ACCENTS IN SPEECH RECOGNITION THROUGH THE LENS OF A WORLD ENGLISHES EVALUATION SET

MIGUEL DEL RÍO[1], COREY MILLER[1], JÁN PROFANT[1], JENNIFER DREXLER-FOX[1], QUINN McNAMARA[1], NISHCHAL BHANDARI[1], NATALIE DELWORTH[1], ILYA PIRKIN[1], MIGÜEL JETTÉ[1], SHIPRA CHANDRA[2], PETER HA[3], RYAN WESTERMAN[4]

[1]Rev.com, [2]Walgreens, [3]Northwestern University, [4]Zoom
miguel.delrio@rev.com, corey.miller@rev.com

Abstract

Automatic Speech Recognition (ASR) systems generalize poorly on accented speech, creating bias issues for users and providers. The phonetic and linguistic variability of accents present challenges for ASR systems in both data collection and modeling strategies. We present two promising approaches to accented speech recognition— custom vocabulary and multilingual modeling— and highlight key challenges in the space. Among these, lack of a standard benchmark makes research and comparison difficult. We address this with a novel corpus of accented speech: Earnings-22, A 125 file, 119 hour corpus of English-language earnings calls gathered from global companies. We compare commercial models showing variation in performance when taking country of origin into consideration and demonstrate targeted improvements using the methods we introduce.

Keywords: accents, dialects, speech recognition, bias, multilingual

1. Introduction

Speech recognition has reached high levels of performance on standardized datasets (Xiong et al., 2016) and widespread use in various real-world settings. However, ASR performance can degrade significantly on certain accents (Koenecke et al., 2020), unsurprisingly echoing accent-related bias in the real world (Lippi-Green, 2012). Thus, accent-robustness is needed for speech recognition to be solved in the wild. Generalizing speech recognition across dialects is a hard problem for real-world speech systems. Dialects are variations within a language that differ in geographical regions and social groups, which can be distinguished by traits of grammar, vocabulary and what can be called "accent": phonetics and phonology. The terms "accent" and "dialect" are sometimes used interchangeably in the context of speech recognition as both can describe speaker-specific variation in a given language.

Accurate speech transcript data is hard to source, with accurately labeled and transcribed speech in non-standard accents even more so. This lack of data means standard approaches to speech recognition perform poorly on lower resource accents. A lack of common training and test suites in the research community makes it difficult to compare similar approaches. This paper starts with a review of the challenges facing the development of accent-robust ASR and then introduces an international English test suite called Earnings-22 that aims to help developers identify accent-related issues with their systems. We then discuss a handful of mitigation strategies to improve accent robustness using Earnings-22 as an example.

2. Challenges of accent variation

This section will review some of the challenges of accent variability for humans and ASR. We suggest two strategies for giving ASR a better chance at handling this: custom vocabulary and multilingual modeling. These strategies will be applied to Earnings-22 later in the paper.

2.1 Are some words more challenging than others?

When humans or machines confront linguistic variation, we seek to understand what words cause the most difficulty. Regarding human perception of foreign accented speech, Levi et al. (2007) distinguish speaker-dependent factors such as age of learning and length of residence, from speaker-independent factors, such as lexical frequency and listening context. They found that lower frequency words were rated as more accented when listeners were presented with audio alone; this effect was lessened when listeners were presented with audio and corresponding text. Incera et al. (2017) found that words in unpredictable sentences were rated as more accented than words in predictable sentences. These findings suggest that rare or less predictable words may be more challenging for listeners.

Palanica et al. (2019) performed an experiment with three voice assistants (Alexa, Siri and Google Assistant) on brand and generic names for the top 50 most dispensed medications in the United States. They observed two of the three assistants had lower comprehension for those with a foreign accent. This indicates that (perhaps somewhat unusual) proper names may present difficulties for ASR when confronting foreign accents.

Speakers of foreign languages may also use their native pronunciations when uttering proper names in English, especially for names of people and places in their native land. These foreign/native pronunciations may be confusing for both humans[1] and ASR, especially when they are not familiar with the foreign language.

2.2 How can we help ASR to deal with these words?

Hinsvark et al. (2021) survey a range of techniques to improve ASR performance in the face of accentual variety in general. When we focus on improving ASR performance on infrequent words and names, we can consider a technique known as contextual recognition or custom vocabulary (Drexler-Fox & Delworth, 2022). ASR users can present the system with lists of words pertinent to their business that are likely to recur in the files they submit for transcription. Kang & Zhou (2020) mention several possible types, including contact names, locations, song playlists and topic-specific or trending terms. We apply this technique below to assess its effectiveness with respect to accent variation.

Baese-Berk et al. (2020) describe both a matched and unmatched interlanguage intelligibility benefit. In the matched condition, speakers of a foreign language are better able to transcribe English of other speakers of that language than English native speakers. In the unmatched condition, they note that this benefit extends across languages, such as a Korean speaker transcribing a Mandarin speaker's English.

Although not necessarily developed due to these sorts of observations, the multilingual speech recognition technique (Zhou et al. 2022) can conceivably provide a similar benefit. This method entails pooling training data from multiple languages. The output of a speech recognition system thus trained can be weighted to produce a single language, or some combination of the languages with which it has been trained. The latter option makes it possible for the system to produce translanguaged/code-switched output, e.g. Spanglish. Our hypothesis is that a multilingual system exposed to the native languages of speakers of foreign-accented English will be able to perform better when confronted with those foreign accents.

