ALIGNMENT IN ASR AND L1 LISTENERS' RECOGNITION OF L2 LEARNER SPEECH: FRENCH EFL LEARNERS & DICTATION.IO

VINCENT CHANETHOM
Princeton University;
vc4@princeton.edu

ALICE HENDERSON
Université Grenoble-Alpes
alice.henderson@univ-grenoble-alpes.fr

Abstract:
This study is an extension of Inceoglu et al.'s (2023) study on *Google Voice Typing* as a pronunciation learning tool. We used the Automatic Speech Recognition (ASR) tool on the *dictation.io* website (Agarwal, 2022), and our participants were L2 English learners of a different L1, but similar proficiency level. Twelve L1 English listeners assessed the L2 English from four L1 French speakers in terms of intelligibility and comprehensibility, measured by word transcription and Likert scale ratings respectively. Their scores were compared to ASR output. The goal was to determine how accurate the tool is, and to what extent its accuracy correlates with human listeners. The results were generally consistent with those of Inceoglu et al. (2023), with few exceptions which we discuss in the current study.

Key words: English, Automatic Speech Recognition, L2 learner speech, replication, intelligibility, comprehensibility

1. Introduction

The development of "easy-to-use and useful software", called for by Derwing (2010, p. 30) is underway, as websites, mobile apps and Massive Open Online Courses (MOOC) have become much more widely used by teachers and learners, inside and outside the classroom. This is due in part to Automatic Speech Recognition (ASR) being a built-in feature of many free programs (e.g., online translators, voice-activated web search, Global Positioning System apps) and ASR's continued improvement (Levis & Suvorov, 2020; McCrocklin et al. 2019). Researchers have explored its potential for L2 pronunciation learning, including speech production, speech perception and attitudes towards ASRs. For example, Liakin et al. (2017) examined learner perceptions of ASR for learning a vowel and a suprasegmental feature in L2 French. Inceoglu et al. (2020) also investigated the beliefs of Korean learners of English about ASR's general usefulness and for learning vowel contrasts. In a more recent study, Inceoglu et al. (2023) examined how *Google Voice Typing* performs compared to native listeners in recognising Taiwanese learners' English. Working with Brazilian Portuguese and Spanish speakers learning English and using *MicrosoftWord* and *VoiceNotebook*, Kivistö de Souza and Gottardi (2022) highlight the advantages in terms of increased exposure and output possibilities, whereby learners are more likely to notice the difference between their speech and a target. Such dictation and transcription also provide a simple measure of how intelligible a speaker is.

Derwing and Munro (1995a, 1995b) proposed three key concepts for the study of pronunciation proficiency, the first one being intelligibility, a measure of what is actually understood when someone speaks. The other two are subjective perceptions, frequently evaluated via Likert scales: accentedness (the degree to which a speaker diverges from a target accent) and comprehensibility (the amount of effort a listener exerts to understand someone). These do correlate,[1] the crucial finding being that one can be intelligible even with marked accentedness. Contrasting the intelligibility ratings of trained listener-raters (similar to a language teacher who is familiar with a certain accent or language) and ASR, exploits the technology's "capabilities for verbatim transcriptions to simulate a naïve listener's understanding of nonnative speech and to serve for the evaluation of intelligibility" (Mroz, 2018, p. 18). This could be especially pertinent for lower proficiency levels, given that Moussalli and Cardoso (2020) found that Amazon's *Echo* adapts well to different degrees of accented speech.

In terms of language instruction, ASR tools can help to promote awareness of these three dimensions of pronunciation proficiency, especially when a teacher can help to decipher their learners' L2 speech output and provide actionable feedback. A key determiner of any technology's usefulness for language learning is its feedback potential, whether it merely indicates that a spoken production is correct or incorrect (binary feedback) or provides targeted feedback, information which is both specific and actionable (Henrichsen, 2021). This is crucial for pronunciation learning because overall, ASR's recognition of non-native speech is not always accurate, even though the accuracy of some ASR output has improved for native speech, for example from 18-20% inaccurate to only 3-5% for *Google Voice Typing* (McCrocklin & Edalatishams, 2020). Nonetheless, while poor accuracy rates may frustrate learners, Golonka et al. (2014) found that speaking with a computer can facilitate willingness to speak the target language with other people. Therefore, the accuracy of different tools as used by different

[1] Comprehensibility shows a moderate-to-strong correlation with accentedness (Munro & Derwing, 1995b).

learners needs to be investigated. For example, McCrocklin and Edalatishams (2020) found no significant correlations between the accuracy of *Google’s* ASR output for L1 Spanish learners and measures of recognition, comprehensibility and accentedness. French-accented English may pose challenges similar to Spanish-accented English, despite the different segmental inventories of the two Romance languages. We therefore chose to extend Inceoglu et al.’s (2023) study, by having participants from a Romance language (L1 French) - thus complementing their findings with non-Romance language participants (L1 Taiwanese). Similar to their study, our participants were at a not-advanced proficiency level. Moreover, their methodology was suitable to the ASR tool our students had suggested (they liked the interface and were happy that no set up or app installation was required). Additionally, we decided to use dictation.io because it forced the students to indicate a variety of English as their 'native' language (i.e. the English the ASR system would try to decode), and we felt that being obliged to choose raised their awareness of this diversity.

In their study, Inceoglu et al. (2023) contrasted the output of *Google’s* ASR dictation system with 12 native English listeners’ transcriptions of 48 words and 24 sentences, as read by 4 Taiwanese EFL English L2 learners at intermediate level. The isolated target words in their stimuli specifically targeted the vowel contrasts /iː-ɪ/ (tense/lax distinction as in beat vs. bit) and /æ-ɛ/ (low/mid front vowel as in bat vs. bet). Their results indicated that overall, their speakers received lower intelligibility scores from both ASR and the L1 listener-based assessments for the word task (40.81% vs. 38.62%, respectively) than the sentence task (75.52% vs. 83.88%, respectively). The proportion of recognised words was larger for the listener-based assessments than ASR for the sentence task, which was the opposite for the isolated words. They also found similarities between ASR and the L1 listeners with respect to the type of errors identified in the speaker’s production. The most common error was a single vowel substitution, followed by the combination of one vowel and one consonant substitution. The third error type was a single consonant substitution. However, despite those similarities, significant weak and moderate positive correlations between ASR output and their L1 listeners’ transcriptions for both word and sentence tasks respectively were only found for one speaker. Thus, they concluded that agreement between ASR and human raters is highly dependent on individual speakers.

