Language-specific effects on ASR errors for world Englishes

June Choe, Yiran Chen, May Pik Yu Chan, Aini Li, Xin Gao

Linguistics, University of Pennsylvania

& Nicole Holliday

Linguistics, Pomona College

COLING 2022

Introduction

Virtual meeting and conferencing are increasingly becoming part of everyday life, driving advancements in ASR.

Yet **transcription performance varies** between speakers, by:

- Dialect and accent (Wheatley & Picone, 1991; Meyer et al., 2020)
- Gender (Adda-Deecker & Lamel, 2005; Sawalha and Shariah, 2013; Tatman, 2017; Tatman & Kasten, 2017)
- Racial background (Koenecke et al., 2020; Martin & Tang, 2020)

Speakers of World Englishes

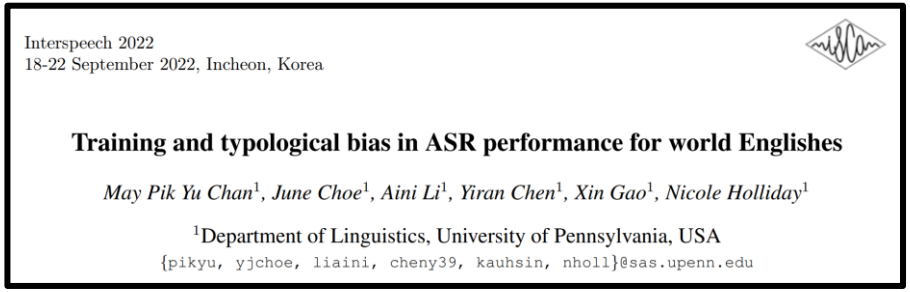
L2 English speakers of different language backgrounds less explored

- 75% of world's English spekers speak it as a second language (Crystal 2002)

L2 speakers may be underserved by ASRs modelled on L1 speakers

Is this the case? If so, in what specific ways?

Data (via Chan et al., 2022)



Interspeech 2022
18-22 September 2022, Incheon, Korea

Training and typological bias in ASR performance for world Englishes

May Pik Yu Chan[1], *June Choe*[1], *Aini Li*[1], *Yiran Chen*[1], *Xin Gao*[1], *Nicole Holliday*[1]

[1]Department of Linguistics, University of Pennsylvania, USA

{pikyu, yjchoe, liaini, cheny39, kauhsin, nholl}@sas.upenn.edu

Performance of **Otter's ASR system** on recordings of World English speakers from the **Speech Accent Archive**.

Otter - ASR used by *Zoom*; reports a list of supported English varieties

Speech Accent Archive - A corpus of >3k speakers around the world reading the same passage containing all sound segments of English.

➔ ⊂ 1.2k speakers of 21 varieties, balanced # of trained vs. untrained

Language-structural effects

Chan et al. (2022) – WER analylsis

1) Effect of training on performance

2) Effect of speaker L1 being a tonal (vs. non-tonal) language

Here, we want to better understand the **source of language-structural effects.**

	Supported	Tonal	Mean WER
English	+		0.035
Hindi	+		0.057
Swedish	+		0.059
German	+		0.065
Swissgerman	+		0.084
French	+		0.098
Italian	+		0.114
Spanish	+		0.115
Russian	+		0.136
Mandarin	+	+	0.157
Cantonese		+	0.162
Thai		+	0.202
Vietnamese		+	0.214
Urdu			0.052
Japanese			0.109
Tagalog			0.109
Arabic			0.114
Korean			0.116
Indonesian			0.122
Dari			0.123
Bengali			0.129
Amharic			0.167

Linguistic analysis of ASR errors

Word Error Rate can be a useful performance metric, but reveals little about the **linguistic nature of errors** - details at the phone-level needed.

When it comes to L2 speakers, we know a bit about **L1->L2 transfer** and the **phonological processes** involved in producing a non-native “accent”.

- E.g., Perceptual Assimilation Model (Best et al., 1994)

Do the type and degree of ASR errors differ across English varieties?

