

Scored but Not Flagged in French: How Azure Pronunciation Assessment Responds to Single-Phoneme Errors in Spanish and French

Jasper Vande Moortel
Orion-Solution, Belgium
jaspervandemoortel@orion-solution.net

Manuscript version 0.5. Reading edition: single column, one section per page.

Abstract

Commercial pronunciation-assessment services document their scores per language, not their sensitivity per language. We measured one, Azure Pronunciation Assessment, on a synthetic corpus of 70 clips in which each item exists as a clean rendition and as a twin with one segmental error introduced through a text-to-speech phoneme override, scored under 4 recogniser locales and 2 modes (280 assessments); phoneme- and syllable-level detail came back for Spanish and French in every assessment, which the documentation does not promise. The paired effect size (d_z) between clean and broken accuracy is clearly positive in es-ES scripted (1.07 [0.62, 1.98]), positive with the pair-level interval excluding zero in es-MX and fr-CA scripted (0.82 and 0.81; sign test $p = 0.031250$ and 0.062500 on the same ten pairs), and near zero with an interval crossing zero in fr-FR scripted (0.27 [-0.49, 0.94]). Word-level flagging is the sharper asymmetry and holds by item as well as by cell: 50.0% of the es-ES scripted broken clips are caught with no false rejections (FRR), against 0 of 10 broken clips in every French locale-mode cell and 0 of 5 French items (Wilson upper 27.8%). A French severity ladder shows the service is insensitive rather than blind: whole-syllable corruption moves the score by 88.3 against 95.4 for clean speech; the recognised text matches the reference on all 10 severe clips, and equally on the clean and mild ones, so this is offered only as a candidate mechanism rather than demonstrated recogniser repair. The stimuli are synthetic, AI-generated Azure neural TTS audio, redistributable under Microsoft's Product Terms; corpus, raw responses and analysis code are released.

Contents

1	Introduction	3
2	Related work	4
3	Method	5
4	Results	7
5	Discussion and conclusion	9
	Limitations	10
A	Field presence	13
B	Item-level view	15
C	Severity ladder	17
D	Candidate mechanisms	18
E	Literature search log	19

1 Introduction

A school buying a pronunciation scorer for Spanish or French has one document to go on: the vendor’s list of supported locales. The published evaluations of automated pronunciation assessment are overwhelmingly about English, the shared benchmark is English (Zhang et al., 2021), and the two independent evaluations of the service studied here outside English, both from one group, concern Arabic (Zainaddin et al., 2026; Zainaddin and Amro, 2026). Whether a service that reports a score for a language can also *detect* the errors a learner of that language makes is a separate question from whether the locale is listed, and nobody has answered it for the two languages Flemish secondary schools teach most after English.

This paper asks that question in the narrowest form answerable without a learner corpus: given a clean rendition and a twin with one planted segmental error, does the service score them apart, and does it flag the error? We synthesised 70 clips with Azure neural voices, planted one segmental error per item (one phoneme in every item but one, a multi-segment vowel-reduction change), added prosodic variants, and scored every clip under two recogniser locales per language and both modes; es-MX and fr-CA re-score the same audio as es-ES and fr-FR under a different recogniser configuration, not a second accent. The answer is a graded asymmetry. Paired effect sizes on the accuracy score are clearly positive under es-ES, positive with the pair-level interval excluding zero under es-MX and fr-CA, near zero and uncertain under fr-FR. Word-level flagging is starker and holds by item as well as by cell: Spanish catches half the planted errors with no false rejections, French flags none in any cell or item. A second French set corrupted at three severities shows the scorer responding to unmistakable corruption while missing single-phoneme errors; its transcripts match the reference at every severity, clean included, so recogniser repair remains a candidate mechanism, not a demonstrated one.

Two contributions follow: a measurement a school can act on (the same API gives usable word-level feedback in Spanish and none in French, on these stimuli) and a method (synthetic segmental corruption with a severity ladder separates “does not flag” from “flags only gross errors” at a cost of a few euro, repeatable for evaluation on any service or locale from the released corpus; the audio may not be used to create, train or improve a competing product).¹

¹Corpus, raw responses, analysis code and manuscript source: <https://github.com/anonymised/spoken-assessment-lab> (DOI at camera-ready). The audio is synthetic Azure TTS output, redistributable under Microsoft’s Product Terms (corpora/SPEECH_DATASHEET.md). Manuscript version 0.5.