Similarly, the objective of the current study was to examine the ability of ASR in *dictation.io* (Agarwal, 2022) to assess the intelligibility of French EFL learners of a low-intermediate proficiency level. To do so, the ASR output is compared to L1 English listeners’ transcriptions and comprehensibility ratings of L2 learners’ productions of both isolated words and sentences. The following research questions (RQ) were posed:

RQ1: How (mis)aligned are ASR outputs from *dictation.io* and L1 listeners' transcriptions?

RQ2: Does the accuracy of ASR outputs from *dictation.io* for L2 speech correlate with human listener recognition (intelligibility) and with their ratings of comprehensibility?

RQ3: What phonetic contrast is associated with the most errors in the ASR output from *dictation.io* and how do those error patterns compare with L1 listener-based assessments?

With respect to RQ1, we hypothesise that the ASR output from *dication.io* will show similar patterns as L1 listener-based assessments in terms of both intelligibility scores and error types recognised. However, we also predict ASR's scores will be lower than those of L1 listeners for the sentence task, as in Inceoglu et al. (2023), possibly due to L1 listeners' exploitation of sentence context. For RQ1, we also hypothesise low correlations between the intelligibility scores of ASR and those of the L1 listeners in our study. As found in Inceoglu et al. (2023), the correlation patterns between the ASR and L1 listeners will vary greatly depending on the speaker and on the task condition (i.e., isolated words vs. sentences). As for RQ2, while research has not yet clarified correlations between intelligibility and comprehensibility, research has shown a weak correlation between accentedness and intelligibility (Munro & Derwing, 1995a; Jułkowska & Cebrian, 2015) and a moderate-to-strong correlation between accentedness and comprehensibility (Munro & Derwing, 1995b). We nonetheless hypothesise that L1 listeners' comprehensibility scores may follow similar patterns as their intelligibility scores and those of the ASR output. Finally, regarding RQ3, we hypothesise that the error type analysis will yield similar results for both the ASR output and L1 listener-based assessments, with a single vowel substitution being the most common error type followed by the combination of a vowel and a consonant, and finally the substitution of a single consonant.

2. Methodology

2.1. Participants: L1 listeners and L2 speakers

To test the hypotheses laid out in the previous section, we recruited 12 listeners who were native anglophones, half of whom were based in France and the other half in the US (see Table 1). Their age ranged from 28 to 65 years old (M = 48.92 years; SD = 11.16 years). Overall, they indicated on a questionnaire, which was part of the listening experiment, that they were very familiar with both spoken French and French-accented English. They also self-reported that they were good spellers. Out of the 12 L1 listeners, 10 were multilingual.

Table 1: Demographic information and language background for English L1 listeners

L1 Listeners (n = 12)	
Age	*M* = 48.92; *SD* = 11.16
Gender	8 F + 4 M
Place of birth	7 US + 1 CAN + 4 GB
Current location	7 US-based + 7 France-based
Languages	10 multiling + 2 English monoling
Familiarity with spoken French*	*M* = 6.88; *SD* = 2.85
Familiarity with French-accented English*	*M* = 6.67; *SD* = 2.12
Spelling competence*	*M* = 7.63; *SD* = 2. 39

*Self-reported scores based on a 1–9-point Likert scale (1 = not at all familiar; 9 = extremely familiar)

The L2 speakers were recruited from a group of 21 undergraduate students (12 females, 9 males; 20-27 years old) taking an obligatory English for Specific Purposes course as part of a food sciences degree at a French university. Those with incomplete data or poor-quality recordings were excluded from the study, and from the remaining 12 data sets 2 female and 2 male voices were used.

The 4 L1 French students indicated having studied English for 8-10 years in school. None had lived in an English-speaking country or used English during lengthy trips (more than 2 weeks), and none had a bilingual French-English upbringing. Their English listening proficiency was evaluated via DIALANG (2022) at A1 for two of them and A2 for the other two, the lowest bands of the Common European Framework of Reference.

2.2. Stimuli and speech recording procedure

The stimuli for this study were speech samples collected as part of a university English course using word- and sentence-reading tasks.[2] The course involves 28 hours of in-person classes held irregularly from October to May (6 months; weeks 1-31), with 18 hours concentrated between mid-January and mid-March. An online platform houses autonomous work in between classes, each of which lasted 3-3.5 hours. This degree involves an obligatory 6-month internship where students often have to interact with both native and non-native English speakers, so one of the main course objectives was to increase confidence in speaking

[2] Recordings of three short texts read aloud were not analysed for the current study, nor was spontaneous speech elicited through questions at the end of each worksheet, about favourite foods and holiday plans.

English by improving intelligibility. Table 2 details the timing of the ASR work, showing its concentration in the second half of the academic year (weeks 15-23):

Table 2: Timing of ASR work

Class Sessions (weeks)	Relevant instructional Content
Class 4 (wk15)	ASR work is explained and practiced
Online	Deadline 24 hours: upload ASR Week 0
	Deadline 10 days: upload ASR Week 1
Class 5 (wk17)	Group feedback on ASR work
Online	Deadline 1 week: upload ASR Week 2
	Deadline 2 weeks: upload ASR Week 3
Class 6 (wk23)	Group feedback on ASR work

Attention was explicitly given to pronunciation through proactive and reactive group feedback in Class 5 and 6. Students also received individualised feedback via email, praising what was recognised correctly by the ASR and explaining mismatches. Homework assignments always required students to prepare texts or videos for discussion in the following class, but they also included pronunciation practice with ASR technology.

In Class 4, students began the ASR work. They used the voice recording app of their choice on their mobile phone or used computer-based software (e.g. *Audacity, Praat*) to record themselves while speaking into the *dictation.io* website, which does not record audio but generates text based on their speech. They recorded a total of 76 words in isolation (or 38 minimal pairs), 10 sentences, and 3 short texts, which were distributed over three worksheets (Week 1-3). A Week 0 was also provided as a training session to help students become familiar with the ASR tool and the recording procedure. The 76 individual words, which were all monosyllabic, featured three vowel and two consonant contrasts (/iː/-/ɪ/, /æ/-/ʌ/, /ɛ/-/eɪ/, /ʧ/-/ʃ/, and /s/-/θ/) that are recognised as challenging for French L1 speakers (Swan & Smith, 2001). To the extent possible, the target vowels were organised in 3 categories: following a word-initial stop consonant, a word-initial fricative, and a word-initial approximant. The target consonants were organised in 2 categories: word-initial or word-final positions. Within each category for the target consonants, the vowel context varied in terms of height (high, mid, and low vowels). However, as was the case in Inceoglu et al. (2023), the current analysis does not include the factor of phonetic context, which we leave for future studies. Table 3 summarises the target words which were included in the current study and which represent the vowel and consonant contrasts of interest.