- **Are certain errors predictable from the L1 phonology?**

Phone-level analysis of errors

3 analyses of **phone-level** substitutions (after converting to ARPABET)

1) Vowel substitution errors:

- E.g., *thick → tech* (“TH IH K” -> “T EH K”)

2) Consonant voicing errors:

- E.g., *slabs → slaps* (“S L AE P S” -> “S L AE B Z”)

3) Consonant cluster errors:

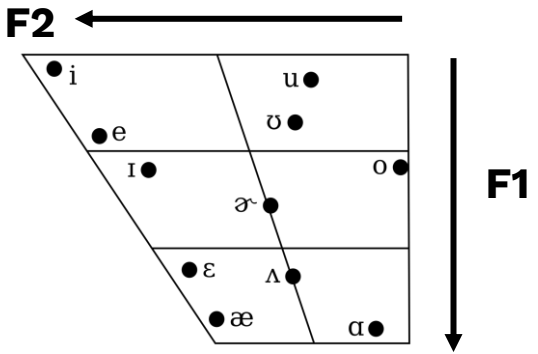
- E.g., *ask → asked* (“AE S K” -> “A S K T”)

1. Vowel substitution error analysis

Acoustic profile of stressed monophthongs, categorized into:

- **Matches**: *thick -> tick*
- **Mismatches**: *thick -> tech*

Speaker-normalized midpoint **F1 and F2 measures** after forced alignment



Vowel substitution data

Korean speaker #2 (Female, 23; Age of Onset: 14):

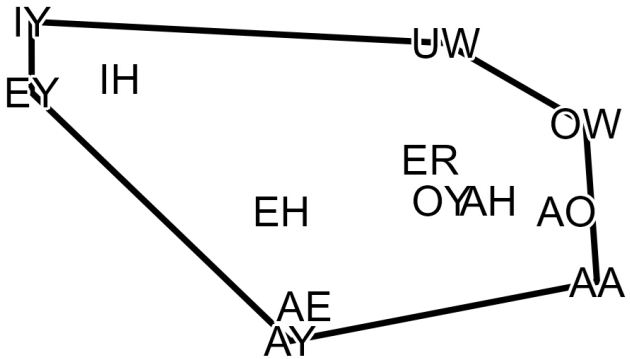
… for the kids … → … for the keys …

Speaker	ID	Observed	Truth	Word	F1	F2
…						
Korean2	150	AO	AO	for	533	1363
Korean2	153	AH	AH	the	397	1975
Korean2	155	IY	IH	kids	437	2118
…						

Vowel substitution visualization

Matches give us “perceived” vowel space by L1 background

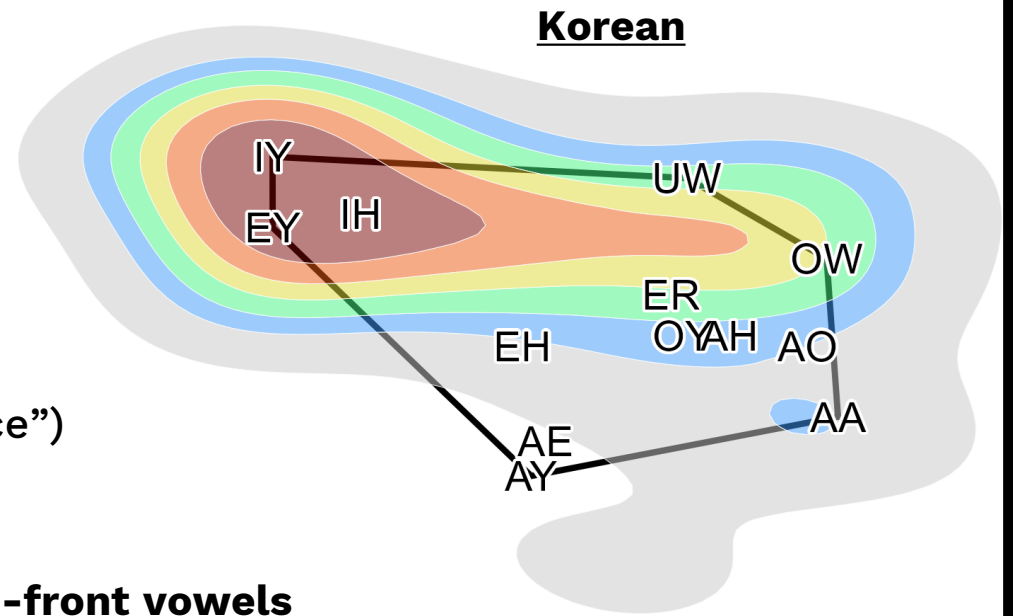
Korean



Vowel substitution visualization

Matches give us “perceived” vowel space by L1 background

Mismatches give us regions of errors (Otter’s “confusion space”)



