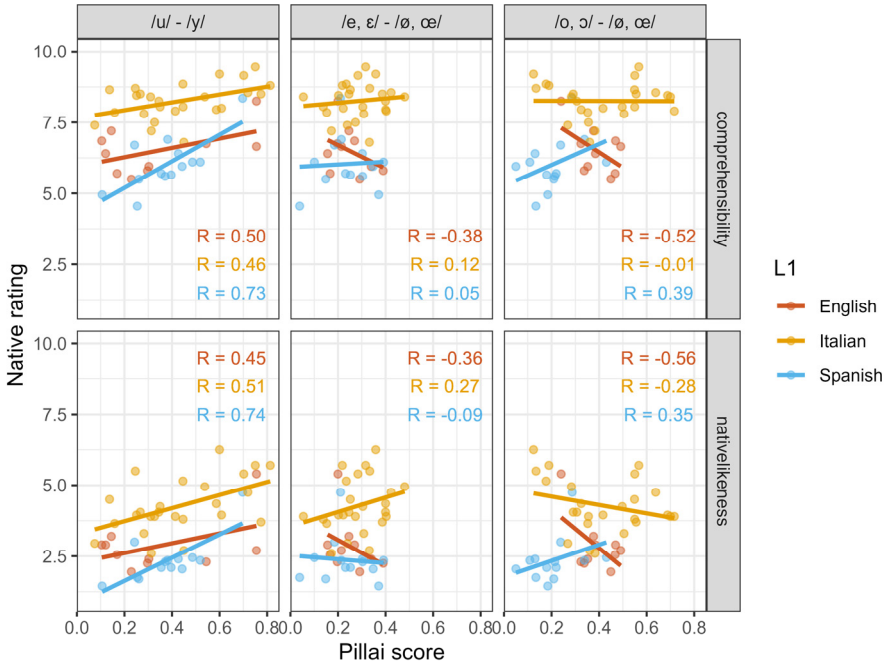

Figure 5. By-participant Pillai scores for French$_{L2}$ target vowel contrasts computed extrinsically (i.e., in reference to L1 data), plotted with native ratings of comprehensibility (**above**) and nativelikeness (**below**).

The results of the extrinsic evaluation do not seem to correlate with native ratings any better than those calculated intrinsically (see Section 3.2). It should be noted that we would expect negative correlations in this case, since lesser overlap (i.e., lower Pillai scores) should correspond to closer-to-target realisations (and therefore higher ratings). Yet, as illustrated in the charts, this is not always the case. Once more, we fitted linear mixed-effects models predicting comprehensibility and nativelikeness with each Pillai score, with a random intercept for L1. The results of the models are reported in Table 5, and show that no extrinsically computed Pillai score seems to be a significant predictor of comprehensibility or nativelikeness.

Table 5. Output of the models predicting comprehensibility (above) and nativelikeness (below) on the basis of Pillai scores calculated extrinsically. Models were fitted with one single fixed effect and with a random intercept for L1, e.g., *Comprehensibility ~ Pillai score for /i:/ + (1 | L1)*.

	/y/	**/ø/**	**/ɛ̃/**	**/ɑ̃/**	**/ɔ̃/**
			Models predicting comprehensibility		
p value for fixed effect	0.43	0.224	0.299	0.307	0.716
Marginal R^2	0.004	0.01	0.007	0.007	<0.001
Conditional R^2	0.694	0.704	0.703	0.695	0.693
			Models predicting nativelikeness		
p value for fixed effect	0.184	0.32	0.358	0.141	0.954
Marginal R^2	0.018	0.01	0.009	0.023	<0.001
Conditional R^2	0.542	0.538	0.54	0.542	0.523

4. Discussion

4.1. Considerations on the Intrinsic Evaluation of L2 Vowels

The results have shown that the intrinsic evaluation of vowels with Pillai scores correlates with native ratings of comprehensibility and nativelikeness for some but not all contrasts. For our English$_{L2}$ data, Pillai scores for /iː/ — /ɪ/, /ɑː/ — /æ/ and to a lesser extent /uː/ — /ʊ/ are significant predictors of comprehensibility and nativelikeness, accounting for a substantial amount of variance in native ratings, but not for /ɜː/ — /ʌ/. For our French$_{L2}$ data, Pillai scores for /y/ — /u/ are the only significant predictor of comprehensibility and nativelikeness, accounting for a substantial amount of variance in ratings, while Pillai scores for /ɛ̃/ — /e, ɛ/ and /ɑ̃/ — /a/ are significant predictors of nativelikeness but only marginal predictors of comprehensibility, and Pillai scores for /ø, œ/ — /o, ɔ/, /ø, œ/ — /e, ɛ/, /ɔ̃/ — /o, ɔ/ are not significant predictors.

