Language-specific Effects on Automatic Speech Recognition Errors for World Englishes

June Choe, Yiran Chen, May Pik Yu Chan, Aini Li, Xin Gao, Nicole Holliday
Department of Linguistics, University of Pennsylvania, USA
{yjchoe, cheny39, pikyu, liaini, kauhsin, nholl}@sas.upenn.edu

Abstract

Despite recent advancements in automated speech recognition (ASR) technologies, reports of unequal performance across speakers of different demographic groups abound. At the same time, the focus on performance metrics such as the Word Error Rate (WER) in prior studies limit the specificity and scope of recommendations that can be offered for system engineering to overcome these challenges. The current study bridges this gap by investigating the performance of Otter's automatic captioning system on native and non-native English speakers of different language background through a linguistic analysis of segment-level errors. By examining language-specific error profiles for vowels and consonants motivated by linguistic theory, we find that certain categories of errors can be predicted from the phonological structure of a speaker's native language.

1 Introduction

A central concern in the ethics of building natural language processing (NLP) tools is ensuring equity in service and representation through a commitment to linguistic justice (e.g., Blodgett et al., 2020; Hovy and Prabhumoye, 2021). This issue is especially pertinent to automatic speech recognition (ASR) systems used to transcribe natural spoken language into text (Markl and McNulty, 2022). As ASR systems are increasingly being adopted into many aspects of social life (e.g., virtual assistants and automatic captioning), various concerns about the equity of ASR systems have been raised. Ideally, these systems should serve speakers of different demographic backgrounds equally well; however, the existing technologies are not entirely satisfactory. ASR performance has been found to vary by users' dialect (Wheatley and Picone, 1991; Meyer et al., 2020), gender (Adda-Decker and Lamel, 2005; Tatman and Kasten, 2017; Tatman, 2017; Sawalha and Abushariah, 2013; Boito et al., 2022) and ethnicity (Koenecke et al., 2020; Martin and Tang, 2020). For instance, ASR systems designed for American English typically perform worse for non-white speakers than for white speakers (Tatman and Kasten, 2017; Koenecke et al., 2020). Such inequalities result in certain groups of speakers being better represented than others, and may even further exacerbate existing inequalities in society.

At the same time, less is known about how the performance of ASR systems can vary for second-language (L2) English speakers, a particularly vulnerable population of English speakers with diverse backgrounds. A recent work by Chan et al. (2022) examined how Otter, a popular automatic transcription system, performs on L1 (native) and L2 (second-language) speakers of 24 English varieties. Not only do the English varieties supported by Otter have lower Word Error Rates (WER) compared to the unsupported varieties, gaps in performance is also driven by an independent effect of the structure of a speaker's first language – Otter performs worse on English spoken by L1 speakers of a tonal language (e.g., Mandarin).

While the Word Error Rate has been widely adopted in these studies due to the ease of quantifying system bias, this one-dimensional measure of performance is inadequate if the aim is to disentangle different types of errors that give rise to discrepancies in performance between speakers. For example, word-level errors can be driven by an error in one or multiple sound segments, and certain errors for consonants and vowels may be more common for speakers of one variety than speakers of another. Studying these details is useful because these linguistic categories are well-studied theoretical constructs and empirical phenomena in the research on sociolinguistic variation (e.g., Koenecke et al., 2020; Wassink et al., 2022) and L2 transfer (e.g., Corder, 1983; Dechert and Raupach, 1989; Best et al., 1994). Moreover, such linguistically-motivated features like conso-

Proceedings of the 29th International Conference on Computational Linguistics, pages 7177–7186
October 12–17, 2022.

nant voicing and vowel height have acoustic correlates, which means that insights from studying system errors in terms of phonological variation and processes can be translated to applied system engineering.

Therefore, the current study aims to systematically investigate whether variations in language structure among native and non-native language speakers of English are tied to different types and degrees of transcriptions errors. First, we introduce and motivate a segment-level error analysis built on traditional error-rate algorithms, which allows an analysis of errors beyond the single dimension of performance. We then investigate the error profile for consonants and vowels across various English varieties, to determine whether the phonological structure of a speaker's native language/variety is predictably tied to certain types of errors that can be captured in terms of phonological processes.