3. Corpora focusing on accents and bias

Corpora or data sets used for speech recognition may be used for training, validation (also known as development) and testing (also known as evaluation). We will concentrate here on corpora for testing or benchmarking. Test sets can focus on particular categories, such as accents, for which the developer wishes to verify performance. Two important questions arise with respect to such corpora: what are the relevant accent categories and how are they determined? The number of variations of a language is as numerous as the number of its speakers;

[1] https://training.npr.org/2019/04/30/pronounce-like-a-polyglot-saying-foreign-names-on-air/

nonetheless, dialectologists and later sociolinguists have been attempting for generations to describe a smaller number of language varieties based on geographical and social features.

Methods of applying accent labels include self-description (subject to individual variation as well as limiting data to that for which the speakers can be consulted), expert labeling (slow, expensive and subject to interrater reliability issues), geographical inference (to be explained below), and machine labeling (subject to availability of accent identification systems with the needed categories). Test sets can also differ in the extent to which their content matches the application areas of the speech technologies that use them. For example, Common Voice (Ardila et al., 2020), is exclusively read speech, which seems like an unusual application for speech recognition—reading implies a transcript is already available, so why the need to generate another one (except perhaps for proofreading)?

An important feature of Common Voice's accent labels is that they are self-descriptions by volunteer speakers. Reid & Williams (2023) reduced 14,822 unique accent entries for English to 164 (only 16 of which came from a dropdown) after applying normalizing heuristics. The Artie Bias dataset (Meyer et al., 2020), is a manually verified subset of English Common Voice containing 1712 audio clips with 17 speaker-supplied geographical accent tags. The dataset is dominated by American, Indian, and British English; the paucity of Chinese, Middle Eastern, European, South and Central American, African, and Australian accents makes it difficult to use as a global benchmark. Additionally, the test set only uses country labels and thus lacks finer regional dialect classifications.

Casual Conversations, now in its second version (Porgali et al., 2023), contains 26,467 videos of 5567 paid participants recorded in Brazil, India, Indonesia, Mexico, Vietnam, Philippines and USA, which was the only location in the first version (Hazirbas et al., 2021). They focus on participant consent, and contain self-reported age, gender, language/dialect, disability status, physical adornments, physical attributes and geo-location. Trained annotators labeled apparent skin tone using the Fitzpatrick Skin Type and Monk Skin Tone scales and voice timbre. Hazirbas et al. (2022) note that race and ethnicity are problematic descriptors and for that reason they opted for skin tone. English data in the corpus comes from speakers of Australian, Canadian, Indian, Philippine, British and American English, as well as speakers of Indonesian, Portuguese and Spanish. The content of the recordings is both scripted text (a reading from a passage by Dostoevsky) and non-scripted text, consisting of monologues about banal topics like how people spend their weekends and the weather.

An alternative to self-labeling of language variety information is what we call geographical inference. The source for the geographical information can be geo-location; for example, Jones (2015) examined regional variations of African American Vernacular English using geotagged tweets and Trinh et al. (2022) used customer device latitude and longitude. O’Neill et al. (2021) used the headquarters country of companies as the label for their speakers (unfortunately these are not present in the downloadable data). While convenient, geographical inference may be misleading for speakers who do not have a linguistic association with the place their location or company’s headquarters would suggest.

4. A case study in earnings calls

In our view, to be useful, accents test suites need to reflect actual speech recognition applications for results to be predictive of future performance in those applications. Since speech recognition systems are utilized in a wide range of domains ranging from finance to media, it will be good for all of these domains to eventually have accent test suites. To this end, we will describe below a test suite in the financial domain consisting of earnings calls— periodic conference calls involving management, analysts, investors, and media to discuss public company financial results. The advantages of this over the existing test suites described above are at least threefold, reflecting the majority of audio submitted to speech recognition by real customers:

- The speech is spontaneous, as opposed to scripted. Spontaneous speech is more challenging for both ASR (Gabler et al., 2023) and human learners (Wagner et al., 2021), so testing on it is crucial in order to obtain realistic performance estimates.
- Each call has multiple speakers rather than one, introducing complexities presented by dialog and different speakers. This is also more reflective of real data; in 20,000 hours of Rev.com audio data, we found that over 90% of 3-minute segments had more than one speaker. Chang et al. (2019) discuss ASR challenges and Arons (1992) summarizes the human challenges of multi-speaker over single-speaker audio.
- The speech belongs to a domain (financial) for which customers actually request transcriptions. In-domain performance is more predictive of future ASR results than out-of-domain performance (Gandhi et al., 2022).

Company earnings calls address each of these desiderata, and, they provide a great deal of speech variety from having many different speakers, differing accents, and complex domain terminology. Although they contain a concentration of financial jargon, earnings calls still provide broad coverage of real-world topics. In the following sections, we present a compiled dataset of 119 hours that covers 7 regional accent groupings and is freely available to the public. We perform WER (word error rate, lower is better) analysis using several industry ASR models and compare their performance across accent regions and word characteristics. In Section 0, we describe the data properties and collection methodology. In Section 0, we provide an initial analysis of accuracy disparities between regional accents, and in Section 0 we apply the accent mitigation strategies mentioned above on Earnings-22 and assess results. Finally, we conclude with a call to action to promote improved accent bias benchmarking in the ASR field.