The first question that arises from these results is whether native ratings are indeed an appropriate point of comparison for Pillai scores of vowel contrasts. Obviously, ratings of comprehensibility and nativelikeness are global scores that reflect a multitude of factors. It is well known that both comprehensibility and nativelikeness are complex and multi-faceted: they depend on many parameters, both at the segmental and suprasegmental levels (Munro and Derwing 1999; Crowther et al. 2022; Kang 2010; Isaacs and Trofimovich 2012; Saito et al. 2017; Saito 2021), and vowel accuracy is just one aspect of such complex constructs. Native ratings of comprehensibility and nativelikeness are global evaluations reflecting this multidimensional complexity, while Pillai scores for vowel contrasts provide a precise measure for a single specific phenomenon. We do not know how much weight native listeners attribute to vowel accuracy vs. other parameters when judging comprehensibility or nativelikeness, and this is likely to vary by listener, as well as by vowel: some speakers may be more sensitive to non-target-like prosody rather than segmental features, or some vowel contrasts may be more relevant than others in affecting native listeners' judgments. So, it would be unreasonable to expect measures of vowel overlap to explain a large amount of variance in global native ratings. On the other hand, various features may progress together during the process of L2 pronunciation acquisition: there is evidence that pronunciation training on some aspects leads to improvements in other aspects too, for example, training in prosody can bring improvements in segmental features (Dahmen et al. 2023). This means that, although vowel accuracy is merely one of many parameters that contribute to shaping listeners' impressions of comprehensibility or nativelikeness, it may itself be correlated to other segmental or suprasegmental parameters. So, with some caveats in mind, we think it is tenable to use global native ratings as a point of comparison for measures of L2 vowel pronunciation, as long as we do not expect them to explain all or most variance in native judgments.

This brings us to another question raised by our results, namely why certain vowel contrasts are significant predictors of comprehensibility and nativelikeness, and others are not. Although we do not have a definitive answer to this issue, it is interesting to note that vowel contrasts at the corners of the vowel chart (/iː/ — /ɪ/, /ɑː/ — /æ/, /uː/ — /ʊ/ for English$_{L2}$; /y/ — /u/, /ɑ̃/ — /a/ for French$_{L2}$) are better predictors than vowel contrasts involving mid or central vowels (/ɜː/ — /ʌ/ for English$_{L2}$; /ø, œ/ — /o, ɔ/, /ø, œ/ — /e, ɛ/, /ɛ̃/ — /e, ɛ/, /ɔ̃/ — /o, ɔ/ for French$_{L2}$). Moreover, among vowel contrasts at the corners of the vowel chart, those involving front vowels (/iː/ — /ɪ/, /ɑː/ — /æ/ in English$_{L2}$; /y/ — /u/ in French$_{L2}$) are more highly significant and account for a larger amount of variance than those involving mainly back vowels. The reason for this asymmetry is not clear, but some explanations are possible. Typological studies have revealed that vowel systems in the world's languages tend to be organised so that vowels are distributed preferably on the periphery of the vowel space, and more packed in the front than in the back of the oral cavity (see, for instance, Crothers 1978). The asymmetry in our results may have its roots in these universal trends, and may depend on a higher sensitivity on the part of listeners to vowels situated at the corners of the vowel space, and to the front vs. back. Alternatively, it may be that these vowels are where learners vary

most in pronunciation, thereby having greater consequences on comprehensibility and nativelikeness.