2 Materials and Methods

2.1 Speech Recognition System

We evaluate the performance of Otter, a speech recognition platform for automatic transcriptions that claims to support multiple varieties of English including "(southern) American, Canadian, Indian, Chinese, Russian, British, Scottish, Italian, German, Swiss, Irish, Scandinavian, and other European accents" (Lai, 2021).[1] Fittingly, Otter's live captioning system is used by popular video conferencing platforms that reach broad international audiences, such as Zoom. Otter's global user-base and its incorporation into educational and professional settings make it an ideal candidate for investigating whether there exists language-specific biases in ASR performance for non-native (L2) speakers of English, and if so, how these biases relate to the phonological structure of the speakers' native (L1) languages.

2.2 Corpus

The data analyzed in this study are collected from the Speech Accent Archive at "http://accent.gmu.edu" (Weinberger, 2015), a comprehensive speech corpus of nearly three thousand recordings from English speakers of diverse geographical and language backgrounds. Each entry in the corpus is a speaker reading the same passage ("Please Call Stella") at a table in a quiet room, seated approximately 8-10 inches from the microphone. The passage is designed to include words that elicit all sound segments in English, which allows for a direct comparison between speakers of different language backgrounds. The passage is 77-words long and recordings were around 30 seconds long on average.

Each recording in the corpus is accompanied by a list of demographic information about the speaker, including birthplace, age at the time of recording, sex, native language, age of English onset, English learning method (naturalistic vs. academic), among others. Notably, some of the recordings are also coded by trained linguists for accent features (e.g., vowel shortening), which used to motivate the error categories for evaluating ASR performance, as will be described in detail in later sections.

Following Chan et al. (2022), we filtered the corpus based on the following criteria: (1) Only varieties of English that are either listed as supported English varieties by Otter, or (2) have recording entries from at least 10 speakers. To balance the effect of system training in the data, we selected eleven varieties supported by Otter and sampled another eleven from the remaining non-supported varieties, for a total of 1,227 speakers/recordings.

All recordings were re-sampled to 22,050 Hz and concatenated with one-second pause inserted between each recording. The resulting 9.5-hour audio file was split into 4-hour chunks (the maximum file size permitted by Otter) before uploading to Otter. Otter's speaker detection system split the transcription output by speaker, though with occasional errors (10%) where multiple recordings of the passage were merged and determined as coming from the same speaker. Two human annotators corrected these errors independently, reaching 99.8% agreement, and a third annotator resolved the conflicts. The transcriptions with corrected alignments to speakers were entered into the error analysis.

2.3 Segment-level Error Analysis

We first ran a Word Error Rate (WER) algorithm by identifying word-level insertions, deletions, and substitutions for each speaker's transcribed string of words ("observed") compared to the reading passage ("truth"). Then, a Phone Error Rate (PER) was calculated for each speaker by first converting the word-level insertions, deletions, and substitutions to phones (sound seg-

[1] Note that while Otter uses the term "accents", we will be adopting the word "varieties" to refer to Englishes from both native and non-native speakers in this paper.

ments) using the CMU English Pronouncing Dictionary ("http://www.speech.cs.cmu.edu/cgi-bin/cmudict"), and passing the resulting string of phones to the same error rate algorithm.

To illustrate, consider the word "ask" from the reading passage, which has an entry in the CMU dictionary as "AE S K", the machine-readable ARPABET transcription representing the International Phonetic Alphabet (IPA) transcription /æsk/. A word-level deletion of "ask" in a transcribed output constitutes three phone errors (counting "AE", "S", and "K"), while a word-level substitution of "ask" with "asked" constitutes just one phone error (word-final insertion of "T" representing /t/). Thus, the phone-level measure of PER can be more precise about the egregiousness of errors made by an ASR system compared to the more traditional benchmark of WER which lacks such sensitivity (Aksënova et al., 2021; Wassink et al., 2022). As discussed earlier, a prior study by Chan et al. (2022) finds an effect of training and language structure across both measures, but the specific source of this disparity remains under-investigated, especially where the distribution of phone-level errors are concerned.