2 Related work

Automated pronunciation assessment descends from goodness-of-pronunciation scoring, which compares the forced-alignment likelihood of the intended phone against a free phone loop with no language model (Witt and Young, 2000) and survived the move to neural acoustic models (Hu et al., 2015). Its sensitivity to alignment is a known weakness: Cheng et al. (2020) name the phone-competition problem and propose assessment without recognition. Two adjacent domains offer the same fix as method precedent, not results: clinical dysarthria assessment (Yeo et al., 2026) and cross-lingual phoneme recognition (Xu et al., 2022) both adopt language-model-free universal phone recognisers to avoid lexical priors that obscure errors. Mispronunciation detection is evaluated on the English benchmark speechocean762 (Zhang et al., 2021) with phone- and word-level precision, recall and correlation with human scores (Lounis et al., 2024); training data are augmented by blending real recordings (EL Kheir et al., 2023) or by treating synthesised speech as the reference (Lo et al., 2025). Korzekwa et al. (2022) report detector AUC rising from 0.928 to 1.00 with the severity of synthetic phoneme substitutions, above the same detector’s results on real non-native speech; synthetic stimuli therefore give an upper bound on sensitivity, which is the sense used here. An open alternative and the other commercial services searched are covered in Appendix E.

No independent evaluation of the service under study exists for Spanish or French, and no published study measures reference-conditioned recogniser repair in scripted assessment or compares scripted with unscripted sensitivity to *planted* segmental errors (search log in Appendix E); the one prior scripted-versus-unscripted comparison, Zainaddin et al. (2026), uses real child speech on unrelated Arabic tasks with no planted errors and finds the opposite direction, no-reference mode inflating scores relative to scripted. Microsoft’s documentation states that pronunciation assessment uses a speech-to-text model version distinct from standard STT, and separately that unscripted assessment also uses a model distinct from standard STT, without comparing the two modes’ models to each other (Microsoft, 2026); edit-distance alignment against the reference text is described only in a support forum answer (Microsoft, 2024), and reference-conditioned decoding is undocumented. For the learner population, the Spanish tap–trill contrast and the palatal nasal are documented for Dutch-speaking learners by Sarmiento and De Vriendt (1977) (*coro/corro*, *bañol/paño*); the jota, theta and vowel-reduction items and every French item are typological predictions (no Dutch-L1 study of French nasal vowels, the uvular rhotic, final devoicing or liaison was found), the dental voicing confusion the source documents was not tested, and the rounded vowel in *tu* against *tout*, a contrast Dutch has, is a control rather than a predicted error.

3 Method

Corpus. Five sentences per language were written to carry one target word each. Every sentence was synthesised with Azure neural text-to-speech in two voices per language as a clean rendition and as a twin carrying one segmental error in the target word through an SSML <phoneme> override with deliberately incorrect IPA, giving 10 clean–broken pairs per language (five items $\times$ two voices). In nine of the ten items the error is a single phoneme; *es_vowel_reduction* reduces the stressed vowel and two of the word’s three unstressed vowels to schwa, keeps the unstressed /i/, and also swaps /θ/ for /s/ (a seseo change duplicating *es_theta_zapato*’s contrast), a distinct multi-segment perturbation kept in the set and reported separately where it matters. The Spanish errors are the tap–trill contrast, the velar fricative, the interdental, the palatal nasal and vowel reduction; the French errors are a nasal vowel, a rounded vowel (*tu* against *tout*, a control since Dutch also has this contrast), the uvular rhotic, postalveolar voicing and a silent final consonant. Prosodic variants at $\pm 45\%$ speech rate and with flattened pitch were added, for 70 clips. Synthesis success and byte-level clean–broken audio difference were checked for every pair before scoring. The corpus is released with a datasheet naming the voices and SSML (SPEECH_DATASHEET.md), a stated audio licence and a paper checklist.

Scoring. Every clip was scored with Azure Pronunciation Assessment under two recogniser locales per language (es-ES, es-MX; fr-FR, fr-CA) and in both modes, scripted and unscripted, for 280 assessments (4 scorings per clip) on 9 September 2026; every response is logged in full. Scripted and unscripted name the API’s ReferenceText parameter, set or absent, not a read-aloud-versus-spontaneous task: every clip is the same read-aloud sentence in both modes. es-MX and fr-CA are not a second accent: they re-score the same es-ES and fr-FR audio under a different recogniser locale, so the four cells per language compare recogniser configurations, not speakers. Field presence was recorded per response. The primary outcome is the accuracy score; the secondary is word-level flagging (recall: true-detection rate on planted errors; FRR: false rejection rate, a correct word wrongly rejected in a clean clip), in which a planted error counts as caught when the target word carries the Mispronunciation error type or a word accuracy below 60.

Severity ladder. A second French set of five sentences in two voices was rendered clean, with one phoneme substituted (mild) and with a whole syllable substituted (severe), 10 clips per condition, scored in fr-FR scripted mode. For each clip the recognised text was compared with the reference after normalising case and punctuation.