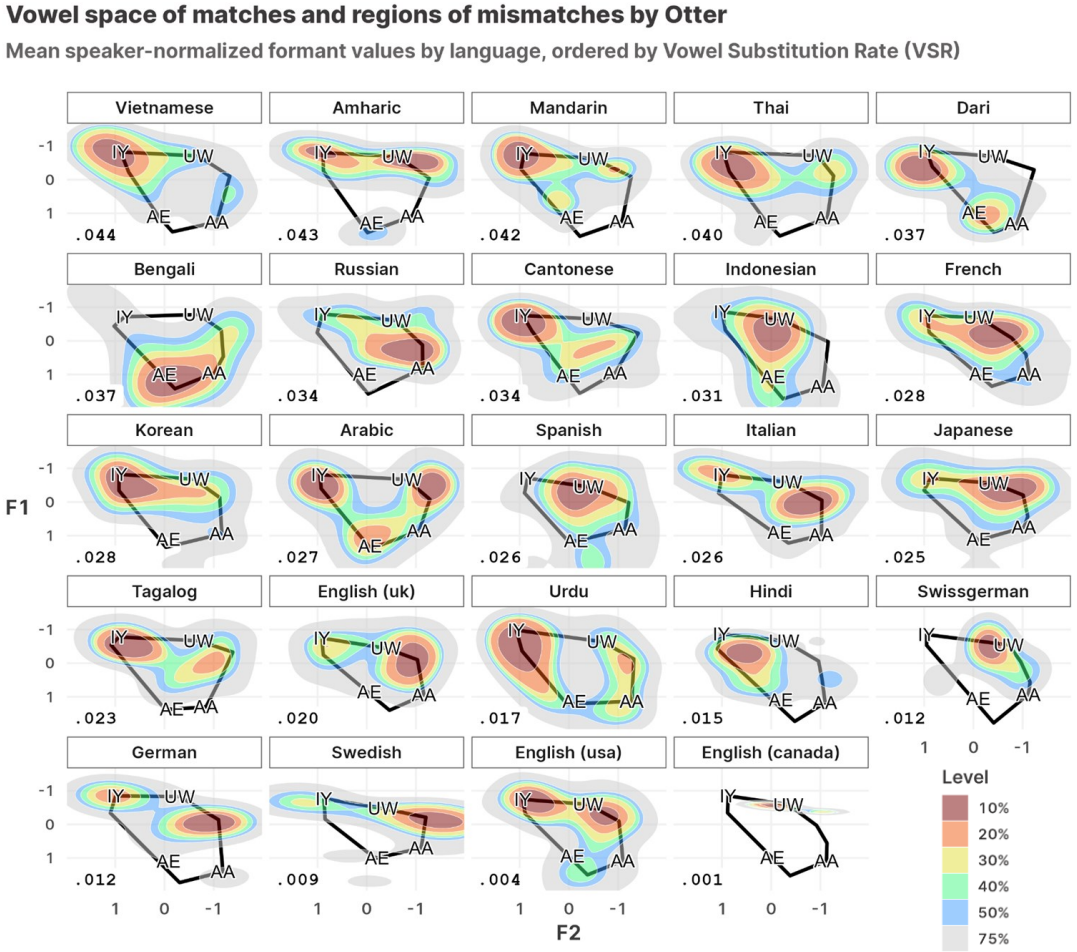
*** Error profile for Korean: high-front vowels**

Crosslinguistic patterns

Vowel errors are **language-specific**; concentrated where the L1 phonology makes **less distinctions** than in English.