Table 3: Target word list (n = 76)

Vowel contrasts (n = 36)		
/iː/ - /ɪ/	/æ/ - /ʌ/	/ɛ/ - /eɪ/
beat – bit	bag – bug	peg – page
deep – dip	tack – tuck	bet – bait
seat – sit	sack – suck	shed – shade
sheep – ship	shack – shuck	fed – fade
leak – lick	lag – lug	let – late
week – wick	yak – yuck	we – wait

Consonant contrasts (n = 40)			
/ʧ/ - /ʃ/		/s/ - /θ/	
word-initial	*word-final*	*word-initial*	*word-final*
cheat – sheet	ditch – dish	sin – thin	pass – path
chip – ship	catch – cash	sigh – thigh	face – faith
chore – shore	latch – lash	some – thumb	miss – myth
cheer – sheer	butch – bush	sank – thank	moss – moth
chair – share	leech – leash	seem – theme	worse – worth

The sentences used in this study contained words that illustrated the same vowel and consonant contrasts, and were designed so that listeners could not rely much on context. To avoid familiarity effects on L2 speaker’s production and L1 listeners’ perception, the target words used in the sentences were different from the words in isolation listed in Table 3. As was the case in Inceoglu et al. (2023), speakers and listeners in this study were exposed to each target word only once (in isolation or in the sentences). Table 4 lists those 10 sentences, which varied from 6-11 words in length (*M* = 9.3 words).

Table 4: Sentence list

Labels	Sentences	Word count
s1	**Pick** up this **green pin** and **stick** it into the **peach**.	11
s2	**Sally thought** about **some things** on the **path**.	8
s3	There is a **big piece** of **cheese** in the kitchen **fridge**.	11
s4	**Stack** the **cups** and the **damp jug** on the other **rack**.	11
s5	**Shaun chose fish** and **chips** instead of a **chop** and **mash**.	11
s6	The **duck swam** up to the **rat**.	7
s7	Last **month Sue saved three pence**.	6
s8	The **child** didn’t **choose** to **share** her **chips**.	8
s9	The **planes** and **elms** were **blamed** for many **wrecks**.	11
s10	They **stayed** and **spread** out to **play** the **game**.	9
	Mean	9.3

To encourage their autonomy, learners worked at home and they could spend as much or as little time as they liked on the exercise. Both the *dictation.io*-generated text file and the sound file were uploaded onto an institutional platform, so that the teacher could monitor the feedback learners received from the ASR tool. A group debriefing was only provided twice (Class 5 and 6) about how to interpret the ASR’s transcription and to ‘work around’ its limitations, but the teacher also provided individualised feedback via email, to explain the ASR output.

2.3. Speech rating procedure

To create the stimuli, two female and two male voices were selected from the 12 complete data sets, as being the best quality recordings and as representative of productions with the ‘best’ and ‘worst’ pronunciation. We chose to use only 2 voices of each gender, as it was easier to rate them at the extremes of a continuum, instead of trying to rate several voices along the continuum. We felt this would give the ASR tool an opportunity to ‘succeed’ and to ‘fail’.

The target words and sentences were extracted into individual audio files which were then normalised for peak intensity and padded with 500 ms of silence before and after each speech sample using Praat (Boersma & Weenink, 2022). In the case of multiple attempts, only the first production of the stimulus was selected. Alternatively, one of the subsequent attempts was chosen if the recording quality of the first one was not adequate (e.g., background noise, coughing).

The extracted stimuli were then presented to L1 listeners in randomised order. Only 1 exposure was provided for each word and sentence, i.e., listeners never heard the same word or sentence twice with different voices, in order to avoid familiarity effects on the stimuli.

The listening test was administered locally via PsychoPy (Peirce, 2019) in four blocks: (1) the demographic questionnaire which included language familiarity questions, (2) the sentence listening task, (3) the first half of the word listening task, and (4) the second half of the word listening task. Participants were given the opportunity to take a short break between each block. To test for intelligibility, listeners were asked to use standard English orthography to transcribe the words or sentences they heard, which were then compared to the ASR output. In addition to the transcriptions, listeners were asked to rate the sentences on a 9-point Likert-scale for comprehensibility, operationalised as the amount of effort required to understand.

2.4. Data coding and analysis

For each sentence, intelligibility scores were calculated as the percentage of correct words transcribed over the total number of words in the target sentence. A mean intelligibility score was then calculated for each speaker by averaging the scores obtained for each sentence and across all listeners. Comprehensibility ratings, on the other hand, were based on a 9-point Likert scale in reply to the question *How much effort did you make to understand the sentence?* (1 =*It was effortless.* and 9 = *I had to try extremely hard*.).

For the word task, intelligibility scores were based on matches between the target word and transcription. A score of 1 was awarded for exact matches, whereas transcriptions with errors were scored zero. Those words which scored zero were then analysed for error type, following Inceoglu et al.'s (2023) coding system as shown in Table 5. Statistical analyses were carried out in R (R Core Team, 2022).

Table 5: Code for error type[3]

0. Whole word missing or not transcribed	h. One consonant + extra segment(s)
a. One vowel	i. One vowel + extra segment(s)
b. One consonant	j. One vowel + one consonant + extra segment(s)
c. One vowel + one consonant	k. One vowel + one consonant + missing consonant
d. Two consonants	l. One vowel + two consonant + missing consonant
e. One vowel + two consonants	m. Missing segment
f. One vowel + missing consonant	n. Two consonant + extra segment(s)
g. Extra segment	

3. Results

3.1. Intelligibility

3.1.1. Word task

As done in Inceoglu et al. (2023), interrater reliability for the 12 L1 listeners was determined using Fleiss Kappa on the word intelligibility task. Based on Landis and Koch's (1977) Kappa coefficient classification table, the test revealed an overall moderate inter-rater agreement for all four speakers (κ = .461, z = 309, *p* < .001). For each speaker individually, the level of agreement between the listeners was fair for speakers SP07 (κ = .244, z = 53.8, *p* < .001) and SP15 (κ = .390, z = 74.0, *p* < .001), moderate for speaker SP06 (κ = .561, z = 79.9, *p* < .001), and substantial for speaker SP14 (κ = .606, z = 89.4, *p* < .001).