Therefore, the analyses in this paper go beyond the singular measures of performance (WER and PER) to investigate the language-specific profile of segment errors, using the phone-level substitutions identified by the PER algorithm. Phone substitutions are interesting from a linguistic standpoint because they allow an analysis of transcription errors in terms of phonological processes; a necessary first step for interdisciplinary work incorporating domain knowledge from linguistics (e.g., second language acquisition, language typology, and sociolinguistic variation). For example, a phone substitution of the vowel /ɪ/ as in "bit" to /i/ as in "beat" represents vowel lengthening. If a system consistently makes errors for a speaker's production of this short /ɪ/ vowel, then that error profile, combined with the language background of the speaker, can be identified as an actionable area of improvement for accent adaptation algorithms in ASR. The next two sections split the analysis of segment errors by consonants and vowels, with the specific methods for each detailed therein.

3 Consonant Analysis

In this section, we zoom in on the consonant errors in the Otter transcription of the same passage read by L2 English speakers with different L1s. It has been established that phonological features of L1 are likely transferred into speakers' L2 (Dechert and Raupach, 1989; Corder, 1983; Best et al., 1994). For example, given that Japanese only allows CV syllables, native speakers of Japanese have more difficulty with complex consonant cluster pronunciation (e.g. "sixth") when speaking English. Therefore, it is not surprising if we find traces of different L1s in our current data set of L2 Englishes. However, what remains under-explored is whether these phonological differences of L1s will be directly reflected in the kinds of consonant errors ASR algorithms make on these nonnative Englishes, especially one that claimed to have trained on non-native accents of English (Lai, 2021). We specifically test this in the following section.

3.1 Methods

With the help of the PER algorithm introduced in Section 2, we identified 2382 errors involving consonant substitutions. Given our primary interest in the relationship between the phonological structure of different L1s and the distribution of consonant errors, we focused on two types of errors that are the most prevalent in the data, as the robust number of tokens allows us to observe cross-linguistic variation. These two types are Cluster errors (e.g. transcribing "ask" as "asked") and Voicing errors (e.g. transcribing "bag" as "back"). We test two hypotheses for the distribution of cluster and Voicing errors, respectively.

First, we hypothesize that the syllable structure of speakers' L1 language (specifically, whether it allows consonant clusters) drives the rate of Cluster errors in their L2 English. The more different a speaker's L1 is from English in this respect, the more likely it is for Otter to make Cluster errors for that speaker. To test this hypothesis, we coded each L1 language in terms of whether they allow consonant clusters at syllable onset (yes vs. no) and syllable coda (yes vs. no) based on descriptions of the phonology of these languages in the literature (Ohala, 1983; Potet, 1995; Mahootian and Gebhardt, 1997; International Phonetic Association, 1999; Mazhar and Ranjha, 2012; Sircar and Nag, 2013; among others). We then categorized languages into four types: those allowing consonant clusters at both syllable onset and coda (Onset-Coda), those allowing such clusters only at syllable onset (Onset-only), only at syllable coda

(Coda-only) and at neither position (Neither) (see Table 1). Since languages that allow consonant clusters at neither position would be the most different from English, we predict Otter to perform the worst on the production of English consonant clusters by native speakers of languages in the Neither category. To statistically test for this effect, we fitted a linear mixed-effect model predicting the number of Cluster errors with L1 syllable structure type (4 levels, sum coded) as a fixed effect and a random intercept by language group, using the *lme4* package in R (Bates et al., 2015).

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*

Second, we hypothesize that the realization of the voicing contrast in the L1 language should drive the rate of Voicing errors in L2 English. To test this hypothesis, we coded each L1 language in terms of how their voicing contrasts is realized into three levels: those that have true voicing contrasts (1), those that have voicing contrast but are not realized as true voicing (2) and those that have no voicing contrast (3). Additionally, we coded whether there is a phonemic aspiration contrast in the L1 language (1: yes vs 2: no). The coding was based on phonetic and phonological descriptions of these languages in the linguistic literature (Henderson, 1972; Thelwall and Sa'Adeddin, 1990; International Phonetic Association, 1999; Fleischer and Schmid, 2006; Petrova et al., 2006; Mikuteit and Reetz, 2007; Soderberg and Olson, 2008; Kramer, 2009; Gallagher, 2010; Labrune, 2012; Tranová, 2016). We further categorized the languages into five types according to these two dimensions. The complete coding and categorization of all L1 languages can be found in Table 2. To statistically test for this effect, we fitted a linear mixed-effect model predicting the number of Voicing errors with errors with L1 stop contrast type (5 levels, sum coded) as a fixed effect and a random intercept by language group. For simplicity, we refer to each group by a representative language in the group (e.g. the Hindi-type) in the following analysis.