Statistics. The ten pairs per cell are five items scored in two voices, and the voices are not independent replicates: several items return near-identical scores across them, so a pair-level interval overstates precision. Two effect sizes are reported per cell: d_z (mean paired difference over the SD of the differences), primary for this matched design, and the pooled-SD d of prior work, for comparison, both with BCa bootstrap intervals over the 10 pairs (2000 resamples, seed 20260909); item-level differences (voices averaged) are in Appendix B. Recall and flagged counts carry Wilson 95% intervals. The analysis plan, including the alternative flag threshold of 70 and the two-locale maximum variant, was fixed before re-derivation; two post hoc exclusions are marked as

such (Appendix B), and every number comes from the logged responses through one script.

locale	mode	clean	broken	d_z [CI]	d [CI]	recall % [CI]	FRR
es-ES	scripted	99.0	93.0	1.07 [0.62, 1.98]	1.54 [0.88, 2.81]	50 [24, 76]	0.00
es-ES	unscripted	96.6	92.8	1.03 [0.56, 1.68]	0.87 [0.43, 1.66]	50 [24, 76]	0.10
es-MX	scripted	95.9	92.0	0.82 [0.47, 1.40]	0.90 [0.40, 1.72]	40 [17, 69]	0.00
es-MX	unscripted	93.6	89.8	0.70 [0.28, 1.34]	0.70 [0.26, 1.31]	40 [17, 69]	0.00
fr-FR	scripted	94.5	93.8	0.27 [-0.49, 0.94]	0.32 [-0.56, 1.07]	0 [0, 28]	0.00
fr-FR	unscripted	91.9	92.3	-0.19 [-0.98, 0.57]	-0.08 [-0.33, 0.29]	0 [0, 28]	0.10
fr-CA	scripted	96.0	94.3	0.81 [0.41, 1.35]	0.32 [0.07, 1.02]	0 [0, 28]	0.00
fr-CA	unscripted	95.5	95.8	-0.44 [-1.16, 0.25]	-0.05 [-0.42, 0.04]	0 [0, 28]	0.00

Table 1: Clean-vs-broken discrimination per recogniser locale and mode, 10 matched item/voice pairs per cell (95% BCa bootstrap intervals over the matched pairs; d_z is the mean paired difference over the SD of the differences (the design’s own effect size), d is the pooled-SD unpaired-formula effect size, kept for comparison to prior literature; Wilson 95% intervals for *recall*, over the 10 broken clips per cell). *recall* is the share of planted errors caught by the word-level rule (ErrorType Mispronunciation or word accuracy < 60). *FRR* (false rejection rate) is the mean count of words flagged by the same rule on the 10 clean clips per cell.

4 Results

What comes back. Phoneme- and syllable-level detail was present in 100.0% of assessments under every recogniser locale and mode, although the documentation lists syllable detail as en-US only; prosody was absent as documented, and fluency sat at exactly 100.0 in 97.5% (Table 2, Appendix A); synthetic speech has no hesitations, so fluency was untestable by design.

Discrimination. Table 1 and Fig. 1 give the clean and broken means with both effect sizes; d_z , the paired effect size the design calls for, is discussed first, at the pair level (five items, two voices, $n = 10$). es-ES scripted is clearly positive ($d_z = 1.07$ [0.62, 1.98]); es-MX and fr-CA scripted are positive with the pair-level interval excluding zero ($d_z = 0.82$ and 0.81). At the item level ($n = 5$) the bootstrap still excludes zero but is wider, and the exact sign test on the ten pairs reaches $p < 0.05$ only for es-ES and es-MX (Appendix B), so “excluding zero” is a pair-level bootstrap property for fr-CA. fr-FR scripted is smaller with an interval crossing zero ($d_z = 0.27$); both French unscripted cells sit near zero. On the pair unit ($n = 8$) the es-MX aggregate moves to $d_z = 1.00$ [0.54, 1.83] once *es_theta_zapato*, whose broken form is the es-MX norm, is excluded (Appendix B). Word-level flagging holds without qualification: es-ES caught 50.0% (23.7–76.3% Wilson, 5 of 10 broken clips scored under the es-ES recogniser in scripted mode) of planted errors and es-MX 40.0%, 0.00 FRR per clean clip; French flagged 0 of 10 broken clips in every one of the four locale-mode cells (Wilson upper 27.8%) and 0 of 5 items, unchanged at threshold 70 (Appendix B). The two Spanish items always caught are *es_nasal_espana* and *es_vowel_reduction*. The primary between-language test is the scripted-only Fisher exact comparison (9 of 20 Spanish vs 0 of 20 French, $p = 0.001229$); each arm scores the same ten clips twice, once per recogniser locale, so the two locales are not independent samples, and the item-level floor above is unit-honest regardless, since no French item is ever caught.