Pairwise comparisons using unweighted Cohen's Kappa were also performed, not only to determine the degrees of interrater reliability within each pair of L1 listeners, but also to examine how each rater compared to ASR. The results yielded interrater agreements that ranged from fair (κ = .282) to substantial (κ = .654). The lowest kappa coefficient was obtained for the agreement between ASR and one of the L1 listeners (R1), whereas the highest coefficient was associated with the comparison between two human raters (R8 and R12). The results indicated that only two L1 listeners (R9 and R10) showed moderate agreement with ASR, while the 10 others showed fair agreement with it. In the great majority, however, the inter-rater agreement among human listeners was moderate (50 instances). Only 15 comparisons within the group of L1 listeners showed fair

[3] Error types k-n were added here because the types of errors we encountered differed to those with the Taiwanese learners involved in Inceoglu et al. (2023). Logically, if the L1 differs, types and proportions of errors will differ.

agreement. The Kappa coefficients from the pairwise comparisons are summarised in Table 6.

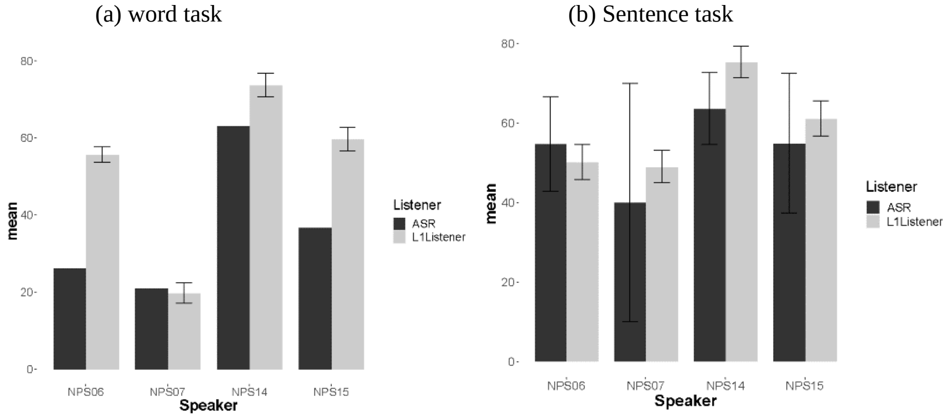
Table 6: Pairwise comparisons for interrater reliability (unweighted Cohen's Kappa) for word intelligibility task

	ASR	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
R1	.284											
R2	.375	**.401***										
R3	.343	.398	.396									
R4	.365	**.404***	**.482***	.399								
R5	.324	**.430***	.374	.384	.390							
R6	.282	**.462***	.376	**.416***	.379	**.434***						
R7	.365	**.482***	**.495***	**.454***	**.522***	**.469***	.393					
R8	.370	**.596***	**.598***	**.503***	**.516***	**.449***	**.440***	**.582***				
R9	**.481***	.329	**.530***	.362	**.516***	.382	.316	**.449***	**.468***			
R10	**.502***	**.416***	**.495***	**.412***	**.522***	**.455***	.379	**.561***	**.555***	**.542***		
R11	.325	**.483***	**.415***	**.454***	.391	**.496***	**.462***	**.443***	**.556***	.396	**.403***	
R12	.362	**.563***	**.537***	**.503***	**.441***	**.495***	**.503***	**.550***	.654*	**.407***	**.536***	**.537***

Cohen's kappa strength: fair (), **moderate (*)** and substantial (*) based on Landis and Koch (1977)

For the word transcription, the results showed that L1 listeners recognised a greater proportion of words produced by the four L2 speakers (52.19 %) than ASR (36.84%), in contrast to Inceoglu et al.'s (2023) findings, in which ASR recognised more words than their L1 listeners. This finding was further confirmed by analysis of each speaker. All L1 listeners attributed higher scores than ASR to all speakers, except for speaker SP07 for whom there was a slight reversal (ASR: 21.05%; L1 listeners: 19.74%). However, there was an agreement between ASR and the L1 listeners regarding which speakers were the most and least intelligible. SP14 received the highest scores of words recognised, whereas SP07 received the lowest scores. The mean percents and standard deviations for recognised words are provided in Table 7. The comparisons between ASR and L1 listeners for the word intelligibility task are also illustrated in Figure 1.

Figure 1: Mean intelligibility scores (%) by speaker for word and sentence tasks



(a) word task

(b) Sentence task

Table 7: Mean percent (M) and standard deviation (SD) for number of words recognised by ASR and L1 listeners for word and sentence tasks

	Word task				Sentence task			
	ASR		L1 Listeners		ASR		L1 Listeners	
Speakers	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
SP06	26.32	-	55.70	6.90	54.77	16.84	50.20	21.54
SP07	21.05	-	19.74	9.01	40.07	51.91	49.04	24.33
SP14	63.16	-	73.68	10.76	63.64	12.86	75.38	19.42
SP15	36.84	-	59.65	10.61	54.92	30.42	61.16	26.67
Overall	36.84	-	52.19	22.09	52.18	30.60	58.18	25.42

Similar to Inceoglu et al. (2023), a series of Kendall's tau-b correlations were carried out for all the words and for each speaker to examine the association between ASR and the L1 listeners. Kendall's tau-b correlations were performed because of the non-parametric data sample obtained for ASR. The results revealed a moderate positive correlation between the two groups for the overall sample which was statistically significant ($\tau\mathrm{b} = .46, p < .001$). At the individual level, statistically significant moderate positive correlations were found for speakers SP14 ($\tau\mathrm{b} = .53, p = .01$) and SP15 ($\tau\mathrm{b} = .56, p < .01$). The results showed weak positive correlations for the other two speakers (PS06: $\tau\mathrm{b} = .28, p < .17$; PS07: $\tau\mathrm{b} = .35, p = .11$), but were not significant.