4.2. Intrinsic vs. Extrinsic Evaluation of L2 Vowels

Interestingly, none of the Pillai scores computed extrinsically to measure the overlap between L1 and L2 realisations of the same vowel in French$_{L2}$ were significant predictors of native judgments of comprehensibility or nativelikeness. This comes as a surprise, as we would expect more target-like vowels to reflect higher ratings of nativelikeness, despite the multifaceted nature of such ratings, as discussed above. We do not have a clear explanation for this. This may be because realisations that are not necessarily target-like but clearly distinct from other categories are perceived as sufficiently comprehensible and nativelike. Or it may be that consistency is more important than target-likeness, so that a speaker producing a vowel category as non-target-like but in a coherent way can be better rated than a learner producing some target-like and some non-target-like realisations for the same vowel. However, we find it encouraging to observe that Pillai scores calculated intrinsically are stronger predictors of comprehensibility and nativelikeness than Pillai scores calculated extrinsically. This is a sign that the intrinsic evaluation of L2 pronunciation is possible, and can provide new and different insights with respect to traditional L1-anchored methods.

4.3. Other Possible Methods for Intrinsic Evaluation

In this article, we have used Pillai scores to quantify the overlap between vowel categories as a measure of L2 pronunciation. As explained in Section 1, Pillai scores have some advantages over other statistics measuring distance and/or overlap across distributions. Yet, other approaches are conceivable. A plethora of simple and complex machine learning algorithms are used for the extrinsic evaluation of L2 pronunciation: models are trained on L1 speech and then used to classify L2 speech. But, it is equally possible to use machine learning techniques and apply the principles of intrinsic evaluation. For instance, Mairano et al. (2019) trained LDA (linear discriminant analysis) models on realisations of vowel pairs produced by each English$_{L2}$ learner, and then let the models re-classify the same L2 data: higher classification accuracy indicated more distinct realisations for different vowel categories. LDA is an extremely simple machine learning technique for supervised classification, but it is of course possible to extend the same principle and use more sophisticated algorithms. It is also conceivable to use techniques for unsupervised classification: we can imagine feeding realisations for vowel pairs produced by an L2 learner to a clustering algorithm, on the assumption that successful clustering (i.e., a clustering that is able to separate realisations for target vowel categories) will be indicative of more distinct L2 vowel realisations. One final word concerns the applicability of intrinsic evaluation to segments other than vowels, and to suprasegmental features. In this article, we focused on vowels as the simplest starting point, since the literature about vowels in L1 and L2 abounds, and acoustic correlates of vowels are widely known and easily measured. However, it is definitely possible to apply the same principles to L2 consonants or suprasegmental characteristics (but consider the limitations outlined in Section 4.4).

4.4. Caveats and Limitations of Intrinsic Evaluation

Algorithms of intrinsic evaluation such as the one presented in this paper have severe limitations. The first limitation is that they cannot be implemented within a CAPT (computer-assisted pronunciation training) system providing real-time feedback for single words: by definition, Pillai scores (and the underlying principle of intrinsic evaluation) compute the overlap between two distributions in a multidimensional space, so they need *distributions*, not single realisations. This brings us to the second limitation, which has to do with data availability: while extrinsic evaluation uses a previously trained model based on L1 data, intrinsic evaluation relies on L2 data produced by learners themselves, and therefore needs a certain number of realisations for each vowel category produced by each

learner; more precisely, it needs to evaluate two distributions for each learner, one for each vowel category.

Finally, the third and most problematic limitation of (raw) intrinsic evaluation algorithms is that they may positively evaluate realisations that are completely off target. For instance, let us consider the English vowel contrast /i:/ — /ɪ/; if a learner consistently realised this as [i:] — [ɑ:], the resulting Pillai score would be high since the two vowel categories are brightly distinct, but this would miss the fact that [ɑ:] is completely off-target and would certainly not be well accepted by listeners as a realisation of /ɪ/. Similarly, if the /i:/ — /ɪ/ contrast is consistently realised as [ɪ] for /i:/ and [i:] for /ɪ/, the Pillai score will again be high, despite the fact that realisations are inverted and therefore probably unacceptable by listeners. These examples are fairly unrealistic, but other cases may be more real and still problematic. Let us consider consonants, and more particularly voice onset time (VOT). It is well-established that English has a distinction between long-lag and short-lag VOT, while Romance languages have a distinction between negative and short-lag VOT (Lisker and Abramson 1964). Romance learners of English$_{L2}$ therefore tend to reinterpret the native (long-lag vs. short-lag) contrast in terms of L1 patterns (short-lag vs. negative VOT), and a raw intrinsic approach would consider this as acceptable, since the two phonological categories are realised as non-native-like, but neatly distinct. We think this limitation can be overcome by specifying some minimal characteristics for L2 realisations to be considered on target: in the /i:/ — /ɪ/ example, one may specify that only realisations in the upper-left quarter of the vowel space count as on target, and that /i:/ needs to be higher and/or more front than /ɪ/; in the VOT example, only positive VOT values may count as on target.