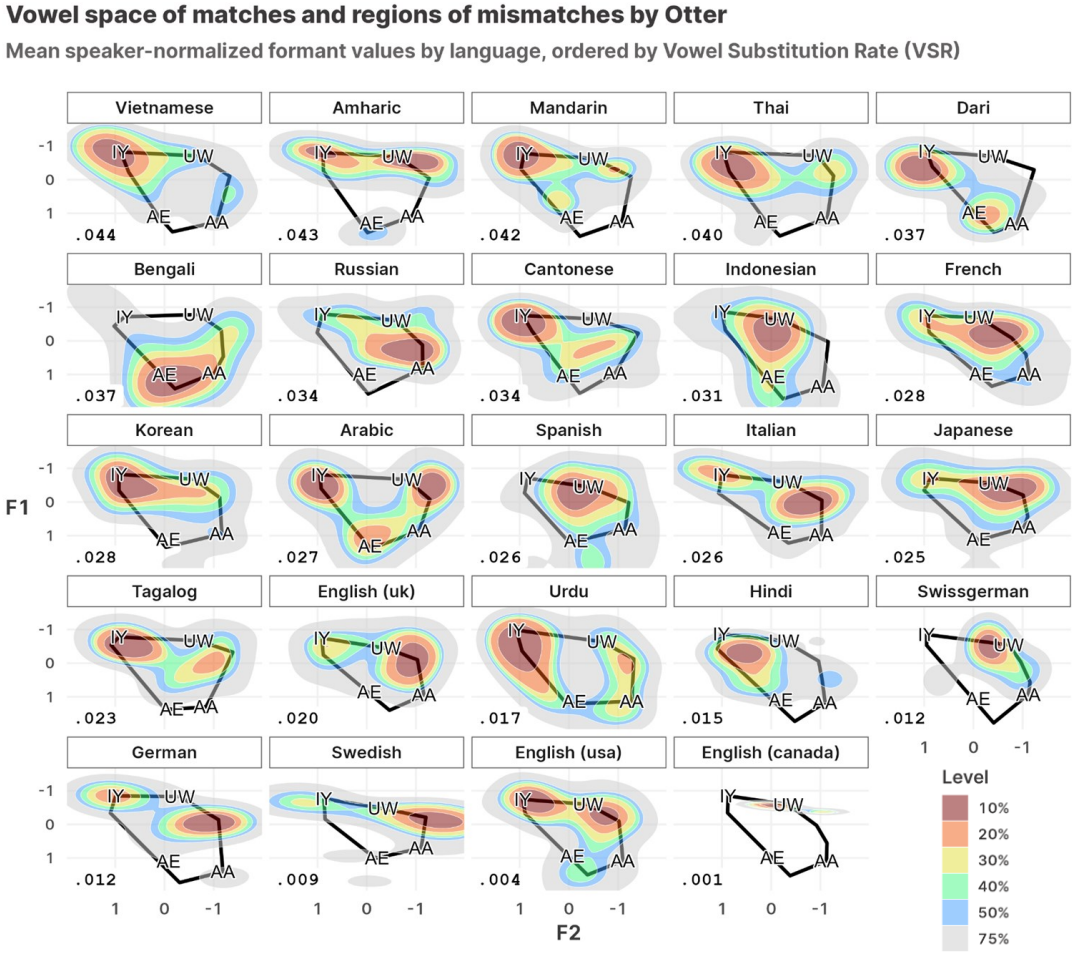
Ex: lack of high-front contrasts in Korean → a particular way of pronouncing /i/ and /ɪ/, in a way that doesn't get picked up.