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

3.2 Results

3.2.1 Syllable structure and Cluster Error

As shown in Figure 1, we find that L1 languages that do not allow consonant clusters at either sylla-

ble onset or coda receive significantly more cluster errors per speaker in Otter transcription (β = 0.429, SE = 0.118, p = 0.002). This is consistent with our prediction that the rate of consonant cluster errors are driven by the extent to which consonant clusters are licensed in the syllable structure of a speaker's native language. At the same time, L1 languages that allow consonant clusters at more restricted positions (Onset only & Coda only) are not significantly different from languages that allow them in both positions (Onset-Coda). In sum, we observe degraded performance on the transcription of words with consonant clusters driven by the degree of difference in syllable structure between English and speakers' L1 languages.

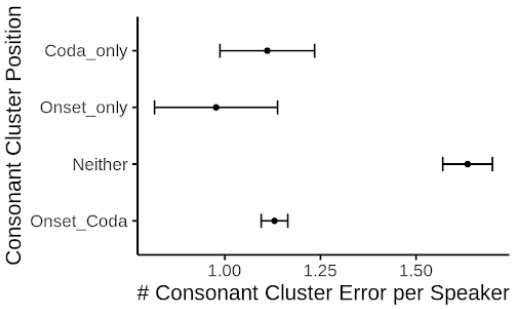

Figure 1: The number of Cluster errors per speaker in L2 English transcription by L1 consonant cluster type.

3.2.2 Stop Voicing and Voicing Error

As shown in Figure 2, we find that L1 languages that realize stop contrasts just like English receive the least voicing errors per speaker in Otter transcription (β = -0.399, SE = 0.138, p = 0.017). Additionally, find a marginal effect of Otter transcriptions generating more voicing errors for Mandarin-type languages (β = 0.302, SE = 0.150, p = 0.075). One explanation for this effect is that, as shown in Table 2, Mandarin-type languages are the most different from English in terms of the phonological structure of the stop consonants - instead of a phonemic voicing contrast, there is a phonemic aspiration contrast. Thus, we find evidence for our hypothesis that the degree of difference in the phonological structure of a speaker's L1 compared to English is directly reflected in the accuracy of automatic transcriptions for their production of L2 English.

4 Vowel Analysis

In the following section, we turn to the distribution of vowel errors in the Otter transcriptions. We ana-

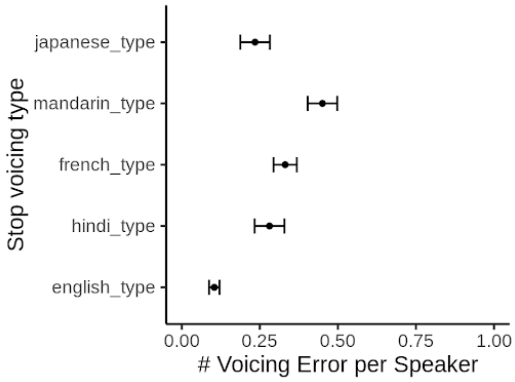

Figure 2: The number of Voicing errors per speaker in L2 English transcription by L1 stop contrast type.

lyze the distribution of vowel errors in the acoustic space in context of the typology of vowel systems among the L1 languages represented in the data. Specifically, we focus on vowel substitution errors (as opposed to insertions and deletions) as they allow us to explore Otter's errors in terms of well-studied phonological processes.

4.1 Methods

All 1,227 sound files selected for analysis were ran through the Penn Phonetics Lab Forced Aligner (Yuan et al., 2008) to align our recordings at the segmental level. We then extracted the first two formants (F1 and F2) at the midpoint all vowel tokens using the LPC (burg) function in Praat (Boersma, 2006) and z-scored all formant measurements by speaker. While formant measurements at the 25% and 75% were also extracted initially to analyze diphthongal patterns, there were insufficient diphthong tokens in the passage to draw conclusive results. Therefore, we focus our analysis on the acoustic measures of monophthong tokens.