5. Conclusions

In summary, on the basis of our results, we find it justifiable to use measures of vowel overlap such as Pillai scores to evaluate vowel pronunciation in an L2 intrinsically. This allows researchers to assess features of L2 pronunciation without the need to refer to L1 data. Such measures are meant to reflect the development of phonological categories in learners' productions, and need to be applied to selected target contrasts. Apparently, the corners of the vowel chart, and more particularly the front axis, are preferable locations for the selection of target vowel contrasts, reflecting comprehensibility and nativelikeness.

That said, we do not claim that intrinsic evaluation is the ultimate solution for L2 pronunciation assessment, and we are fully aware of the limitations of this approach. We merely propound it as a viable alternative to avoid referring to an L1 model, in the spirit of considering the interlanguage as a system in its own right, which does not need to be evaluated by reference to a native model. It is important to stress that intrinsic evaluation via Pillai scores is not designed to reflect the target-likeness of vowel realisations. It is meant to reflect the development of phonological categories for new L2 sounds, and as such may be particularly relevant within acquisitional research. Some learners aim to sound native-like, and will prefer to be assessed in reference to a given L1 model. Intrinsic evaluation will not suit those learners, but (i) will possibly be fairer to learners who do not aim for a native-like pronunciation; (ii) will meet the practical needs of some L2 speech researchers having to assess specific phonetic features of L2 learners' speech, and not disposing of (or not wishing to refer to) comparable L1 data; (iii) may be useful within longitudinal studies analysing the effects of training on specific contrasts, as it will allow researchers to track the progressive development of new phonological categories in learners' productions.

Author Contributions: Conceptualization, P.M., F.S., and L.C.R.; methodology, P.M.; software, P.M.; validation, F.S. and L.C.R.; formal analysis, P.M.; investigation, P.M., F.S., and L.C.R.; resources, P.M., F.S., and L.C.R.; data curation, P.M., F.S., and L.C.R.; writing—original draft preparation, P.M.; writing—review and editing, P.M., F.S., and L.C.R.; visualization, P.M.; supervision, P.M.; project administration, P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study by the authors of each corpus.

Data Availability Statement: This study analyses data from pre-existing speech learner corpora. $English_{L2}$ data are part of the IPCE-IPAC corpus (https://www.pacprogramme.net/IPCE-IPAC), which is presently not freely available. $French_{L2}$ data are part of three various corpora, some of which made be made available by authors (see Section 2.2 for details about each corpus).

Acknowledgments: We are grateful to Nadine Herry-Bénit, Audrey Gros-Bonfiglioli and Ioana Trifu-Dejeu for sharing their data of 5 French participants with us. We also express our gratitude to the authors or the corpora used for analysis, and to Caroline Bouzon and Marc Capliez, who performed a manual segmentation verification for the $French_{L1}$ data of $English_{L2}$ published in our initial exploratory study.

Conflicts of Interest: The authors declare no conflict of interest.

References

Baills, Florence, Fabián Santiago, Paolo Mairano, and Pilar Prieto. 2022. The Effects of Prosodic Training with Logatomes and Prosodic Gestures on L2 Spontaneous Speech. Paper presented at Speech Prosody 2022, Lisbon, Portugal, May 23–26; pp. 802–6. [CrossRef]

Barreda, Santiago. 2015. PhonTools: Functions for Phonetics in R. R Package, Version 0.2-2.1. Available online: https://cran.r-project.org/web/packages/phonTools/citation.html (accessed on 20 November 2023).

Bartoń, Kamil. 2023. MuMIn: Multi-Model Inference. R Package, Version 1.47.5. Available online: https://CRAN.R-project.org/package=MuMIn (accessed on 20 November 2023).