5. The Earnings-22 dataset

The Earnings-22 benchmark dataset[2] is developed with the intention of providing real-world audio focused on identifying bias in ASR systems. To that end we aggregated public[3] earnings calls exhibiting a range of World Englishes from global companies. We collected a total of 125 calls, totaling 119 hours downloaded from various sources[4]. The earnings calls in the Earnings-22 corpus are sourced from a total of 27 countries which we categorize into 7 linguistic regions defined in Table 1 (see Section 0 for explanation on how these regions were defined).

To produce a broad range of speakers and accents, we focused our efforts on finding earnings calls from global companies. The process of properly labeling speaker accents is a difficult task that requires language experts to rate accents and techniques to deal with any disagreements, in addition to implicit bias we add as a result of the rating. We opted instead to follow the geographical inference method used in O’Neill et al. (2021) and associate an earnings call with the country where the company was headquartered.

As mentioned previously, we recognize that in a given call there will be speakers who are linguistically distinct from the company’s headquarters. Our effort to partially reduce this possible bias was to attempt a minimum of 5 calls from as many countries as were available to us; in this way, we hoped that we could normalize the “global voices” in every country we included while highlighting the linguistic variances unique to the country’s earnings calls.

One exception to this were calls from Ghana and Nigeria, which we actively sought out to increase the coverage of African accents in this dataset. Despite best efforts for these accents, we were only able to find 1 Nigerian and 4 Ghanaian earnings calls. Although these countries were the least represented in the dataset, their inclusion was crucial to improve the overall analytical value of the corpus.

[2] This benchmark is available on Github at https://github.com/revdotcom/speechdatasets/tree/master/earnings22

[3] Earnings calls fair use legal precedent in Swatch Group Management Services Ltd. v. Bloomberg L.P. https://www.copyright.gov/fair-use/summaries/swatchgrp-bloomberg-2dcir2014.pdf

[4] Most calls are from https://seekingalpha.com/. A few come directly from the company websites: https://www.mtn.com.gh/ and https://transcorpnigeria.com/

Table 1: Earnings-22 language regions

Language Region	**Countries**	**Time (in Hours)**	**Number of Files**
African	Nigeria, Ghana	5.06	5
Asian	Indonesia, Turkey, India, Japan, South Korea, China	25.27	28
English (inner circle)	United Kingdom, Canada, Australia, United States, South Africa	22.85	26
Germanic	Denmark, Sweden, Germany	13.53	12
Other Romance	France, Italy, Greece	15.61	13
Slavic	Russia, Poland	7.72	10
Spanish / Portuguese	Argentina, Brazil, Chile, Spain, Colombia, Portugal	28.87	31

5.1 Creating and preparing the transcripts

To ensure high quality transcripts, we submitted our files to the Rev.com human transcription platform. Once completed, the quality of each transcript is also verified by a separate group of graders[5]. Following our work in Del Rio et al. (2021), we chose to produce verbatim[6] transcriptions to best characterize real human speech. We processed the transcripts produced by Rev.com and removed atmospheric information[7] using our internal processing tools. These files are then converted into our .nlp file format[8] to allow for WER scoring using *fstalign* (see below).

[5] See https://www.rev.com/blog/revver-tips/revver-levels-freelance-transcriptionist for description of different transcriptionist levels

[6] https://www.rev.com/ blog/resources/verbatim-transcription

[7] Examples include information about background music, coughs, and other non-speech noises

[8] https://github.com/revdotcom/fstalign/blob/develop/docs/NLP-Format.md

5.2 Defining regions

Due to the large number of countries, we defined linguistic regions to make bias analysis more practical (we show some statistics for those regions in Table 1). Our primary divisions are inspired by Kachru's (1992) circles of English. The inner circle includes the English homeland (UK) and the first wave of colonization (US, Canada, Australia, New Zealand, South Africa). The outer circle includes the later wave of colonization including India and Africa. The expanding circle is everywhere else.

Our region grouping includes a mixture of language family information and geographical location. Due to the overwhelmingly large number of Spanish-speaking countries in the dataset compared to other Romance languages, we felt the separation of Spanish / Portuguese and Other Romance was helpful in order to get a better view on the accents. Portuguese was grouped with Spanish because of its linguistic similarity as both are a part of the Ibero-Romance group of Romance languages (Pharies, 2006).

Previous work has found that despite a more distant genetic relationship between Greek and Romance languages, their proximity and long-term contact has resulted in significant association both lexically and phonologically (Ralli, 2020) — as a result, we felt they best fit together in the Other Romance region. We chose to split South African earnings calls from the African region to follow the distinctions made in previous works (e.g. Wells, 1982) that include South African English in the inner circle in contrast with other countries in the African region. Of course, South African English in truth exhibits wide variation across the peoples of that multiethnic society (van Rooy, 2020), so like all of our grouping attempts, it can be subject to revision and refinement.