* Mean age: ~30; Mean AoA: ~10



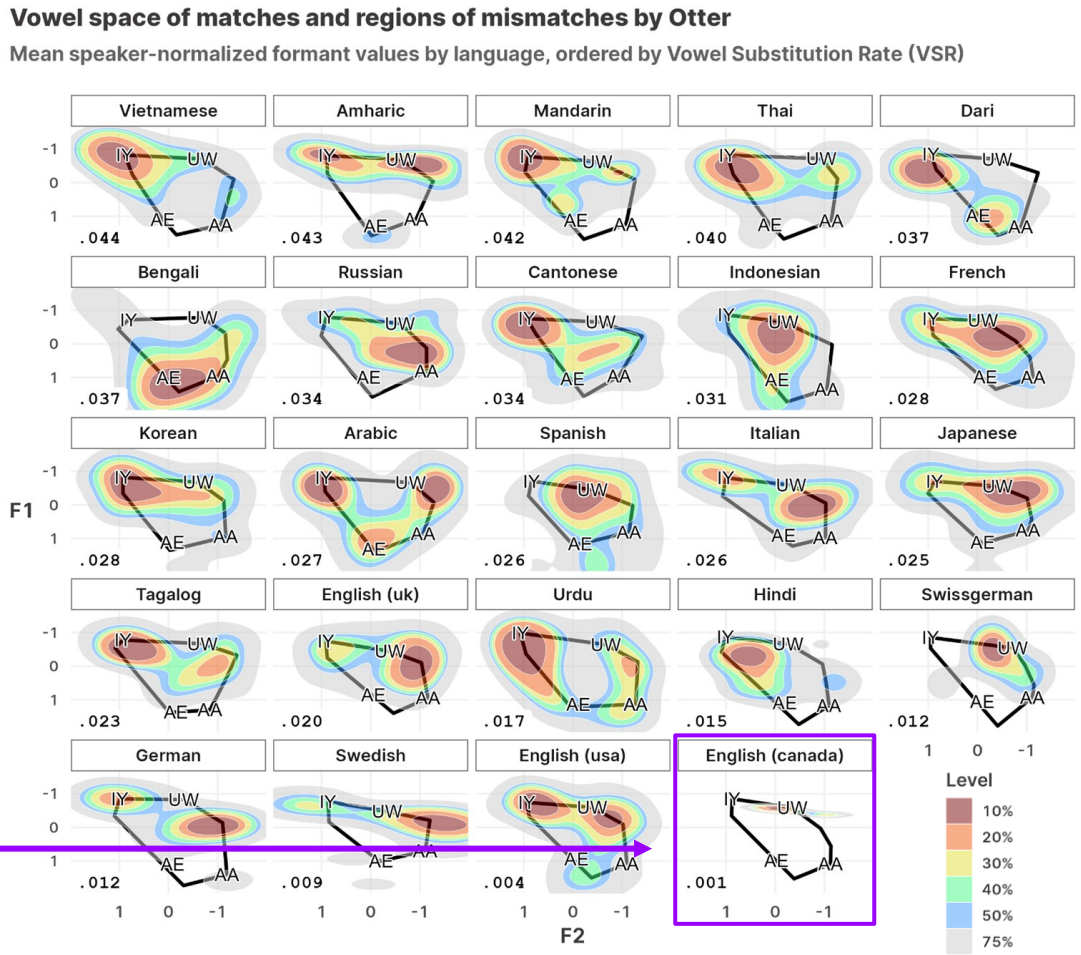
Crosslinguistic patterns

Language	Error Category
Vietnamese	High Front Vowels
Mandarin	High Front/Point Vowels
Thai	High Front Vowels
Korean	High Front Vowels
Hindi	High Front Vowels
Amharic	High Vowels
Cantonese	High Vowels
French	High Vowels
Italian	High Vowels
Tagalog	High Vowels
German	High Vowels
English (USA)	High Vowels
Bengali	Low Vowels
Russian	Low Vowels
Dari	Low/Point Vowels
Indonesian	High Back Vowels
Spanish	High Back Vowels
Japanese	High Back Vowels
English (UK)	High Back Vowels
Swiss German	High Back Vowels
Swedish	High Back Vowels
English (Canada)	High Back Vowels
Arabic	Point Vowels
Urdu	Point Vowels



Crosslinguistic patterns

Language	Error Category
Vietnamese	High Front Vowels
Mandarin	High Front/Point Vowels
Thai	High Front Vowels
Korean	High Front Vowels
Hindi	High Front Vowels
Amharic	High Vowels
Cantonese	High Vowels
French	High Vowels
Italian	High Vowels
Tagalog	High Vowels
German	High Vowels
English (USA)	High Vowels
Bengali	Low Vowels
Russian	Low Vowels
Dari	Low/Point Vowels
Indonesian	High Back Vowels
Spanish	High Back Vowels
Japanese	High Back Vowels
English (UK)	High Back Vowels
Swiss German	High Back Vowels
Swedish	High Back Vowels
English (Canada)	High Back Vowels
Arabic	Point Vowels
Urdu	Point Vowels