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]

Bent, Tessa, and Ann R. Bradlow. 2003. The Interlanguage Speech Intelligibility Benefit. *The Journal of the Acoustical Society of America* 114: 1600–10. [CrossRef] [PubMed]

Best, Catherine T., and Michael D. Tyler. 2007. Nonnative and Second-Language Speech Perception: Commonalities and Complementarities. In *Language Learning & Language Teaching*. Edited by Ocke-Schwen Bohn and Murray J. Munro. Amsterdam: John Benjamins Publishing Company, vol. 17, pp. 13–34. [CrossRef]

Boersma, Paul, and David Weenink. 2023. Praat: Doing Phonetics by Computer [Computer Program] (Version 6). Available online: http://www.praat.org/ (accessed on 20 November 2023).

Cebrian, Juli. 2006. Experience and the Use of Non-Native Duration in L2 Vowel Categorization. *Journal of Phonetics* 34: 372–87. [CrossRef]

Cook, Vivian. 2016. *Second Language Learning and Language Teaching*, 5th ed. New York: Routledge. [CrossRef]

Crothers, John. 1978. Typology and Universals of Vowel Systems in Phonology. *Universals of Human Language* 2: 95–152.

Crowther, Dustin, Daniel Holden, and Kristen Urada. 2022. Second Language Speech Comprehensibility. *Language Teaching* 55: 470–89. [CrossRef]

Cucchiarini, Catia, Helmer Strik, and Lou Boves. 2000. Quantitative Assessment of Second Language Learners' Fluency by Means of Automatic Speech Recognition Technology. *The Journal of the Acoustical Society of America* 107: 989–99. [CrossRef]

Dahmen, Silvia, Martine Grice, and Simon Roessig. 2023. Prosodic and Segmental Aspects of Pronunciation Training and Their Effects on L2. *Languages* 8: 74. [CrossRef]

Darcy, Isabelle, Laurent Dekydtspotter, Rex A Sprouse, Justin Glover, Christiane Kaden, Michael McGuire, and John Hg Scott. 2012. Direct Mapping of Acoustics to Phonology: On the Lexical Encoding of Front Rounded Vowels in L1 English–L2 French Acquisition. *Second Language Research* 28: 5–40. [CrossRef]

De Jonge, Keryn, Olga Maxwell, and Helen Zhao. 2022. Learning on the Field: L2 Turkish Vowel Production by L1 American English-Speaking NGOs in Turkey. *Languages* 7: 252. [CrossRef]

Delais-Roussarie, Elisabeth, and Hiyon Yoo. 2010. The COREIL Corpus: A Learner Corpus Designed for Studying Phrasal Phonology and Intonation. Paper presented at 6th New Sounds, Poznan, Poland, May 1–3; pp. 100–5.

Delais-Roussarie, Elisabeth, Tanja Kupisch, Paolo Mairano, Fabian Santiago, and Frida Splendido. 2018. ProSeg: A Comporable Corpus of Spoken L2 French. Paper presented at EuroSLA 2018, Münster, Germany, September 5–8.

Escudero, Paola, and Daniel Williams. 2012. Native Dialect Influences Second-Language Vowel Perception: Peruvian versus Iberian Spanish Learners of Dutch. *The Journal of the Acoustical Society of America* 131: EL406–EL412. [CrossRef]

Escudero, Paola, Ellen Simon, and Holger Mitterer. 2012. The Perception of English Front Vowels by North Holland and Flemish Listeners: Acoustic Similarity Predicts and Explains Cross-Linguistic and L2 Perception. *Journal of Phonetics* 40: 280–88. [CrossRef]

Farrús, Mireia. 2023. Automatic Speech Recognition in L2 Learning: A Review Based on PRISMA Methodology. *Languages* 8: 242. [CrossRef]

Ferragne, Emmanuel, and François Pellegrino. 2010. Formant Frequencies of Vowels in 13 Accents of the British Isles. *Journal of the International Phonetic Association* 40: 1–34. [CrossRef]