We visualize our acoustic analysis in Figure 3. The x-axis is the z-scored F2 and the y-axis is the z-scored F1; axes are plotted in reverse to match the height and backness dimensions of the physical vowel space. The figure highlights two vowel analyses conducted for each language background. First is Otter's perceived vowel space, represented by the black solid lines that connect the four edges of the vowel space. The perceived vowel space is constructed from cases where the "truth" vowel and the "observed" vowel match - i.e., when Otter correctly transcribes the intended vowel that was produced. Second is Otter's regions of error, represented by a color contour imposed on top of the perceived vowel space. The error regions are constructed

from the formant values of vowels involved in substitution errors - i.e., when the transcribed vowel does not match the intended vowel that a speaker produced. In sum, we analyzed the acoustic profile of matches and mismatches between "true" and "observed" vowel tokens for each language background.

Based on the distribution of vowel errors from the acoustic analysis, we categorized the L1 language varieties into five vowel error types. Depending on where errors are concentrated in the vowel space, each language was assigned to at least one of the following: (i) high front vowels, (ii) high back vowels, (iii) high vowels, (iv) low vowels, and (v) point vowels. The vowel error categories are summarized in Table 3.

Table 3: Vowel Error Category.

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

4.2 Results

The concentration of errors within the vowel spaces reveals two main trends. First, languages with fewer vowel distinctions than English at a phonemic level, in either the entire vowel space or in certain parts of the vowel space, have higher vowel substitution errors. As predicted, the distribution of the types of vowel errors are language-specific, such that errors concentrate on specific regions of the vowel space where the L1 phonology makes less distinctions than in English. This is not surprising given the literature on second language acquisition such as the Perceptual Assimilation Model (Best et al., 1994), which posits that listeners perceive non-native phones in terms of the similarities or dissimilarities of the phones to their native phonemic contrasts. For instance, Vietnamese (Kirby, 2011) and Thai (International Phonetic Association, 1999) have comparable phonological vowel spaces, and neither has a tense-lax contrast for high front vowels (which English does have). Consequently, we find vowel substitution errors concentrated in the high front region of the vowel space. At the same time, though both languages also lack the tense-lax contrast in the high back region, the existence of a roundness contrast may have been used to disambiguate high back vowels in their L2 English pronunciation. When a Vietnamese speaker says the words "thick slabs" (transcribed as /θɪk slæbz/ in the CMU dictionary) for example, Otter transcribes it as "techs lab" /tɛks læb/, replacing the /ɪ/ vowel with /ɛ/, which is in the Vietnamese vowel inventory. Another example of languages with fewer vowel distinctions than English is Arabic, which has been described to contrast three main monophthongal vowel qualities, also referred to as point vowels. We find that Otter primarily misidentifies tokens spoken at the point regions of the vowel space by Arabic speakers of English. For example, when an Arabic speaker produces "Stella" (/stɛlə/), the Otter transcription confused /ɛ/ as /i/, yielding the transcription "stealer" /stilɚ/. These results confirm existing findings for speakers localizing their pronunciation of non-native phones to the categories available in their language. Critically, we find that this also drives language-specific phone-level transcription errors from Otter.

Second, languages that have a similar phonological structure to that of English in its vowel inventory have an overall lower vowel substitution rate (and lower word error rate, writ large). Interestingly, our predictions are still borne out in the vowel error profile of these languages for which Otter performs well. An example of this is German, which was categorized into the high vowel error category. The vowel inventory of German is simi-