3.1.2. Sentence task

Figure 1 and Table 7 also illustrate the comparisons between ASR and L1 listeners for the sentence intelligibility task. As for the word task, the recognition scores for the sentences produced by the speakers were higher for L1 listeners (58.18%) than ASR (52.18%), in line with Inceoglu et al.'s (2023) finding. However, the overall proportions of words recognised in the sentence task for both ASR and L1 listeners in Inceoglu et al.'s study were much higher (ASR: 75.52%; L1 listeners: 83.88%) than in the current study.

Another compelling result is that the gap between ASR's and L1 listeners' scores is greater for the word task (15.35% difference) than for the sentence task (6% difference). To the extent possible, the sentence stimuli for this study were created so that listeners could not rely on context to recognise the words, which may have contributed to that small 6% difference.

At the individual level, the results also revealed that speakers SP07 and SP14 received the lowest and highest scores respectively from both ASR (SP07: 40.07%; SP14: 63.64%) and the L1 listeners (SP07: 49.04%; SP14: 75.38%), consistent with the word task analysis. This further corroborates the agreement between ASR and the human raters that SP07 is the least intelligible, whereas SP14 is the most intelligible speaker in our sample. One noteworthy difference, however, is that the L1 listeners' intelligibility scores showed high variability for speakers SP07 (as well as SP15), echoing the fair inter-rater agreement results found in the previous section. Because of SP07's low intelligibility, it is possible that for their assessment of speaker SP07 the L1 listeners relied more on their familiarity with French-accented speech or the French language, which was different across the group of listeners.

Given the small sample size, Spearman's rank correlation was performed to examine the relationship between ASR's and L1 listeners' intelligibility scores for the sentence task. The test revealed a statistically significant strong positive correlation between the two groups (*r*(8) = .76, *p* = .025). However, because of the limited number of observations in our data sample, correlations for each individual speaker could not be performed, as was done in Inceoglu et al.'s (2023).

3.2. Comprehensibility

With respect to comprehensibility, L1 listeners were asked to indicate how much effort they made to understand each of the sentences, using a 9-point Likert scale with the score of 1 corresponding to no effort and 9 corresponding to considerable effort. Overall, all four speakers were perceived as relatively hard to understand by the L1 listeners, with average scores reaching beyond the midpoint of the Likert scale (minimum value: 5.88, maximum value: 7.57). However, consistent with the findings in both the word and sentence intelligibility analyses, the results showed that SP14 is the most comprehensible speaker in our sample (*M* = 5.88, *SD* = 1.33). Surprisingly, however, SP07 was not rated as the least comprehensible speaker by the L1 listeners (*M* = 7.51, *SD* = 1.63), even though this speaker received the lowest intelligibility scores in both the word and sentence task.

Instead, the results showed that speaker SP06 was the hardest speaker to understand in our sample (*M* = 7.57, *SD* = 1.90). This result shows that intelligibility may not be the only factor to influence a speaker's comprehensibility. It may, for instance, be affected by the types of errors produced by the speakers, which is analysed in the next section. Moreover, in relation to instructed L2 pronunciation learning, this supports other findings which confirm that several factors impact comprehensibility, other than the speaker's pronunciation: " ... if the goal of instruction is to achieve both global intelligibility and comprehensibility, a focus on developing a better command of vocabulary and more fluidity in accessing it will make L2 accented speech easier to understand." (Thomson, 2018, 23)." Finally, the fourth speaker, SP15, received comprehensibility scores that were intermediate within the range of values obtained for our sample (*M* = 6.2, *SD* = 2.45), but also showed higher variability than the other speakers, as shown by the relatively large standard deviation.

Due to the small data size, the Shapiro-Wilk test of normality was performed. The test showed that the comprehensibility data deviated significantly from a normal distribution ($W = .89$, $p < .001$). Thus, the Kruskal-Wallis test for non-parametric data was used to compare the comprehensibility scores obtained for each speaker. The test yielded statistically significant differences between the speakers with respect to their comprehensibility scores ($H = 18.61$, $p < .001$). Post-hoc analyses using multiple pairwise Wilcoxon comparisons with Benjamini-Hochberg correction revealed that speaker SP06 did not differ significantly from SP07, and speaker SP14 did not differ significantly from SP15. However, SP14 and SP15 had comprehensibility scores that were statistically different from both SP06 ($p < .01$ for SP14 and $p = .04$ for SP15) and SP07 ($p < .001$ for SP14 and $p = .04$ for SP15).

3.3. Error types

Types of errors were also examined in ASR output and the L1 listeners' transcriptions of the individual words in the word task produced by the four speakers in our sample. Transcription errors were classified following the code in Table 5. Errors due to an incorrect vowel were the most frequently found in ASR output (27.08%) and in L1 listeners' transcriptions (36.01%), followed closely by errors due to a combination of both an incorrect vowel and an incorrect consonant (25% for ASR, 22.71% for L1 listeners). The third most common type of error was also the same for both ASR and L1 listeners, namely errors due uniquely to an incorrect consonant (12.50% and 15.37% respectively). The proportions for each error type are summarised in Table 8. Overall, while the proportions are quite comparable across the two groups regarding the second and third most common error types, L1 listeners showed greater proportions than ASR for the most common error type, namely those due to an incorrect vowel. Given the limited data, the proportions for the other types of error were not substantial enough to draw conclusions.

Table 8: Percentage for each error type in the word intelligibility task

Error type	ASR (n = 48)	L1 Listeners (n = 436)	Examples of error
a. One vowel	27.08	36.01	ship → sheep
b. One consonant	12.50	15.37	thank → sank
c. One vowel + One consonant	25.00	22.71	suck → shook
d. Two consonants	2.08	0.92	dish → this
e. One vowel + two consonants	4.17	1.83	path → bathe
f. One vowel + missing consonant	0.00	1.38	yuck→ eck
g. Extra segment	4.17	4.36	fade → frayed
h. One consonant + extra segment(s)	2.08	0.23	bet → beds
i. One vowel + extra segment(s)	6.25	6.65	shade → shield
j. One vowel + One consonant + extra segment(s)	4.17	3.90	worse → walls
k. One vowel + One consonant + missing consonant	2.08	2.29	yuck → egg
l. One vowel + two consonant + missing consonant	2.08	0.00	worth → ross
m. Missing segment	0.00	0.46	yuck → uck
n. Two consonant + extra segment(s)	0.00	0.23	fade → thrate
0. Whole word missing or not transcribed	6.25	2.98	