Phone-level analysis of errors

3 analyses of **phone-level** substitutions (after converting to ARPABET)

1) Vowel substitution errors:

- E.g., *thick* → *tech* (“TH IH K” -> “T EH K”)

2) Consonant cluster errors:

- E.g., *ask* → *asked* (“AE S K” -> “A S K T”)

3) Consonant voicing errors:

- E.g., *slabs* → *slaps* (“S L AE B Z” -> “S L AE P S”)

2. Consonant errors: Clusters

Onset cluster: please (CCV*)

Coda cluster: asked (*VCC)

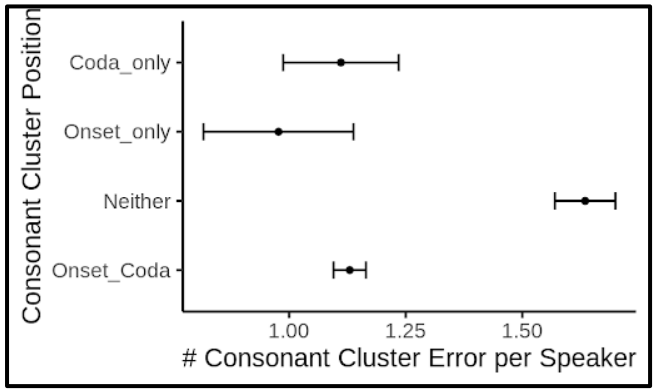


Table 1: Coding of whether a language allows consonant clusters at syllable onset or coda.

	Onset	Coda	Type
English	+	+	*OnsetCoda*
German	+	+	*OnsetCoda*
French	+	+	*OnsetCoda*
Spanish	+	+	*OnsetCoda*
Russian	+	+	*OnsetCoda*
Swedish	+	+	*OnsetCoda*
Swissgerman	+		*OnsetCoda*
Italian	+		*OnsetOnly*
Bengali	+		*OnsetOnly*
Hindi		+	*CodaOnly*
Urdu		+	*CodaOnly*
Dari		+	*CodaOnly*
Mandarin			*Neither*
Cantonese			*Neither*
Japanese			*Neither*
Korean			*Neither*
Thai			*Neither*
Vietnamese			*Neither*
Indonesian			*Neither*
Arabic			*Neither*
Amharic			*Neither*
Tagalog			*Neither*

2. Consonant errors: Clusters

Onset cluster: please (CCV*)

Coda cluster: asked (*VCC)

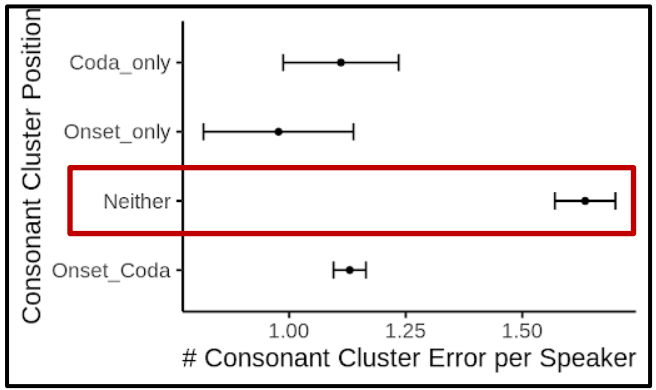


Table 1: Coding of whether a language allows consonant clusters at syllable onset or coda.