Flege, James E. 1995. Second Language Speech Learning: Theory, Findings, and Problems. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Edited by Winifred Strange. Timonium: York Press, pp. 233–77.
Flege, James Emil, and Ocke-Schwen Bohn. 2021. The Revised Speech Learning Model (SLM-r). In *Second Language Speech Learning*, 1st ed. Edited by Ratree Wayland. Cambridge: Cambridge University Press, pp. 3–83. [CrossRef]
Flege, James Emil, Ian R. A. MacKay, and Diane Meador. 1999. Native Italian Speakers' Perception and Production of English Vowels. *The Journal of the Acoustical Society of America* 106: 2973–87. [CrossRef] [PubMed]
Flege, James Emil, Ocke-Schwen Bohn, and Sunyoung Jang. 1997. Effects of Experience on Non-Native Speakers' Production and Perception of English Vowels. *Journal of Phonetics* 25: 437–70. [CrossRef]
Fullana Rivera, Natalia, and Ian R. A. MacKay. 2003. Production of English Sounds by EFL Learners: The Case of /i/and/ɪ/. Paper presented at 15th International Congress of Phonetic Sciences, Barcelona, Spain, August 3–9; pp. 1525–28.
Gallant, Jordan. 2023. Typed Transcription as a Simultaneous Measure of Foreign-Accent Comprehensibility and Intelligibility: An Online Replication Study. *Research Methods in Applied Linguistics* 2: 100055. [CrossRef]
Gendrot, Cédric, Kim Gerdes, and Martine Adda-Decker. 2016. Détection Automatique d'une Hiérarchie Prosodique Dans Un Corpus de Parole Journalistique. *Langue Française* 191: 123–49. [CrossRef]
Georgiou, Georgios P. 2022. The Acquisition of /ɪ/−/I:/ Is Challenging: Perceptual and Production Evidence from Cypriot Greek Speakers of English. *Behavioral Sciences* 12: 469. [CrossRef] [PubMed]
Goldman, Jean-Philippe. 2011. Easyalign: An Automatic Phonetic Alignment Tool under Praat. Paper presented at Twelfth Annual Conference of the International Speech Communication Association (ISCA), Florence, Italy, August 27–31; pp. 3233–36. [CrossRef]
Herment, Sophie, Anne Tortel, Brigitte Bigi, Daniel J. Hirst, and Anastassia Loukina. 2014. AixOx, a Multi-Layered Learners Corpus: Automatic Annotation. In *Specialisation and Variation in Language Corpora*. Edited by Francisco Javier Diaz-Pérez and Ana Díaz-Negrillo. Bern: Peter Lang, pp. 41–76.
Herry-Bénit, Nadine, Stéphanie Lopez, Takeki Kamiyama, and Jeff Tennant. 2021. The *Interphonology of Contemporary English Corpus* (IPCE-IPAC). *International Journal of Learner Corpus Research* 7: 275–89. [CrossRef]
Isaacs, Talia, and Pavel Trofimovich. 2012. Deconsctructing Comprehensibility: Identifying the Linguistic Influences on Listeners' L2 Comprehensibility Ratings. *Studies in Second Language Acquisition* 34: 475–505. [CrossRef]
Isaacs, Talia, and Ron I. Thomson. 2013. 'Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions'. *Language Assessment Quarterly* 10: 135–59. [CrossRef]