To further investigate the error patterns in the word intelligibility task, the errors in ASR output and L1 listeners' transcriptions were also divided into the relevant phonetic contrast categories to which they were associated. The individual words selected for this study illustrated three vowel and two consonant contrasts that are particularly challenging for L1 French speakers to produce: /iː/-/ɪ/ (tense vs. lax vowels), /æ/-/ʌ/ (low-front vs. mid-central vowel), /ɛ/-/eɪ/ (mid vowel diphthongization), /ʧ/-/ʃ/ (affricate vs. postalveolar fricative in both word-initial and final positions), /s/-/θ/ (alveolar vs. interdental fricatives in both word-initial and final positions). The results revealed very similar patterns between ASR and L1 listeners, showing a possible agreement between the two groups. As shown in Table 9, ASR's proportions of errors by phonetic contrast and for each speaker are very comparable to those of human listeners. For both ASR and L1 listeners, SP14 was associated with the least number of errors and

SP07 the largest number of errors, which is consistent with the results that SP14 was found to be the most intelligible and SP07 the least intelligible speaker of the sample.

Another interesting result is that each speaker showed a different error pattern. Subject SP06 mostly had trouble with word-initial interdental fricatives and lax vowels. By contrast, SP15 had the most issues with word-final interdental fricatives and some difficulty with mid-vowel diphthongization. As for the other two speakers, the proportions of errors were distributed over multiple categories. SP07, who was found to be the least intelligible speaker, had most issues with low front vowels, then word-final postalveolar fricatives/affricates, mid-vowel diphthongization, and word-final interdental fricatives in that order. On the other hand, SP14, the most intelligible speaker in the sample, mostly had trouble with word-initial postalveolar fricatives/affricates, then low front vowels, word-initial interdental fricatives, and lax vowels in that order.

Table 9: Percentage of errors recognised by ASR and L1 listeners by phonetic contrast

	ASR				L1 Listeners			
Speakers	SP06	SP07	SP14	SP15	SP06	SP07	SP14	SP15
(n)	(14)	(15)	(7)	(12)	(101)	(183)	(60)	(92)
/iː/ - /ɪ/	**35.71**	0.00	**14.29**	0.00	**38.61**	0.00	**13.33**	0.00
/æ/ - /ʌ/	0.00	**40.00**	**28.57**	0.00	0.00	**39.34**	**26.67**	0.00
/ɛ/ - /eɪ/	0.00	**20.00**	0.00	**33.33**	0.00	**16.94**	0.00	**25.00**
/tʃ/ - /ʃ/ (initial)	0.00	0.00	**42.86**	0.00	0.00	0.00	**43.33**	0.00
/tʃ/ - /ʃ/ (final)	0.00	**33.33**	0.00	**8.33**	0.00	**34.97**	0.00	**11.96**
/s/ - /θ/ (initial)	**64.29**	0.00	**14.29**	0.00	**61.39**	0.00	**16.67**	0.00
/s/ - /θ/ (final)	0.00	**6.67**	0.00	**58.33**	0.00	**8.74**	0.00	**63.04**

4. Discussion

The goal of the current study was to explore the accuracy of one ASR-based tool, *dictation.io,* compared to L1 listeners in understanding the L2 English of native French speakers of a low-intermediate proficiency level. In this section we will address each of the three research questions from the Introduction.

In response to RQ1, concerning the accuracy of *dictation.io* for these learners' L2 English, overall, the data revealed relatively low intelligibility scores for ASR and L1 listeners for both task conditions, namely the isolated words and the sentences. Like Inceoglu et al. (2023), however, the scores were higher for the sentence task than the word task. One notable difference, however, is that the L1 listeners overall showed higher proportions of recognised words than the ASR for both task conditions (word and sentence tasks), whereas Inceoglu et al. (2023) found this was only true for the sentence task.

Regarding the word intelligibility task, our L2 speakers received higher scores from L1 listeners than from the ASR. Moreover, in contrast with Inceoglu et al. (2023), the ranking of our speakers is consistent across the two groups (S14 > S15 > S06> S07) for intelligibility scores as well as sentence comprehensibility scores. The one speaker who received higher scores with ASR also received the lowest score with the L1 listeners; Inceoglu et al. (2023) had two such cases. One possible explanation is that listeners reacted to features which the ASR tool ignored, such as voice quality or prosodic traits.

The sentence task revealed the greatest differences between the two studies, as we found lower mean intelligibility scores (≈ 50%) than Inceoglu et al. (2023) at 75%, and much lower than McCrocklin and Edalatishams (2020) with self-declared native English, Spanish and Chinese speakers (over 90% for all groups), using *Google Voice Typing*. This could be due to the importance of linguistic context in speech perception; our target sentences were deliberately created so that listeners could not rely on context. This is in line with Kennedy and Trofimovich (2008), where decontextualization lowers intelligibility.

Explanations for those patterns may be found in the results related to RQ2, about the extent to which this L2 speech correlated with human listener recognition (intelligibility) and with their ratings of comprehensibility. Our analysis showed that overall, ASR's responses positively correlated with those of L1 listeners, moderately for the word task and strongly for the sentence task. This pattern is in clear contrast with Inceoglu et al. (2023), where significant weak to moderate correlations were found for only one speaker. In the current study, significant positive correlations between ASR and the human raters were found for two speakers, namely SP14 and SP15. These two speakers' overall scores also showed that they were the most intelligible and comprehensible speakers in our sample. Therefore, one implication of the current study could be that learners who are found to be most intelligible and comprehensible by L1 listeners tend to align well with ASR methods.

RQ3 queried whether L1 listeners would have difficulties with the same phonetic contrasts as *dictation.io*, and our results are consistent with those of Inceoglu et al. (2023): segment substitutions dominate. The most common error for both the ASR system and the L1 listeners was vowel substitution (*ship/sheep*), then the combination of an incorrect vowel and an incorrect consonant (*suck/shook*), and finally, consonant substitutions (*thank/sank*). Other error types were considerably rarer.