6. WER analysis on Earnings-22

To fully showcase the dataset and its characteristics, we provide an initial benchmark along the accent dimension. We utilized our previously released opensource-toolkit, *fstalign*[9], in order to calculate WER, since it provides the most holistic and fair comparison for speech recognition models that may output different, but equivalent, transcriptions (Del Rio et al., 2021). For example, alternate but equivalent transcriptions of “2023” are “twenty twenty-three” or “two thousand twenty-three”, as both are possible ways of saying that year. Equivalent transcriptions extend to vernacular speech which can be “standardized” by different recognition models; saying “gonna” can result in a direct transcription of “gonna” as well as “going to”. By using *fstalign*, we allow

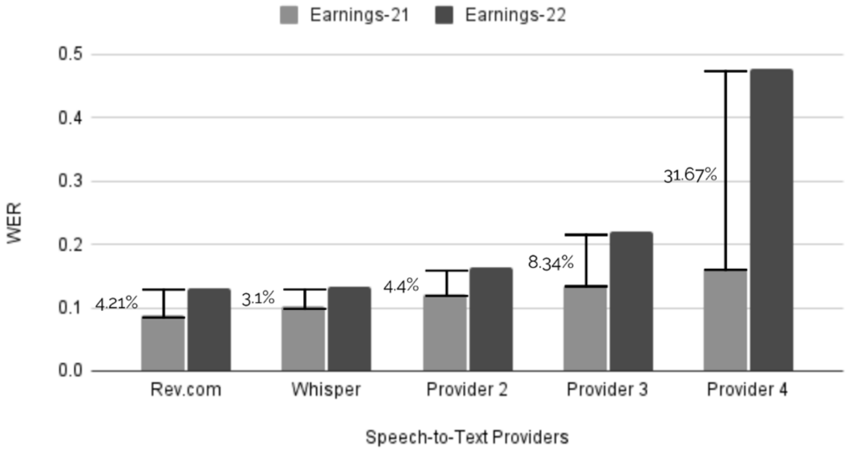
[9] https://github.com/revdotcom/fstalign

for these transformations at evaluation time and reduce bias against providers that do not conform to Rev's verbatim transcription rules.

6.1 Provider and regional file breakdown

We used four cloud-based ASR providers to submit our evaluation audio and obtain hypothesis transcripts. We evaluated our V2[10] speech recognition engine currently available to customers, Open AI's open source Whisper Large model (Radford et al., 2022), and three industrial engines hosted by other speech recognition providers. As this work is not meant to impugn competitor business but rather to highlight an important problem that must be addressed, we have chosen to obfuscate the names of all industrial engines except the one we service. To highlight the discrepancy in performance between "inner circle English" and "outer/expanded circle English" we compare all speech engines on the Earnings-21 (Del Rio et al., 2021) corpus containing largely inner circle English region countries, to the new Earnings-22 corpus with the broader accent regions in Figure 1. For each provider, we show the difference in performance in WER between Earnings-21 and Earnings-22.

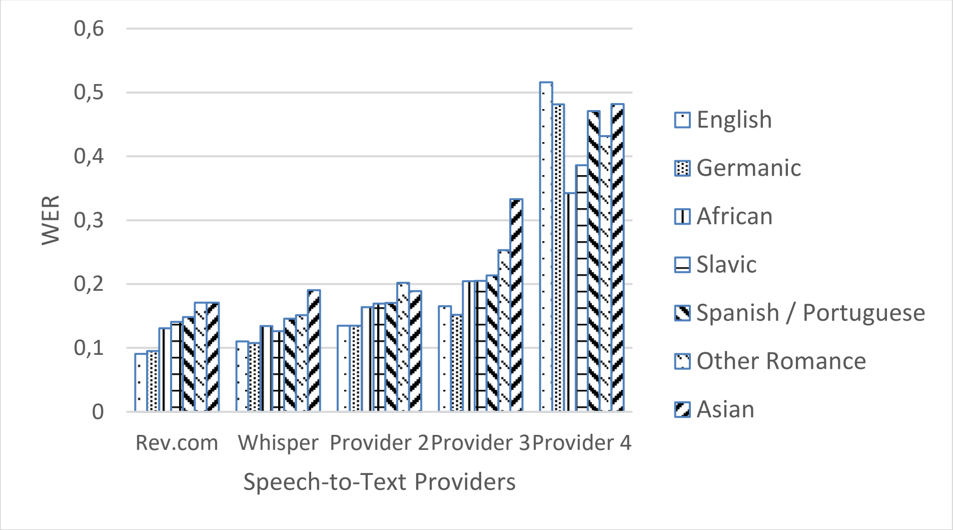
Figure 1: WER of each evaluated speech-engine on both Earnings-21 and Earnings-22.



It is immediately evident that every provider has a considerable gap in performance caused by "accented English". Most notably, the performance of Provider 4 is so degraded that accented speech would result in almost *every other word* being incorrectly transcribed.

[10] https://www.rev.com/blog/product-features/rev-improves-accuracy-by-over-25-with-launch-of-new-v2-asr-model

Figure 2: WER on Earnings-22 for each speech engine and conditioned by language region.



Looking at the discrepancy in more detail, Figure 2 shows the performance of each engine by conditioning the output by different language regions defined in the Earnings-22 corpus. With the exception of Provider 4, the performance on the inner circle English region for each speech engine is approximately equivalent to its performance on the Earnings-21 corpus. The accent regions are presented in an order generally reflecting the ASR difficulty for the remaining providers. Comparing over all the speech engines reveals:

- English and Germanic regions perform best with ASR relative to the other language regions
- Asian and Romance Languages other than Spanish / Portuguese perform the worst relative to the other language regions.