	Onset	Coda	Type
English	+	+	*OnsetCoda*
German	+	+	*OnsetCoda*
French	+	+	*OnsetCoda*
Spanish	+	+	*OnsetCoda*
Russian	+	+	*OnsetCoda*
Swedish	+	+	*OnsetCoda*
Swissgerman	+		*OnsetCoda*
Italian	+		*OnsetOnly*
Bengali	+		*OnsetOnly*
Hindi		+	*CodaOnly*
Urdu		+	*CodaOnly*
Dari		+	*CodaOnly*
Mandarin			*Neither*
Cantonese			*Neither*
Japanese			*Neither*
Korean			*Neither*
Thai			*Neither*
Vietnamese			*Neither*
Indonesian			*Neither*
Arabic			*Neither*
Amharic			*Neither*
Tagalog			*Neither*

3. Consonant errors: Voicing

(1) True voicing contrast

(2) Voicing contrast not realized as true voicing

(3) No contrast

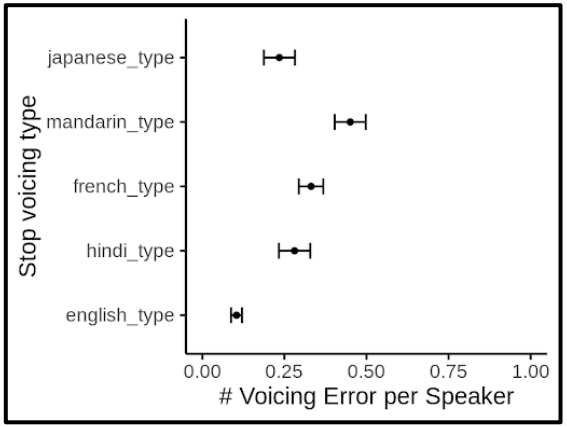


Table 2: Coding of language stop voicing and aspiration contrasts (language in bold is used as group name). Numbers represent each language's category in the typology of voicing and aspiration contrasts.

	Voicing	Aspiration
Hindi	1	1
Vietnamese	1	1
Thai	1	1
Bengali	1	1
Indonesian	1	1
Swedish	1	1
Urdu	1	1
French	1	2
Amharic	1	2
Russian	1	2
Italian	1	2
Arabic	1	2
Dari	1	2
Spanish	1	2
Tagalog	1	2
Japanese	2	1
Korean	2	1
English	2	2
German	2	2
Swissgerman	2	2
Mandarin	3	1
Cantonese	3	1

3. Consonant errors: Voicing

(1) True voicing contrast

(2) Voicing contrast not realized as true voicing

(3) No contrast

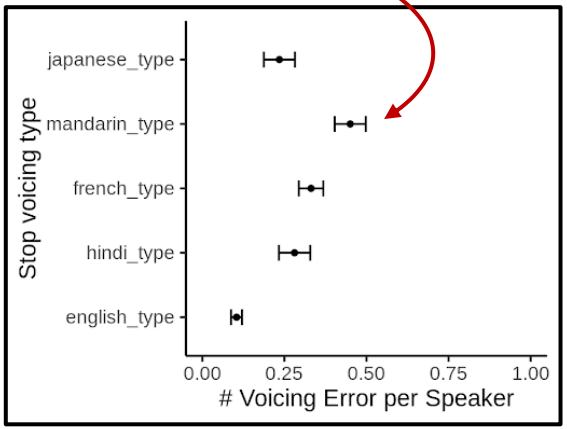


Table 2: Coding of language stop voicing and aspiration contrasts (language in bold is used as group name). Numbers represent each language's category in the typology of voicing and aspiration contrasts.

	Voicing	Aspiration
Hindi	1	1
Vietnamese	1	1
Thai	1	1
Bengali	1	1
Indonesian	1	1
Swedish	1	1
Urdu	1	1
French	1	2
Amharic	1	2
Russian	1	2
Italian	1	2
Arabic	1	2
Dari	1	2
Spanish	1	2
Tagalog	1	2
Japanese	2	1
Korean	2	1
English	2	2
German	2	2
Swissgerman	2	2
Mandarin	3	1
Cantonese	3	1

Conclusion

Not all non-native “accents” are equal: ASR errors vary in type and degree depending on speaker’s L1 (= *language-specific error profiles*).

Otter may be expecting **native-like contrasts**, not just native-like sounds → lower performance even for competent L2 English speakers.

Further exploring the strategies that non-native speakers use to produce L2 phonemic contrasts may help address this performance gap.

Thank you!

Contact: June Choe (yjchoe@sas.upenn.edu)