Jenkins, Jennifer. 2000. *The Phonology of English as an International Language*. Oxford: Oxford University Press.
Jenkins, Jennifer. 2006. The Spread of EIL: A Testing Time for Testers. *ELT Journal* 60: 42–50. [CrossRef]
Kabakoff, Heather, Gretchen Go, and Susannah V. Levi. 2020. Training a Non-Native Vowel Contrast with a Distributional Learning Paradigm Results in Improved Perception and Production. *Journal of Phonetics* 78: 100940. [CrossRef]
Kang, Okim. 2010. Relative Salience of Suprasegmental Features on Judgments of L2 Comprehensibility and Accentedness. *System* 38: 301–15. [CrossRef]
Kang, Okim, and Lucy Pickering. 2013. Acoustic and Temporal Analysis for Assessing Speaking. In *The Companion to Language Assessment*. Edited by Antony John Kunnan. Hoboken: John Wiley & Sons, Inc., pp. 1047–62. [CrossRef]
Kang, Okim, Ron I. Thomson, and Meghan Moran. 2018. Empirical Approaches to Measuring the Intelligibility of Different Varieties of English in Predicting Listener Comprehension: Measuring Intelligibility in Varieties of English. *Language Learning* 68: 115–46. [CrossRef]
Kartushina, Natalia, and Ulrich H. Frauenfelder. 2014. On the Effects of L2 Perception and of Individual Differences in L1 Production on L2 Pronunciation. *Frontiers in Psychology* 5: 1246. [CrossRef]
Kelley, Matthew C., and Benjamin V. Tucker. 2020. A Comparison of Four Vowel Overlap Measures. *The Journal of the Acoustical Society of America* 147: 137–45. [CrossRef] [PubMed]
Kisler, Thomas, Uwe Reichel, and Florian Schiel. 2017. Multilingual Processing of Speech via Web Services. *Computer Speech & Language* 45: 326–47. [CrossRef]
Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. LmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82: 1–26. [CrossRef]
Leppik, Katrin, Pärtel Lippus, and Eva Liina Asu. 2023. The Perception and Production of Estonian Vowels and Vocalic Quantity Contrasts by Spanish L1 Learners. *Ampersand* 11: 100147. [CrossRef]
Liakin, Denis, Walcir Cardoso, and Natallia Liakina. 2015. Learning L2 Pronunciation with a Mobile Speech Recognizer: French/y/. *Calico Journal* 32: 1–25. [CrossRef]
Lisker, Leigh, and Arthur S. Abramson. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. *Word* 20: 384–422. [CrossRef]
Mairano, Paolo, Caroline Bouzon, Marc Capliez, and Valentina De Iacovo. 2019. 'Acoustic Distances, Pillai Scores and LDA Classification Scores as Metrics of L2 Comprehensibility and Nativelikeness'. Paper presented at 19th International Congress of Phonetic Sciences, Melbourne, Australia, August 5–9; pp. 1104–8. Available online: https://www.icphs2023.org/programme/proceedings/ (accessed on 20 November 2023).
Mairano, Paolo, and Fabian Santiago. 2020. What Vocabulary Size Tells Us about Pronunciation Skills: Issues in Assessing L2 Learners. *Journal of French Language Studies* 30: 141–60. [CrossRef]