It may be that tense-lax contrasts bear more weight in comprehensibility ratings and some work has explored which specific segmentals facilitate easy comprehensibility. For example, Munro and Derwing (2006) tested the impact of functional load (FL) errors and found that low FL errors such as /d/ replacing /ð/ did not impact greatly on comprehensibility ratings but that high FL errors did, for example /l/ substituted for /n/. Kang and Moran (2014), analysing Cambridge ESOL examination candidate speech files, found similar results, with high FL errors decreasing drastically as overall proficiency levels increased. Using the International English Language Testing System (IELTS) Pronunciation Scale to assess the English of L1 Japanese speakers, Suzukida and Saito (2022) found that distinguishing mid-level proficiency learners from low-level ones was most influenced by the frequency and quality of errors which had high FL. Their results also led them to conclude that accuracy in low FL segmentals (combined with word stress and syllable accuracy) may be required for speakers to be perceived as highly proficient. These findings echoed their earlier study (2021) where only high FL consonant contrasts (of 7 tested) impacted comprehensibility ratings, and not the three vowel contrasts tested (one with high FL, two with low FL). In our study, which uses Brown's original FL ranking (1988), the three vowel pairs tested all have a high FL (/iː ɪ/ 10/10; /æ ʌ/ 8/10; /e eɪ/ 9/10) whereas the /s/ - /θ/ pair is ranked 5/10 and /ʧ/ - /ʃ/ has low FL (2/10). Thus our results are in line with previous work.

The current study went beyond Inceoglu et al. (2023) in analysing the proportions of errors by phonetic contrasts (e.g., tense/lax, monophthong/diphthong, etc.). While similar patterns appeared between the ASR tool and the L1 listeners, nonetheless the highest proportion of errors by target phonetic feature is different for each speaker. The fact that SP06 had difficulty with lax-tense vowel contrasts whereas SP07 had none, confirms the uniqueness of each speaker's trajectory in mastering the contrasts, a uniqueness which for pedagogical reasons should not be ignored.

Analysing individual results in addition to group trends highlights important pedagogical insights about learning trajectories. Knowing a vowel is not an 'all or nothing' phenomenon; control of production routines varies with different lexical phonetic environments, lexical items and speaking situations (Munro, 2019). Addressing the complexity of establishing a hierarchy of 'difficult' vowels for second language acquisition models, Munro argued that quantitative applied linguistics research remains too much in thrall to null hypothesis statistical testing. Despite

> … the 21st century rethink of statistics in the social sciences […], that emphasis appears to be at odds both with common anecdotal reports and with longitudinal data (Munro et al., 2015) indicating large inter-learner differences in pronunciation learning trajectories. (Munro, 2021, p. 2)

The seemingly idiosyncratic performances of many learners may therefore be evidence of variation as the norm, in which case establishing a pedagogically useful hierarchy would be infeasible. Rather, Munro (2021, p. 12) advises instructors to "lower their expectations of L1-based error hierarchies and instead focus on identifying and addressing individual learner needs," as well as those segments that can potentially interfere the most with intelligibility and comprehensibility ratings.

5. Limitations

The current study has examined a limited number of speakers (N = 4) and phonetic contrasts (N = 5), and excluded prosodic targets, even though other research has shown ASR to be effective in intonation instruction (Verdugo, 2006). Future work should also include more analysis of instructional setting or learner variables (i.e., affective variables, multilingual background, etc.), to see how these impact upon their English learning trajectories. One example of such research is Mroz (2018), which examined how learner factors in mobile-based ASR-enhanced instruction influenced intelligibility and proficiency ratings. Nonetheless, this study revealed that there was agreement between this ASR-based tool and the L1 listeners, as shown by the strong correlation with the sentences and moderate with the words, as well as common error types and phonetic features identified.

6. Conclusion

To conclude, despite the lower scores, the output patterns of *dictation.io* mirrored those of the L1 listeners concerning intelligibility-based speaker ranking, error type frequency, and error frequency by target phonetic feature. The error types and proportions were also generally consistent with Inceoglu et al.'s (2023) findings, with a few exceptions. Support was found for their statement that current ASR technology may be particularly useful for lower proficiency learners. Yet any learners working on their own, seeing their intended words repeatedly misspelt by an ASR system, may succumb to frustration (Putri Yaniafari et al., 2022). Teachers could intervene by explaining the meaning and implications of output or feedback. More generally, teachers could make learners of all levels aware of the distinction between accentedness and intelligibility, especially when their learners will interact with a variety of speakers of English, whether native or non-native. Such awareness might also mitigate foreign language anxiety; if everyone has an accent, then aiming for comfortable intelligibility is a valid goal.

ASR has great potential to raise learners' awareness about the type of errors they make during oral production. The combination of ASR with explicit/implicit feedback from the teacher could improve L2 pronunciation learning in instructed settings. Thus, while ASR has not (yet) come of age for use in self-study, a future ideal ASR-based, language-learning programme could “recognise everything the user says, point out those areas that are most problematic … and then offer explicit feedback indicating how to improve” (Fouz-Gonzalez, 2015, p. 324). Such a tool would offer hope, for example, in the very real challenge of transferring a new routine from carefully controlled speech to spontaneous speech. Furthermore, we join Coulange (2023) in calling for greater collaboration between teachers and engineers, for such pedagogically effective tools to become a reality.

References

Agarwal, A. (2022). dictation.io [Online app]. Digital Inspiration. https://dictation.io/

Boersma, P. and D. Weenink. 2022. *Praat: Doing Phonetics by Computer* [Computer program]. Version 6.3.09, retrieved 14 August 2022 from http://www.praat.org/

Brown, A. (1988). *Functional Load and the Teaching of Pronunciation.* TESOL Quarterly, 22(4), 593. https://doi.org/10.2307/3587258; DIALANG 2022. [https://dialangweb.lancaster.ac.uk/]

Coulange, S. 2023. *Computer Aided Pronunciation Training in 2022: When Pedagogy Struggles to Catch Up*. In A. Henderson and A. Kirkova-Naskova (Eds.), *Proceedings from the 7th International Conference English Pronunciation: Issues & Practices*. Université Grenoble-Alpes, May 2022. (pp11-22), Grenoble, France. https://hal.science/hal-04159763

Derwing, T. M. 2010. *Utopian Goals for Pronunciation Teaching.* In J. Levis and K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference*. Iowa State University, Sept. 2009. (pp.24-37), Ames, IA: Iowa State University. https://www.iastatedigitalpress.com/psllt/article/id/15147/