For English and Germanic, this distinction makes sense, as English is a Germanic language, and most modern ASR models are trained on inner-circle English data. We speculate that the excellent performance on Germanic may be due to highly skilled second-language speakers who may articulate more clearly and speak more slowly than first language speakers, a result also noted in Miller et al. (2021). The results on other regions echo the linguistic distance results of Chiswick et al. (2004). For example, the Asian region has the biggest WER gap from the English region.

6.2 Regional word-level breakdown

With the ASR transcripts of all providers, we want to understand what words were the most difficult to capture. Our first exploration aggregated the results over all providers and filtered for words that occurred at least 5 times within a transcript and were most commonly missed by all providers. To measure how often a word is mistaken, we look at a word's F1-score (the harmonic mean of Precision and Recall and also referred to as F1) to get one number that measures not only how well a word is captured when it occurs within the data (Recall) but also how accurately the providers predict the same word (Precision); for this analysis, we considered words with an F1 less than 0.3 as "commonly missed by all providers". We found that some domain-specific finance words were commonly incorrect, as well as a few contractions, whose processing often causes trouble for ASR systems (Goldwater et al., 2010). The results of this analysis are shown in Table 2.

Table 2: Top 10 word-level errors across all files.
Format of cell is: **word-category** word (number of files)
Categories: Contractions **(')** | Financial **($)**

$ Capex (27)	an (9)
' we'll (17)	too (8)
yeah (16)	**$** sales (8)
$ EBITDA (12)	**'** there's (8)
' it's (10)	**$** CD (8)

We next dug deeper into commonly mistaken words on a regional level by repeating the previous experiment but removing words that were shared in every region. By doing so, we hoped to identify words that ASR systems struggled with that were particular to each region. The results of this analysis are shown in Table 3. Across the regions, we see that errors involve both common words as well as names, abbreviations, or acronyms. Thus we see that the degradation of WER seen across region groups is not simply due to more industry-specific jargon or terminology, but also poor recognition around common words spoken in regional accents.

Table 3: Word-level errors by region.
Format of cell is: **word-category** word (number of files)
Categories: Contractions **(‘)** | Names **(☺)** | Abbreviations/Acronyms **(ABC)** |
Disfluencies **(…)** | Financial **($)**

African	**Asian**	**English**	**Germanic**	**Other Romance**	**Slavic**	**Spanish / Portuguese**
☺ momo (6)	an (7)	‘ we’ll (7)	‘ there’s (4)	**$** euro (5)	☺ cd (8)	**…** eh (9)
affect (4)	**$** quarters (6)	**$** stock (5)	☺ reinhard (3)	engineers (3)	☺ projekt (5)	yeah (8)
ABC gh (2)	part (4)	**$** rand (4)	federal (3)	grew (3)	☺ cyberpunk (3)	☺ alejandro (6)
‘ won’t (2)	there (4)	yeah (4)	☺ lars (3)	☺ carlo (3)	red (3)	☺ chile (6)
better (2)	**ABC** tl (4)	☺ gus (3)	☺ aker (3)	renovation (3)	**$** rubles (3)	**‘** it’s (6)
‘ we’ll (1)	**$** won (4)	‘ we’re (3)	lead (3)	**ABC** d (3)	continued (3)	too (5)
ABC mtn (1)	**ABC** mr (4)	day (3)	☺ roland (3)	‘ we’ll (3)	**…** hmm (2)	had (5)
their (1)	**$** business (3)	**$** breakeven (3)	☺ coke (3)	**$** asset (3)	higher (2)	**$** euros (5)
$ merchant (1)	**$** monetization (3)	☺ allen (3)	**$** billing (3)	☺ tarek (3)	**‘** they’re (2)	slide (5)
… oh (1)	**$** provisions (3)	☺ billie (3)	sea (3)	☺ publicis (2)	**ABC** gen (2)	these (5)

7. Application of accent mitigations to Earnings-22

We seek to apply two of the methods we discussed above, custom vocabulary and multilingual modeling, in order to see whether they can be employed to improve the Rev.com results on Earnings-22.

7.1 Applying Custom Vocabulary to Earnings-22

As described above, Custom Vocabulary (CV) is a technique whereby lists of likely or in-domain words can be presented at recognition time to boost the likelihood that such words will be correctly recognized. These words could come from ancillary materials; for example, a list of names of employees of a company could be presented when automatically transcribing that company’s meetings. Or they could come from prior transcripts from the company or domain that have been validated. In the latter case, we believe it is helpful to identify a subset of words that distinguish audio in this domain from others. To derive such a list, Drexler-Fox and Delworth (2022) used the spaCy named entity tagger[11].

We ran this entity tagger on each Earnings-22 file's transcript to generate a custom vocabulary list for that file. We included spans of words the tagger predicted were organization or person names as well as other proper nouns. The list is also filtered to remove duplicates and stopwords[12]. Of course, "in real life", it would be strange to have a validated transcript prior to using ASR to generate another one. Therefore, the results we show below are a kind of "oracle" best-case scenario for the use of this technology. We distinguished the following different conditions:

- English vs. Spanglish (see below) speech recognition model
- All Earnings-22 vs. English (inner circle) partition vs. Spanish/Portuguese partition
- With and without spaCy-derived CV
- In-vocabulary (IV) vs. Out-of-vocabulary (OOV): For CV words, are they in the model's lexicon or not?