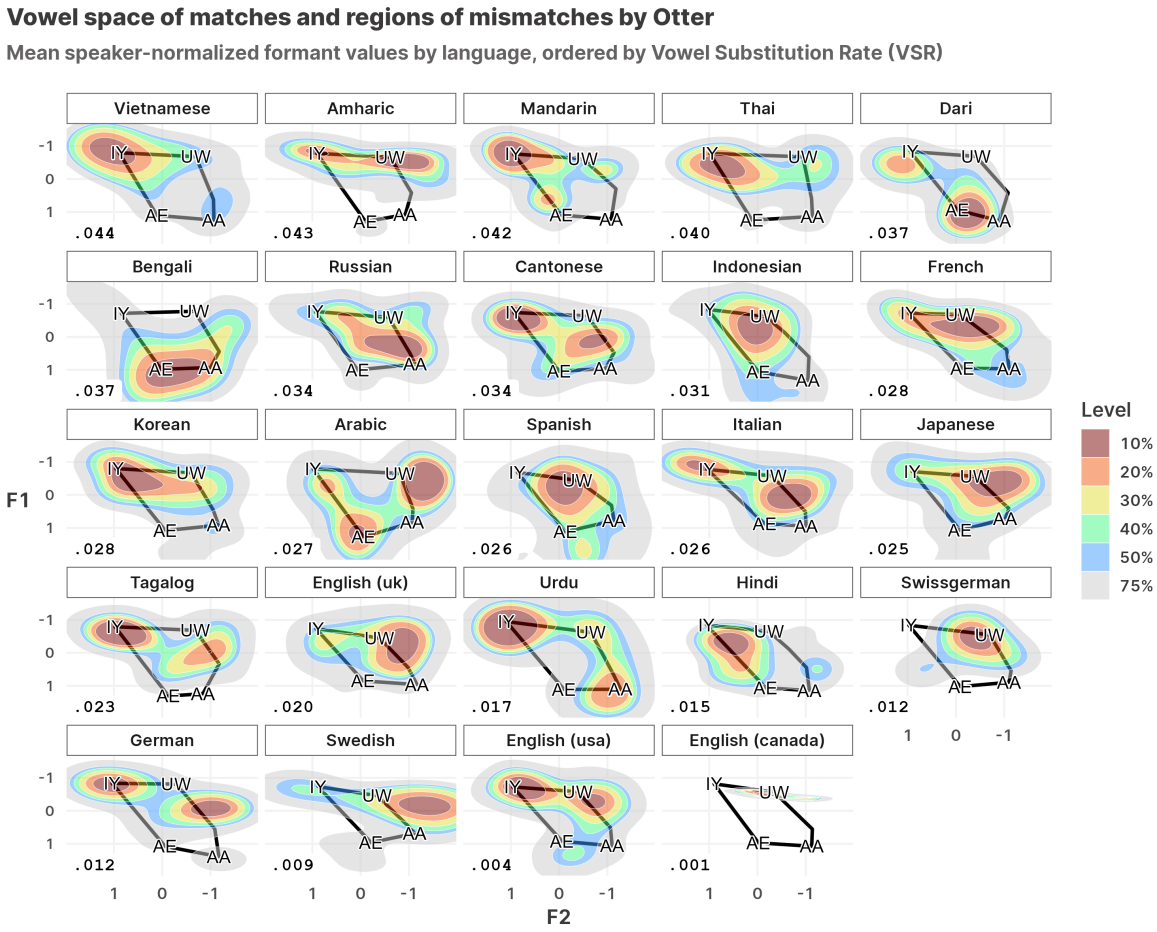

Figure 3: A speaker-normalized (z-scored) F1-by-F2 vowel plot, split by speakers' L1. The black polygon represents the vowel space detected by Otter, composed of correctly identified vowels. The density contours are constructed from the distribution of incorrectly identified vowels via a highest-density-region (HDR) estimation (Otto and Kahle, 2022), where the fill color represents the percentage of vowel errors occurring within a specific region. Languages are ordered by their vowel substitution rate (annotated in the bottom left of each panel), where a higher number reflects worse performance on vowel identification.

lar to that of English in many ways; for example, it contrasts both tense and lax high vowels, contrasts open-mid and closed-mid cardinal vowels, and includes the unstressed vowel schwa (International Phonetic Association, 1999). Unlike English, however, German lacks the back vowel /ʌ/, which may explain the concentration of back vowel errors in the vowel space. For example, one German speaker's pronunciation of "brother" /brʌðɚ/ was transcribed by Otter as "product" /pɹɑdəkt/ instead, where the /ʌ/ vowel was replaced by another, lower back vowel /ɑ/.

Lastly, we also find predictable errors that reflect well-documented sociolinguistic variation among L1 English speakers. For example, the vowel substitution errors made by Otter in Canadian English are highly localized in the high vowel region, which likely reflects Canadian raising (Chambers, 1973). In sum, the acoustic analysis of vowel substitution errors show that the phonological structure of a native language's vowel spaces can inform us of specific gaps in the performance of ASR systems.

5 Discussion

This study investigated the performance of Otter's automatic captioning system on native and non-native English speakers of different language backgrounds through a linguistic analysis of segment-level substitution errors. We proposed that understanding language-specific error profiles is crucial to preempting predictable system errors. In our analysis of consonant and vowel errors motivated by phonological theory, we report the following findings.