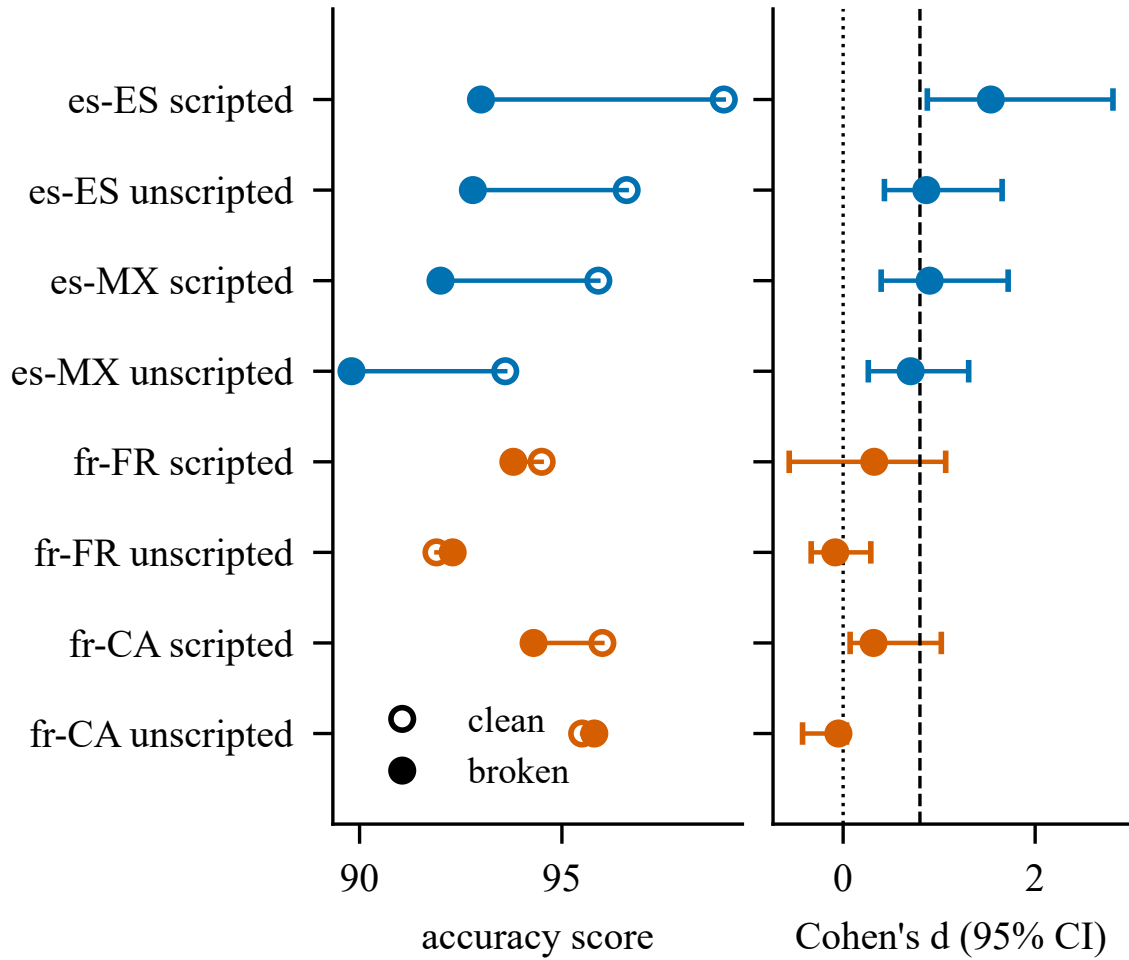


Figure 1: Clean (hollow) and broken (filled) accuracy per recogniser locale and mode over 10 clean–broken pairs, and Cohen’s d with 95% BCa bootstrap intervals; Table 1 gives d_z , the paired effect size, alongside. es-ES ($d_z = 1.07$) and es-MX and fr-CA scripted ($d_z = 0.82$ and 0.81) exclude zero; fr-FR scripted ($d_z = 0.27$) and the French unscripted cells cross zero.

Scripted beats unscripted. Every recogniser locale scored higher and discriminated better with a reference text supplied, matching the two-pass design Microsoft’s documentation recommends (Microsoft, 2026); the two-locale-maximum variant is in Appendix B.