For IV and OOV, we provide two measures, WER and F1. F1 is used to characterize the precision (P) and recall (R) of IV and OOV terms (higher is better). As we can see in Table 4, the English speech recognition model outperforms the Spanglish speech recognition model in all cases except Spanish/Portuguese partition OOV. Indeed, the OOVs in the Spanish/Portuguese partition contain many words with Spanish origins (e.g. *bicentenario, cartagena's, cementos, corredores, cuadrado, energia's, inteligencia*). The facility with which the Spanglish model can interpret Spanish pronunciations likely improves recognition of such words. In contrast, the OOVs in the English (inner circle) partition are often compounds or words from other languages (e.g. *bamberg, betgenius, blostein, breithaupt)*. We find that the English model outperforms Spanglish on these non-Spanish kinds of words, just as it does on the English general vocabulary.

[11] https://spacy.io/api/entityrecognizer

[12] https://gist.githubusercontent.com/sebleier/554280/raw/7e0e4a1ce04c2bb7bd41089c9821dbcf6d0c786c/NLTK's%2520list%2520of%2520english%2520stopwords + "wear, okay, head, may, yes, sorry, year, hmm, mm, sir, kinda, must".

Table 4: Custom Vocabulary Statistics by Data Partition

Partition	**Measure**	**English acoustic model**	**English acoustic model + CV**	**Spanglish acoustic model**	**Spanglish acoustic model + CV**
All	WER	12.09	11.98	13.08	12.99
	IV F1	90	90.8	89.5	90.3
	IV WER	19.2	17.8	20.2	18.8
	OOV F1	0	53.8	3.3	48.3
	OOV WER	100	91.3	98.6	85.9
English (inner circle)	WER	8.72	8.92	8.99	9.15
	IV F1	92.3	93	92	92.8
	IV WER	14.8	13.5	15.3	13.9
	OOV F1	0	69.6	0.6	56.4
	OOV WER	100	64.2	99.7	73.6
Spanish/Portuguese	WER	12.57	12.61	13.91	13.97
	IV F1	89.4	90.2	89	89.7
	IV WER	20.5	19	21.1	20
	OOV F1	0	45	10.1	49
	OOV WER	100	110.7	95.2	81.9

7.2 Applying multilingual speech recognition to Earnings-22

We trained a multilingual model with English, French and Spanish layers in the manner of Zhou et al. (2022). Whereas a "monolingual" mode of operation would be to weight the English, French or Spanish layer 1 and the other layers 0, we chose to create a "Spanglish" model where English was weighted 0.5 and Spanish was weighted 0.5 (and French 0). The Spanglish model is capable of transcribing either English or Spanish, but not quite as well as monolingual models. However, it is also capable of transcribing mixed English and Spanish data, such as is found in English-Spanish code-switching or translanguaging. While the Spanish/Portuguese partition of Earnings-22 is English, it does feature Spanish native speakers speaking accented English with a fair amount of Spanish/Portuguese-origin words, due to the nature of their businesses. We sought to explore this model's performance in comparison to our monolingual (non-multilingual) English speech recognition model.

While the Spanglish model does not yet perform as well as the English model (trained on the same English data included in the Spanglish model) on general data, we saw above how it appears to have an advantage in the CV OOV case in the Spanish/Portuguese data partition; a set of words that appears to be heavy in Spanish origin words, which Spanish native speakers likely pronounce in a Spanish manner. We sought to compare the performance of two CV words, *Colombia* and *EBITDA* (generally pronounced [ˈɛbɪɾə] in American English, an acronym for "earnings before interest, taxes, depreciation and amortization"), in order to look more closely at the performance of the two acoustic models on one Spanish-origin word and one non-Spanish origin word.

For most native speakers of American English, *Colombia* (country) is homophonous with *Columbia* (university, District of Columbia): [kəˈlʌmbiə]. However, for speakers of Spanish, *Colombia* may often be pronounced [koˈlombia]; the most perceptually salient difference being the stressed vowel [o] rather than [ʌ]. The word *Colombia* only appears in the Spanish/Portuguese partition of Earnings-22 and the word *Columbia* does not appear at all, so we only analyze the performance of this word in the Spanish/Portuguese partition, where in Table 5 we compare the same four conditions as above: English with and without CV, Spanglish with and without CV. The word *Colombia* appears in the CV list—both *Columbia* and *Colombia* are in the training vocabulary for both the English and Spanglish models.

In the English conditions, CV reduces false positives (FP) for *Columbia* and increases correct accepts (CA) for *Colombia*. The Spanglish condition outperforms the English condition and achieves the same F1 as the English CV condition. The Spanglish CV condition is best of all, with the lowest FP *Columbia*, the highest CA for *Colombia*, along with the highest F1 for *Colombia*. As above, we presume the better ability of the Spanglish model to correctly interpret native Spanish pronunciations of *Colombia* accounts for its superiority in this respect.