Results from our consonant analysis show that

the phonological structure of a non-native English speaker's first language predicts the types of consonant errors that are dominant in the automatic transcription of their production of English. Specifically, we find higher rates of consonant cluster errors in the transcription of speakers whose native language does not allow consonant clusters. Similarly, we find higher rates of stop voicing errors for speakers whose first language has aspiration contrasts, as opposed to voicing contrasts like English.

Results from our vowel analysis show that the distribution of vowel substitution errors patterns with the structure of the vowel inventory of a non-native English speaker's first language. Specifically, we find three main patterns. First, the transcriptions of non-native speakers whose first language make fewer vowel distinctions than English show predictable regions of error in the vowel space. Second, when languages have similar phonological structures compared to English, the frequency of vowel substitution errors tend to be lower. Third, known sociolinguistic variation even among native speakers of English also predict vowel errors in a similar manner.

Understanding how the sound patterns of a speaker's native language and/or variety of English affect the performance of ASR systems on specific categories of sounds is an important first step towards designing robust yet flexible acoustic models which can detect and adapt to varieties without reducing structural differences to correlated demographic information like race and ethnicity (Tatman, 2020). Recent successes in the design of accent adaptation algorithms support this view, such as the automatic accent identification model by Najafian and Russell (2020) trained on regional varieties of British English. Furthermore, the focus on adapting to differences in linguistic structure grounded in phonological theory is crucial to the generalizability of ASR systems, which is an important consideration for providing transcription services to speakers of minority languages. For example, an ASR model designed to learn and leverage language-specific phonemic contrasts by Li et al. (2020) vastly improved phone-level accuracy on very small corpora (1k utterances) of two indigenous low-resource languages, Inuktitut and Tusom. Our study contributes to this momentum by offering insights into the phonological and acoustic nature of transcription errors for speakers of different language backgrounds. Such research on language-specific error profiles can motivate the design of adaptation algorithms for supporting non-native English speakers of various language backgrounds.

Lastly, improving the performance of ASR systems for non-native speakers of English is an important task not simply for the sake of catering to a large user-base given the status of English as the de facto lingua franca with far more L2 than L1 English speakers in the world (SIL International, 2022), but also because L2 speakers of English are an especially vulnerable population facing specific, and often invisible, harms from the prevailing stereotypes of being uninterpretable in speech and lacking education and proficiency in English (Lippi-Green, 2011). Many existing applications of ASR systems are ill-equipped to work with speech input from non-standard varieties, as evidenced by accumulating cases of discrimination against non-native speakers of English across all levels of harm (Blodgett et al., 2020). For example, allocational harms have been reported for even life-or-death situations such as in voice command systems for roadside vehicle assistance (Wassink et al., 2022) and for medical diagnoses and records management in healthcare systems (Lee, 2021). Moreover, representational harms from stereotypes of unintelligibility are perpetuated by systems that claim to work on a language while neglecting how the system might perform differently among sociolects (Aksënova et al., 2021). As an extreme example of this ideology, some recruiting firms have claimed to screen and rank job applications by passing their voice data through off-the-shelf ASR systems and using the interpretability of the transcription output itself as proxies for friendliness and communication skills, putting non-native English speakers at a disadvantage (Raghavan et al., 2020). In these ways, unchecked bias against non-native speakers of English in ASR systems reinforce social inequalities and simultaneously dismiss real cases of need for accommodation.

6 Conclusion

In this study, we examined the language-specific error profiles of native and non-native English speakers of diverse language backgrounds. A segment-level analysis of consonant and vowel errors made by Otter's transcription system reveals that certain categories of errors are predictable from the phonological structure of a speaker's native language.

Thus, we demonstrate the fruitfulness of applying a linguistic analysis to transcription errors, informed by general phonological theory as well as specific literature from relevant domains such as sociolinguistic variation and second language acquisition. Findings inform the design and maintenance of new and existing ASR systems for adapting to non-native speakers of English and speakers of non-standard English varieties.

References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Ninth European Conference on Speech Communication and Technology*.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, Online. Association for Computational Linguistics.

International Phonetic Association, International Phonetic Association Staff, et al. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Catherine T. Best et al. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, 167(224):233–277.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint. arXiv:2005.14050*.

Paul Boersma. 2006. Praat: doing phonetics by computer. *http://www. praat. org/*.

Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève. 2022. A study of gender impact in self-supervised models for speech-to-text systems. *arXiv preprint arXiv:2204.01397*.