Insensitive, not broken. The severity ladder (Table 5 and Fig. 2, Appendix C) moves monotonically: 95.4 clean, 93.0 with one phoneme substituted, 88.3 with a whole syllable substituted, 2 and 6 of 10 target words flagged. One mild-condition item, *fx_matin*, was synthesised with its orthographic spelling where an IPA override was meant to go; without it the mild mean is 94.0 and 0 of 10 flagged (Wilson upper 31.2%), so every mild-condition flag traces to that item (datasheet documents it). On all 10 severely corrupted clips, and equally on the clean and mild clips, the recognised text matched the reference; a rate that does not vary with severity cannot distinguish recogniser repair from reference-biased decoding, so repair stays a candidate mechanism (Appendix D). The per-word flags still saw the corruption in 6 of 10 severe clips, scoring what the transcript hid.

5 Discussion and conclusion

Two languages share an API, a documentation page and a price, and the accuracy score separates broken from clean unevenly: clearly under es-ES, at the pair level under es-MX and fr-CA, barely under fr-FR. Word-level flagging is not a matter of degree: half the planted Spanish errors are caught and none of the French ones, in every cell and item. Spanish scoring on this service can carry word-level feedback; French should ship as a coarse band with a caveat, or with the criterion disabled, until an engine that does not repair its input has been validated.

Three explanations remain open and the data cannot separate them: a less sensitive French acoustic model, recogniser repair before scoring, or the mode-specific speech-to-text model Microsoft documents (Appendix D). A planned follow-up scores the same clips with a language-model-free phone recogniser, which cannot repair and so discriminates among them; a second uses teacher-graded recordings of Flemish pupils, turning each upper bound into an estimate.

Limitations

All stimuli are synthetic. Text-to-speech renders one phoneme wrong in an otherwise native voice, which is easier to detect than a learner's accented production (Korzekwa et al., 2022); every effect size here is an upper bound on sensitivity, not an estimate of it, and the absence of hesitation means fluency was untested by construction. There are five items per language, two voices each, so the pair-level intervals reported alongside the item-level ones are wide at $n = 5$ and the French intervals include zero as well as moderate effects; the asymmetry is supported by the Spanish intervals excluding zero, the French word-level count of 0 of 10 in every cell (0 of 5 items), and the scripted-only Fisher exact test between languages ($p = 0.001229$), whose two arms both re-score the same ten clips, not independent samples. There is no human listener criterion: a detected substitution is not shown to be one a listener would notice. Whether a broken clip is a well-formed mispronunciation or a text-to-speech synthesis artefact was not checked by ear or by spectral analysis; the only checks are that synthesis succeeded, that clean and broken audio differ byte-for-byte, and that the severe-set transcripts read as the intended sentence. The acoustic magnitude of the phoneme overrides is not measured or matched across the two languages, so item selection remains an alternative explanation to a genuine language asymmetry, and the severity ladder uses a different item set from the main discrimination cells, so it cannot control for this directly. The French items are typological predictions rather than corpus-documented Dutch-L1 errors. The Spanish set is documented for Dutch-speaking learners for the tap–trill contrast and the palatal nasal only (Sarmiento and De Vriendt, 1977); the dental voicing confusion that source also documents (*dado/dato*) was not tested here. Results reflect one service, one account and one day; the service's models change without notice, and the phoneme-level detail that is undocumented for these locales could be withdrawn as silently as it appeared. A scorer that flags nothing reads to a French learner as confirmation the pronunciation was correct, a false reassurance rather than a neutral non-result; a classroom deployment on the current French configuration would need this stated to teachers and pupils. A scorer used to evaluate pupils may fall under the EU AI Act's education high-risk category; this paper evaluates the engine's sensitivity, not its regulatory compliance. An independent reimplementation of the BCa bootstrap (`stats_helpers.bootstrap_ci`, seed 20260909, 2000 resamples) reproduces every lower bound and the zero floor but gives upper bounds that differ by 0.1–0.2 at $n = 4$ –10 with tied differences, so the exact upper-bound value at this sample size is implementation-sensitive; the direction and the zero-crossing are not.