Mairano, Paolo, Leonardo Contreras Roa, Marc Capliez, and Caroline Bouzon. 2021. 'The /s/ ~ /z/ Voice Contrast in L1 French, L1 Spanish and L1 Italian Learners of L2 English'. *Language, Interaction and Acquisition* 12: 210–50. [CrossRef]
Marinescu, Irina. 2013. Native Dialect Effects in Non-Native Production: Cuban and Peninsular Learners of English. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique* 58: 415–41. [CrossRef]
Méli, Adrien, and Nicolas Ballier. 2019. Analyse de La Production de Voyelles Anglaises Par Des Apprenants Francophones, l'acquisition Du Contraste /ɪ/–/I:/ à La Lumière Des k-NN. *Anglophonia* 27: 1–16. [CrossRef]
Melnik-Leroy, Gerda Ana, Rory Turnbull, and Sharon Peperkamp. 2022. On the Relationship between Perception and Production of L2 Sounds: Evidence from Anglophones' Processing of the French /u/–/y/ Contrast. *Second Language Research* 38: 581–605. [CrossRef]
Munro, Murray J., and Tracey M. Derwing. 1999. Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning* 49: 285–310. [CrossRef]
Neri, Ambra, Catia Cucchiarini, and Helmer Strik. 2008. The Effectiveness of Computer-Based Speech Corrective Feedback for Improving Segmental Quality in L2 Dutch. *ReCALL* 20: 225–43. [CrossRef]
Nycz, Jennifer, and Lauren Hall-Lew. 2013. Best Practices in Measuring Vowel Merger. *The Journal of the Acoustical Society of America* 134: 4198. [CrossRef]
Perry, Scott James, and Benjamin V. Tucker. 2019. L2 Production of American English Vowels in Function Words by Spanish L1 Speakers. *Canadian Acoustics* 47: 94–95.
Pillot-Loiseau, Claire, and Martina Grando. 2020. Apport des comptines pour la prononciation du/y/français chez des enfants italophones: Une étude perceptive pilote. Paper presented at Actes de la 6e conférence conjointe JEP TALN RÉCITAL, Nancy, France, June 8–19; vol. 1, pp. 507–15.
Racine, Isabelle, and Sylvain Detey. 2018. Production of French Close Rounded Vowels by Spanish Learners: A Corpus-Based Study. In *Romance Phonetics and Phonology*. Oxford: Oxford University Press, pp. 381–94. [CrossRef]
Rathcke, Tamara, Jane Stuart-Smith, Bernard Torsney, and Jonathan Harrington. 2017. The Beauty in a Beast: Minimising the Effects of Diverse Recording Quality on Vowel Formant Measurements in Sociophonetic Real-Time Studies. *Speech Communication* 86: 24–41. [CrossRef]
R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 20 November 2023).
Ruellot, Viviane. 2011. Computer-Assisted Pronunciation Learning of French/u/and/y/at the Intermediate Level. Paper presented at 2nd Pronunciation in Second Language Learning and Teaching Conference, Ames, IA, USA, September 10–11; vol. 2, pp. 199–213.
Saito, Kazuya. 2021. What Characterizes Comprehensible and Native-like Pronunciation Among English-as-a-Second-Language Speakers? Meta-Analyses of Phonological, Rater, and Instructional Factors. *TESOL Quarterly* 55: 866–900. [CrossRef]
Saito, Kazuya, Pavel Trofimovich, and Talia Isaacs. 2017. Using Listener Judgments to Investigate Linguistic Influences on L2 Comprehensibility and Accentedness: A Validation and Generalization Study. *Applied Linguistics* 38: 439–62. [CrossRef]
Santiago, Fabián, and Paolo Mairano. 2021. 'La Prononciation Des Voyelles /e/, /ɛ/, /ə/, /ø/, /œ/ En FLE Chez Les Hispanophones et Le Rôle de l'orthographe'. In *La Prononciation Du Français Langue Étrangère: Perspectives Linguistiques et Didactiques*. Edited by Elisa Pustka. Romanistische Fremdsprachenforschung Und Unterrichtsentwicklungen. Tübingen: Narr, pp. 113–32.
Simon, Ellen, Mathijs Debaene, and Mieke Van Herreweghe. 2015. The Effect of L1 Regional Variation on the Perception and Production of Standard L1 and L2 Vowels. *Folia Linguistica* 49: 521–53. [CrossRef]
Strik, Helmer, Khiet Truong, Febe De Wet, and Catia Cucchiarini. 2009. Comparing Different Approaches for Automatic Pronunciation Error Detection. *Speech Communication* 51: 845–52. [CrossRef]
Limesurvey Project Team. 2012. Limesurvey: An Open Source Survey Tool. Available online: http://www.limesurvey.org (accessed on 20 November 2023).
Tejedor-Garcia, Cristian, David Escudero-Mancebo, Valentin Cardenoso-Payo, and Cesar Gonzalez-Ferreras. 2020. Using Challenges to Enhance a Learning Game for Pronunciation Training of English as a Second Language. *IEEE Access* 8: 74250–66. [CrossRef]
Valenzuela Farias, Maria Gabriela. 2022. Developing English Vowel Contrasts: An Analysis of Spanish L1 Learners' Speech Production over Time in the UK. Ph.D. thesis, University of York, Heslington, UK.
Witt, Silke M. 2012. Automatic Error Detection in Pronunciation Training: Where We Are and Where We Need to Go. Paper presented at International Symposium on Automatic Detection on Errors in Pronunciation Training, Stockholm, Sweden, June 6–8; vol. 1, pp. 1–8.
Witt, Silke M., and Steve J. Young. 2000. Phone-Level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication* 30: 95–108. [CrossRef]
Xi, Xiaoming. 2010. Special Issue: Automated Scoring and Feedback Systems for Language Assessment and Learning. *Language Testing* 27: 291–440. [CrossRef]
Yoon, Su-Youn, Mark Hasegawa-Johnson, and Richard Sproat. 2010. Landmark-Based Automated Pronunciation Error Detection. Paper presented at Eleventh Annual Conference of the International Speech Communication Association (ISCA), Chiba, Japan, September 26–30; pp. 614–17. [CrossRef]