Jack K. Chambers. 1973. Canadian raising. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 18(2):113–135.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. 2022. Training and typological bias in asr performance for world englishes. In *Proceedings of the 23rd Conference of the International Speech Communication Association*.

Stephen Pit Corder. 1983. A role for the mother tongue. *Language transfer in language learning*, 1:85–97.

Hans-Wilhelm Dechert and Manfred Raupach. 1989. *Transfer in language production*. Praeger.

Jürg Fleischer and Stephan Schmid. 2006. Zurich german. *Journal of the International Phonetic Association*, 36(2):243–253.

Gillian Elizabeth Scott Gallagher. 2010. *The perceptual basis of long-distance laryngeal restrictions*. Ph.D. thesis, Massachusetts Institute of Technology.

Michael Magnus Thyne Henderson. 1972. *Dari (Kabul Persian) Phonology*. The University of Wisconsin-Madison.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

James P. Kirby. 2011. Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association*, 41(3):381–392.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Martin Kramer. 2009. *The phonology of Italian*. OUP Oxford.

Laurence Labrune. 2012. *The phonology of Japanese*. Oxford University Press.

Allen Lai. 2021. Supported languages. *Otter.ai Help Center*.

Dave Lee. 2021. The next big tech battle: Amazon's bet on healthcare begins to take shape. *Financial Times*.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253.

Rosina Lippi-Green. 2011. *English with accents: Language, ideology, and discrimination in the United States*. Routledge, New York.

Shahrzad Mahootian and Lewis Gebhardt. 1997. Persian (descriptive grammars). *London: Routledge. doi*, 10:9780203192887.

Nina Markl and Stephen Joseph McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in asr. *arXiv preprint arXiv:2202.12603*.

Joshua L. Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual "be". In *INTERSPEECH*, pages 626–630.

Iqbal Mazhar and Mazhar Ranjha. 2012. Urdu syllable: Templates and constraints. In *Proceeding of the Conference on Language and technology 2012*.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6462–6468.

Simone Mikuteit and Henning Reetz. 2007. Caught in the ACT: The timing of aspiration and voicing in East Bengali. *Language and speech*, 50(2):247–277.

Maryam Najafian and Martin Russell. 2020. Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*, 122:44–55.

Manjari Ohala. 1983. *Aspects of Hindi phonology*. Motilal Banarsidass Publishers Pvt. Limited.

James Otto and David Kahle. 2022. *ggdensity: Interpretable Bivariate Density Visualization with 'ggplot2'*. Https://jamesotto852.github.io/ggdensity/, https://github.com/jamesotto852/ggdensity/.

Olga Petrova, Rosemary Plapp, Catherine Ringen, and Szilárd Szentgyörgyi. 2006. *Voice and aspiration: Evidence from Russian, Hungarian, German, Swedish, and Turkish*. Walter de Gruyter.

Jean-Paul G. Potet. 1995. Tagalog monosyllabic roots. *Oceanic linguistics*, pages 345–374.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481.

Majdi Sawalha and Mohammad Abushariah. 2013. The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus. In *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*. Leeds.

SIL International. 2022. Ethnologue: English.

Shruti Sircar and Sonali Nag. 2013. 19 akshara–syllable mappings in bengali: a language-specific skill for reading. *South and Southeast Asian psycholinguistics*, page 202.

Craig D. Soderberg and Kenneth S. Olson. 2008. Indonesian. *Journal of the International Phonetic Association*, 38(2):209–213.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

Rachael Tatman. 2020. Sociolinguistic variation and automatic speech recognition: Challenges and approaches. In *2020 Annual Meeting*. AAAS.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech*, pages 934–938.

Robin Thelwall and M. Akram Sa'Adeddin. 1990. Arabic. *Journal of the International Phonetic Association*, 20(2):37–39.

Lenka Tranová. 2016. *The voicing contrast in Vietnamese English*. Univerzita Karlova, Filozofická fakulta.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

Steven Weinberger. 2015. Speech accent archive. george mason university. *http://accent.gmu.edu*.

Barbara Wheatley and Joseph Picone. 1991. Voice Across America: Toward robust speaker-independent speech recognition for telecommunications applications. *Digital Signal Processing*, 1(2):45–63.

Jiahong Yuan, Mark Liberman, et al. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.