References

- Sitong Cheng, Zhixin Liu, Lantian Li, Zhiyuan Tang, Dong Wang, and Thomas Fang Zheng. 2020. ASR-free pronunciation assessment. arXiv:2005.11902. *arXiv preprint*.
- Yassine EL Kheir, Shammur Chowdhury, Ahmed Ali, Hamdy Mubarak, and Shazia Afzal. 2023. SpeechBlender: Speech augmentation framework for mispronunciation data generation. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 26–30. ISCA.
- Halleck45. 2024. Open pronunciation scoring with wav2vec2 and DTW: OpenPronounce. <https://github.com/Halleck45/OpenPronounce>. Accessed September 10, 2026.
- Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166.
- Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, and Bozena Kostek. 2022. Computer-assisted pronunciation training. speech synthesis is almost all you need. *Speech Communication*, 142:22–33.
- Tien-Hong Lo, Meng-Ting Tsai, Yao-Ting Sung, and Berlin Chen. 2025. Zero-shot text-to-speech as golden speech generator: A systematic framework and its applicability in automatic pronunciation assessment. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 26–30. ISCA.
- Meriem Lounis, Bilal Dendani, and Halima Bahi. 2024. Mispronunciation detection and diagnosis using deep neural networks: a systematic review. *Multimedia Tools and Applications*, 83(23):62793–62827.
- Microsoft. 2024. How does scripted pronunciation assessment use the reference text (Microsoft Q&A answer). <https://learn.microsoft.com/en-us/answers/questions/1542372>. Accessed September 10, 2026.
- Microsoft. 2026. Use pronunciation assessment (Azure AI Speech documentation, model-version note). <https://learn.microsoft.com/azure/ai-services/speech-service/how-to-pronunciation-assessment>. Accessed September 10, 2026.
- Sarmiento and De Vriendt. 1977. Corrección fonética de los neerlandófonos que estudian español. https://cvc.cervantes.es/ensenanza/biblioteca_ele/aepe/pdf/boletin_17_10_77/boletin_17_10_77_09.pdf. Accessed September 10, 2026.
- S. M. Witt and S. J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95–108.
- Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Interspeech 2022*, pages 2113–2117. ISCA.
- Eunjung Yeo, Julie M. Liss, Visar Berisha, and David R. Mortensen. 2026. Multilingual dysarthric speech assessment using universal phone recognition and Language-Specific phonemic contrast modeling. arXiv:2601.21205. *arXiv preprint*.

- Jalal Zainaddin, Mohammad Amro, and Wasfi G. Al-Khatib. 2026. Evaluating Azure pronunciation assessment for mispronunciation detection in children learning Arabic. In *Proceedings of the International Conference on Industrial Engineering and Operations Management*. IEOM Society International.
- Jalal Ali Zainaddin and Mohammad Ismail Amro. 2026. Evaluation of Azure pronunciation assessment tool for Arabic-Speaking learners. *Journal of Undergraduate Research International*, 2(2):65–73.
- Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Interspeech 2021*, pages 3710–3714. ISCA.

A Field presence

Table 2 gives the per-cell field-presence rates behind the "What comes back" paragraph in Section 4: phoneme and syllable detail at 100.0% under every recogniser locale and mode, prosody at 0.0% everywhere, and the fluency-score exceptions confined to one unscripted-mode item.

locale	mode	Acc.	Flu.	Comp.	Pron.	Pros.	Phon.	Syll.
es-ES	scr.	100	100	100	100	0	100	100
es-ES	unscr.	100	100	100	100	0	100	100
es-MX	scr.	100	100	100	100	0	100	100
es-MX	unscr.	100	100	100	100	0	100	100
fr-FR	scr.	100	100	100	100	0	100	100
fr-FR	unscr.	100	100	100	100	0	100	100
fr-CA	scr.	100	100	100	100	0	100	100
fr-CA	unscr.	100	100	100	100	0	100	100

Table 2: Field-presence rate (%) per recogniser locale and mode (*scr.* = scripted, *unscr.* = unscripted), 280 assessments of 70 clips. *Phon.* and *Syll.* are word-level Phonemes/Syllables presence over all words in the cell. The fluency score is exactly 100.0 in 97.5% of all 280 assessments (range 74.0–100.0); the 7 exceptions are unscripted-mode assessments of a single item, detailed in the text.

item	es-ES sc	es-ES un	es-MX sc	es-MX un	fr-FR sc	fr-FR un	fr-CA sc	fr-CA un
es_trill_perro	5.5	6.0	1.5	-0.5	–	–	–	–
es_jota_trabajo	1.0	0.5	0.0	0.5	–	–	–	–
es_theta_zapato	0.0	0.0	0.0	-0.5	–	–	–	–
es_nasal_espana	11.0	6.0	6.5	7.5	–	–	–	–
es_vowel_reduction	12.5	6.5	11.5	12.0	–	–	–	–
fr_nasal_pain	–	–	–	–	1.5	-0.5	3.0	-1.0
fr_y_vs_u_tu	–	–	–	–	4.5	3.0	5.0	0.0
fr_uvular_r_rouge	–	–	–	–	0.0	-1.0	0.5	-0.5
fr_zh_vs_sh_journal	–	–	–	–	0.5	-0.5	0.0	0.0
fr_final_consonant	–	–	–	–	-3.0	-3.0	0.0	0.0

Table 3: Voice-averaged clean-minus-broken AccuracyScore difference per item and per (recogniser locale, mode) cell (A2): five Spanish items, then five French items. A Spanish item was never scored under a French recogniser locale and vice versa, marked –.