Table 5: *Colombia* and *Columbia*

Condition	word	FP	CA	P	R	F1
English	colombia	16	79	0.83	0.52	0.64
	columbia	90	0	0	n/a	0
English CV	colombia	6	101	0.94	0.66	0.78
	columbia	28	0	0	n/a	0
Spanglish	colombia	36	121	0.77	0.79	0.78
	columbia	36	0	0	n/a	0
Spanglish CV	colombia	47	136	0.74	0.89	0.81
	columbia	23	0	0	n/a	0

We next look at the word *EBITDA*, which does not have a Spanish origin. Our hypothesis is that the Spanglish model will not offer a performance gain over English for this word. In this case, in Table 6 we compare both English inner circle and Spanish/Portuguese partitions, since the word appears in both. *EBITDA* does appear in the CV list; however, there is no obvious homophonous word with which to compare it as in the *Colombia*/*Columbia* case. The English partition outperforms the Spanish/Portuguese partition in all conditions, indicating that inner circle English pronunciations are easier for the recognizer. CV improves performance on the English model for both partitions. The English model outperforms the Spanglish model for both partitions. CV improves Spanglish performance on both partitions. Interestingly, the Spanglish CV condition outperforms the English CV condition for the English partition, but not the Spanish partition.

Table 6: *EBITDA*

Model	Partition	FP	CA	P	R	F1
English	English	1	31	0.91	0.74	0.82
	Spanish/Portuguese	8	40	0.83	0.22	0.35
English CV	English	2	35	0.9	0.9	0.9
	Spanish/Portuguese	11	62	0.85	0.32	0.47
Spanglish	English	1	28	0.93	0.68	0.79
	Spanish/Portuguese	4	30	0.88	0.16	0.27
Spanglish CV	English	2	35	0.92	0.9	0.91
	Spanish/Portuguese	7	49	0.88	0.24	0.38

8. Conclusion

Accent-robust ASR is increasingly important in practical applications. Significant challenges such as data sparsity and lack of a standard benchmark impede research progress, and many open questions exist. We release the Earnings-22 corpus to begin to address the lack of standard benchmarking. Using Earnings-22 we have shown that, despite major WER improvements in ASR in general, gaps in performance on some language varieties indicate there is more work to be done. Through the use of both custom vocabulary and multilingual models, we have shown possibilities for improving performance on some accented terms. While custom vocabulary appears to offer promise regardless of language variety, the multilingual approach seems to offer particular promise to those varieties that are influenced by other languages. We hope our experiments lead to further exploration of multilingual ASR research as a way to build more inclusive models, building on the encouraging work of Wrembel et al. (2020, p. 2) in the human sphere: “Multilingual learners, thus, have at their disposal a broadened phonetic repertoire, a raised level of metalinguistic awareness, and potentially enhanced perceptual sensitivity, which may facilitate the learning of a subsequent phonological system”. With the release of this new corpus, we hope to motivate researchers to work on the problem of real-world accented audio. We challenge all industry and academic leaders to find new techniques to improve model recognition on all voices to create more equitable and fair speech technologies.

9. Acknowledgments

We would like to thank the transcriptionists without whose help and hundreds of hours of work, none of this dataset release would be possible.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G.. (2020). Common Voice: A massively-multilingual speech corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation*, pp. 4218-4222.

Arons, B. (1992). A Review of the Cocktail Party Effect. *AVIOS*.

Baese-Berk, M. M., McLaughlin, D. J. & McGowan, K. B. (2020). Perception of non-native speech. *Language and Linguistics Compass*, pp. 1-20. https://doi.org/10.1111/lnc3.12375

Chang, X., Qian, Y., Yu, K. & Watanabe, S. (2019). End-To-End Monaural Multi-Speaker ASR System Without Pretraining. *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP.2019.8682822

Chiswick, B. R. and Miller P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development,* vol. 26, no. 1, pp. 1–11. https://doi.org/10.1080/14790710508668395

Del Río, M., Delworth, N., Westerman, R., Huang, M., Bhandari, N., Palakapilly, J., McNamara, Q., Dong, J., Zelasko, Z., and Jetté, M. (2021). “Earnings-21: A Practical Benchmark for ASR in the Wild,” in *Proc. Interspeech 2021*, pp. 3465–3469. https://doi.org/10.21437/Interspeech.2021-1915

Drexler-Fox, J. & Delworth, N. (2022). Improving contextual recognition of rare words with an alternate spelling prediction model. *Proceedings of Interspeech*.

Gabler, P., Geiger, B. C., Schuppler, B. & Kern, R. (2023). Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition. *Information*, 14, 137. https://doi.org/10.3390/info14020137

Gandhi, S., Von Platen, P., & Rush, A. M. (2022). ESB: A Benchmark for Multi-Domain End-to-End Speech Recognition. *arXiv preprint arXiv:2210.13352*.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). “Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates,” *Speech Communication*, vol. 52, no. 3, pp. 181–200. https://doi.org/10.1016/j.specom.2009.10.001

Good, P. I. (2004). Permutation, Parametric, and Bootstrap Tests of Hypotheses. *Springer Series in Statistics*. Springer-Verlag.

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A. & Ferrer, C. C. (2021). Towards measuring fairness in AI: the Casual Conversations dataset. *ArXiv*.