Fouz-González, J. 2015. *Trends and Directions in Computer-Assisted Pronunciation Training*. In J.A. Mompéan and J. Fouz-González (Eds), *Investigating English Pronunciation: Trends and Directions*. Basingstoke and New York: Palgrave Macmillan: 314–342. https://doi.org/10.1057/9781137509437_14

Golonka, E. M., A. R. Bowles, V. M. Frank, D. L. Richardson, and S. Freynik. 2014. *Technologies for Foreign Language Learning: A Review of Technology Types and Their Effectiveness.* Computer Assisted Language Learning 27(1): 70–105. https://doi.org/10.1080/09588221.2012.700315

Henrichsen, L. E. 2021. *An Illustrated Taxonomy of Online CAPT Resources*. RELC Journal *52*(1): 179-188. https://doi.org/10.1177/0033688220954560

Inceoglu, S., W-H. Chen, and H. Lim. 2023. *Assessment of L2 Intelligibility: Comparing L1 Listeners and Automatic Speech Recognition*. ReCALL 35(1): 89-104. https://doi:10.1017/S0958344022000192

Inceoglu, S., H. Lim, and W-H. Chen. 2020. *ASR for EFL Pronunciation Practice: Segmental Development and Learners' Beliefs.* The Journal of Asia TEFL 17(3): 824-840. https://doi.org/10.18823/asiatefl.2020.17.3.5.824

Jułkowska, I. A., & Cebrian, J. (2015). Effects of Listener Factors and Stimulus Properties on the Intelligibility, Comprehensibility and Accentedness of L2 Speech. Journal of Second Language Pronunciation, 1(2), 211–237. https://doi.org/10.18823/asiatefl.2020.17.3.5.824

Kennedy, S. and P. Trofimovich. 2008. *Intelligibility, Comprehensibility and Accentedness of L2 Speech: The Role of Listener Experience and Semantic Context.* Canadian Modern Language Review 64(3): 459–489. https://doi.org/10.3138/cmlr.64.3.459

Kivistö de Souza, H. and W. Gottardi. 2022. How Well Can ASR Technology Understand Foreign-accented Speech? Trabalhos Em Linguística Aplicada, 61(3):, 764-781. https://doi.org/10.1590/010318138668782v61n32022

Landis, J. R. and G. G. Koch. 1977. *An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers.* International Biometric Society *33*(2): 363-374. https://doi.org/10.2307/2529786

Levis, J.W. and R. Suvorov. 2020. *Automatic Speech Recognition.* In C.A. Chapelle (ed), *The Concise Encyclopedia of Applied Linguistics.* Hoboken: Wiley-Blackwell: 149–156. https://doi.org/10.1002/9781405198431.wbeal0066.pub2

Liakin, D., W. Cardoso, and N. Liakina. 2017. *Mobilizing Instruction in a Second-Language Context: Learners’ Perceptions of Two Speech Technologies.* Languages 2(3): 11. https://doi.org/10.3390/languages2030011

McCrocklin, S. and I. Edalatishams. 2020. *Revisiting popular speech recognition software for ESL speech.* TESOL Quarterly 54(4): 1086–1097. https://doi.org/10.1002/tesq.3006

McCrocklin, S., A. Humaidan, and I. Edalatishams. 2019. *ASR Dictation Program Accuracy: Have Current Programs Improved?* In J. M. Levis, C. Nagle, and E. Todey (Ed.s), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*. Ames, IA, Sept.2018. (pp. 191–200). Ames, IA: Iowa State University. https://iastate.box.com/shared/static/wtnv3yg890ze2ibtkihwdpts7bfojt8h.pdf

Moussalli, S. and W. Cardoso. 2020. I Computer Assisted Language Learning 33(8): 865-890. https://doi.org/10.1080/09588221.2019.1595664

Mroz, A. 2018. *Seeing How People Hear You: French Learners Experiencing Intelligibility through Automatic Speech Recognition.* Foreign Language Annals 51(3): 617–637. https://doi.org/10.1111/flan.12348

Munro, M. J. 2019, December 12. *Where to Next? Thoughts on the Future of Pronunciation Research.* [Plenary]. 13th International Conference on Native and Non-native Accents of English, University of Łódź, Poland.

Munro, M. J. 2021. *On the Difficulty of Defining “Difficult” in Second-Language Vowel Acquisition.* Frontiers in Communication 6*: 1–15* https://doi.org/10.3389/fcomm.2021.639398

Munro, M. J. and T. M. Derwing. 1995a. *Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-accented Speech.* Language and Speech 38(3): 289–306. https://doi.org/10.1177/002383099503800305

Munro, M. J. and T. M. Derwing,. 1995b. *Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners.* Language Learning 45(1): 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Peirce, J. W., J. R. Gray, S. Simpson, M. R. MacAskill, , R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv. 2019. *PsychoPy2: Experiments in Behavior Made Easy*. Behavior Research Methods 51(1): 195–203. https://doi.org/10.3758/s13428-018-01193-y

Putri Yaniafari, R., V. Olivia, and Suharayadi. 2022. *The Potential of ASR for Improving English Pronunciation : A Review.* KnE Social Sciences 7(7): 281-289. https://doi.org/10.18502/kss.v7i7.10670

R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Suzukida, Y., & Saito, K. 2022. *What Is Second Language Pronunciation Proficiency? An Empirical Study.* System, 106: 102754. https://doi.org/10.1016/j.system.2022.102754

Suzukida, Y., & Saito, K. 2021. *Which Segmental Features Matter for Successful L2 Comprehensibility? Revisiting and Generalizing the Pedagogical Value of the Functional Load Principle*. Language Teaching Research 25(3): 431–450. https://doi.org/10.1177/1362168819858246

Swan, M and B. Smith. 2001. *Learner English: A Teacher’s Guide to Interference and Other Problems (2nd ed).* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667121

Thomson, R. I. 2018. Measurement of Accentedness, Intelligibility and Comprehensibility. In Kang, O., & Ginther, A. (Eds.). *Assessment in second language pronunciation*. (pp. 11-29). Routledge. https://doi.org/10.4324/9781315170756-2

Verdugo, D. R. 2006. *A Study of Intonation Awareness and Learning in Non-native Speakers of English.* Language Awareness 15(3): 141–159.