B Item-level view

Voice is not an independent replicate of an item: several items return near-identical scores across the two voices, so the pair-level intervals in Table 1 overstate precision relative to the five items per language that actually vary. Table 3 gives the voice-averaged clean-minus-broken difference per item and cell. At the item level, 4 of 5 es-ES scripted items move in the broken direction, against 3, 3 and 3 of 5 for es-MX, fr-FR and fr-CA scripted; Table 4 isolates the es-ES/es-MX comparison per item, showing that *es_theta_zapato* carries a zero diff under es-MX because its broken form is the es-MX recogniser’s own norm rather than an error, and that excluding it moves the es-MX scripted gap from 3.9 points to 4.9. Recall at the alternative word-flag threshold of 70 is unchanged: es-ES 50.0%, es-MX 40.0%, fr-FR 0.0%, fr-CA 0.0%.

The item-cluster bootstrap (resampling the five items, voice-averaged, 2000 resamples, seed 20260909) gives, per scripted cell: es-ES $d_z = 1.06$ [0.45, 2.31]; es-MX $d_z = 0.78$ [0.45, 1.52]; fr-FR $d_z = 0.26$ [-0.95, 1.34]; fr-CA $d_z = 0.76$ [0.45, 1.79]. The exact sign test on the ten pairs per cell gives: es-ES $p = 0.007812$ (8 of 8 usable pairs positive), es-MX $p = 0.031250$ (6 of 6), fr-FR $p = 0.453125$ (5 of 7), fr-CA $p = 0.062500$ (5 of 5); tied differences are dropped before the test, which is why the usable count is below 10. Every cell’s item-cluster interval excludes zero except fr-FR; only es-ES and es-MX reach the conventional $p < 0.05$ threshold on the sign test. Running both locales and keeping the higher score, sometimes recommended when the learner’s variety is unknown, shrank the es-ES gap only slightly (6.0 to 5.7 points), driven by one broken clip flipping locale while no clean clip did; the locale should be fixed per assignment.

item	es-ES scripted			es-MX scripted		
	clean	broken	diff	clean	broken	diff
es_jota_trabajo	99.0	98.0	1.0	96.0	96.0	0.0
es_nasal_espana	100.0	89.0	11.0	94.0	87.5	6.5
es_theta_zapato	99.0	99.0	0.0	99.0	99.0	0.0
es_trill_perro	98.5	93.0	5.5	95.0	93.5	1.5
es_vowel_reduction	98.5	86.0	12.5	95.5	84.0	11.5

Table 4: Per-item, voice-averaged AccuracyScore under es-ES scripted and es-MX scripted (A3). *es_theta_zapato* is the seseo item: its “broken” substitution of /s/ for Castilian /θ/ is the es-MX recogniser’s own norm, so its es-MX diff is 0.

condition	mean acc.	gap	flagged	heard = ref
clean	95.4	0.0	0/10	10/10
mild	93.0	2.4	2/10	10/10
severe	88.3	7.1	6/10	10/10

Table 5: French severity ladder: 5 items $\times$ 2 voices = 10 clips per condition, whole-syllable substitutions unmistakable to a human listener. *gap* is the drop from the clean-condition mean (95% BCa bootstrap intervals on the means are given in the text, omitted here for width). *flagged* is the word-level rule catching the target word. *heard = ref* is how many recognised transcripts exactly match the reference sentence after normalising case, punctuation and whitespace.

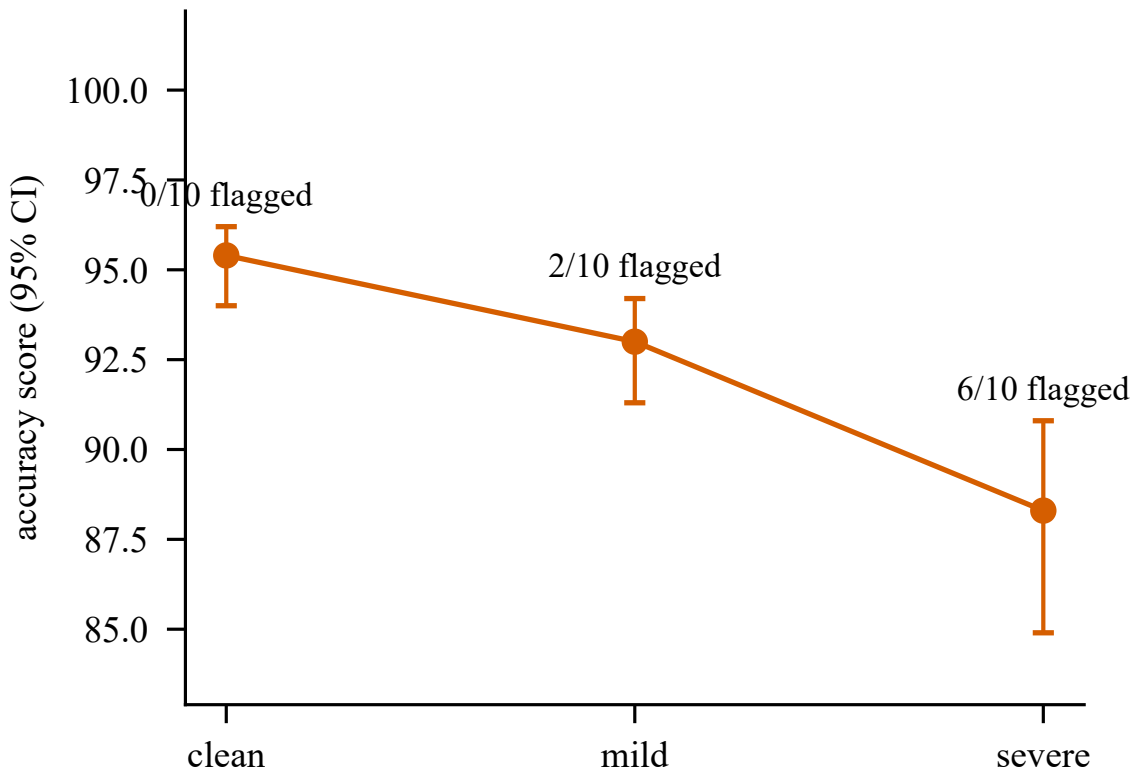


Figure 2: French severity ladder: mean accuracy with 95% BCa intervals over 10 clips per condition and the number of target words flagged.

C Severity ladder

A percentile bootstrap over the same resamples gives the four headline d_z intervals as: es-ES scripted [0.63, 2.03], es-MX scripted [0.47, 1.40], fr-FR scripted [-0.36, 1.05] and fr-CA scripted [0.45, 1.46]. Bounds shift by up to a fifth of the BCa value at $n = 10$, most visibly on fr-FR scripted's lower bound, but the two methods agree on which cells exclude zero.

D Candidate mechanisms

The French insensitivity admits three readings the present data cannot separate. First, a genuinely less sensitive French acoustic model. Second, forced alignment plus a strong French language model repairing the input before scoring; the transcripts cannot adjudicate this, since the recognised text equals the reference at every severity, clean included (Section 4). Third, the speech-to-text model behind pronunciation assessment, which Microsoft documents as a specific version different from standard speech to text, with unscripted assessment using a different model again (Microsoft, 2026); the documentation compares neither model to the other nor to the French and Spanish locales, so this reading explains the mode split within a language and only by assumption the gap between languages. A language-model-free phone recogniser (Xu et al., 2022; Yeo et al., 2026) cannot repair and so discriminates among the three; goodness-of-pronunciation, whose denominator is such a free phone loop (Witt and Young, 2000), is the natural baseline for that follow-up.

E Literature search log

Searches were run on 9 September 2026 against OpenAlex, Crossref, arXiv abstract pages, the ISCA archive and general web search; the full log is in the packet. The absence claims are: no independent evaluation of Azure Pronunciation Assessment in Spanish or French (OpenAlex title and abstract search “azure pronunciation assessment”, 2019–2026, nine hits, none in either language); no published measurement of reference-conditioned recogniser repair in scripted pronunciation assessment and no published comparison of scripted with unscripted sensitivity to *planted* segmental errors (five targeted queries), against Zainaddin et al. (2026)’s comparison of the two modes on unplanted errors in real Arabic child speech; no Dutch-L1 study of French nasal vowels, the uvular rhotic, final devoicing or liaison production (OpenAlex, web search and the IPFC project site); and no SLaTE or Interspeech paper from 2022 to 2025 evaluating a commercial service across Spanish and French on single-phoneme errors (five queries on the ISCA archive). No published cross-language sensitivity measurement was found for any other commercial pronunciation-assessment service either (SpeechAce, ELSA, Speechmatics, Google, iFlytek, Versant, Duolingo; two OpenAlex queries and two web-search passes, two English-only user studies excluded). OpenPronounce (Halleck45, 2024), an open alternative, is validated on speechocean762 only and makes no claim for Spanish or French. Each is a statement about what these searches found.