Hazirbas, C., Bang, Y., Yu, T., Assar, P., Porgali, B., Albiero, V., Hermanek, S., Pan, J., McReynolds, E., Bogen, M., Fung, P. & Ferrer, C. C. (2022). Casual Conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness. https://doi.org/10.1109/TBIOM.2021.3132237

Hinsvark, A. J., Delworth, N., Del Río, M., McNamara, Q., Dong, J., Westerman, R., Huang, M., Palakapilly, J., Drexler, J., Pirkin, I., Bhandari, N. & Jetté, M. (2021). Accented Speech Recognition: A Survey. *ArXiv*.

Holmes, J. (2013). *An introduction to sociolinguistics*. Routledge. https://doi.org/10.4324/9781315833057

Incera, S., Shah, A. P., McLennan, C. T. & Wetzel, M. T. (2017). Sentence context influences the subjective perception of foreign accents. *Acta Psychologica* 172, pp. 71-76.

Jones, T. (2015). Toward a description of African American Vernacular English dialect regions using “Black Twitter”. *American Speech*, Vol. 90, No. 4. https://doi.org/10.1215/00031283-3442117

Kachru, B. (1992). *The Other Tongue: English across cultures*. University of Illinois Press.

Kang, Y. M. & Zhou, Y. (2020). Fast and robust unsupervised contextual biasing for speech recognition. *ArXiv*.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z , Toups, C., Rickford, J. R., Jurafsky, D. & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kosmala, L., and Crible, L. (2021). The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. *Language and Speech*, May 2021. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-03225622 https://doi.org/10.1177/00238309211010862

Levi, S. V., Winters, S. J. & Pisoni, D. B. (2007). Speaker-independent factors affecting the perception of foreign accent in a second language. *Journal of the Acoustic Society of America*, 121(4), pp. 2327-2338. https://doi.org/10.1121/1.2537345

Lippi-Green, R. (2012). English with an Accent: Language, Ideology and Discrimination in the United States. Routledge. https://doi.org/10.4324/9780203348802

Meyer, J., Rauchenstein, L., Eisenberg, J. D. & Howell, N. (2020). Artie bias corpus: An open dataset for detecting demographic bias in speech applications. *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6462–6468.

Miller, C., Tzoukermann E., Doyon J., and Mallard, E., (2021). Corpus creation and evaluation for speech-to-text and speech translation. *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pp. 44–53.

O’Neill, P. K., Lavrukhin, V., Majumdar, S., Noroozi, V., Zhang, Y., Kuchaiev, O., Balam, J., Dovzhenko, Y., Freyberg, K., Shulman, M. D., Ginsburg, B., Watanabe, S., and Kucsko, G. (2021). “SPGISpeech: 5,000 Hours of Transcribed Financial Audio for Fully Formatted End-to-End Speech Recognition,” in *Proc. Interspeech*, pp. 1434–1438.

Palanica, A., Thommandram, A., Lee, A., Li, M. & Fossat, Y. (2019). Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. *NPJ Digital Medicine*, vol. 55, pp. 1-6. https://doi.org/10.1038/s41746-019-0133-x

Pharies, D. A. (2007). *A Brief History of the Spanish Language*. University Of Chicago Press. https://doi.org/10.1038/s41746-019-0133-x

Porgali, B., Albiero, V., Ryda, J., Ferrer, C. C. & Hazirbas, C. (2023). The Casual Conversations v2 Dataset. *ArXiv*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Ralli, A. (2020). Greek in Contact with Romance. In M. Loporcaro & F. Gardani (eds.) *The Oxford Encyclopedia of Romance Linguistics*. Oxford. https://doi.org/10.1093/acrefore/9780199384655.013.422

Reid, K. & Williams, E. T. (2023). Common Voice and accent choice: Data contributors self-describe their spoken accents in diverse ways. *EasyChair*. https://doi.org/10.1145/3617694.3623258

Trinh, V. A., Gharemani, P., King, B., Droppo, J., Stolcke, A. & Maas, R. (2022). Reducing geographic disparities in automatic speech recognition via elastic weight consolidation. *Proceedings of Interspeech*. https://doi.org/10.21437/Interspeech.2022-11063

van Rooy, B. (2020). English in Africa. In D. Schreier, M. Hundt & E. W. Schneider (eds.), *The Cambridge Handbook of World Englishes*, pp. 210-235. Cambridge University Press.

Wagner, E., Liao, Y.-F. & Wagner, S. (2021). Authenticated Spoken Texts for L2 Listening Tests. *Language Assessment Quarterly* 18:3, pp. 205-227. https://doi.org/10.1080/15434303.2020.1860057

Wells, J. C. (1982). *Accents of English: Volume 3: Beyond the British Isles*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611766

Wrembel, M., Gut, U., Kopečková, R. & Balas, A. Cross-linguistic interactions in third language acquisition: Evidence from multi-feature analysis of speech perception. (2020). *Languages* 5:52, pp. 1-21. https://doi.org/10.3390/languages5040052

Yang, X., Audhkhasi, K., Rosenberg, A., Thomas, S., Ramabhadran, B., and Hasegawa-Johnson, M. (2018). "Joint modeling of accents and acoustics for multi-accent speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018, pp. 1–5. https://doi.org/10.1109/ICASSP.2018.8462557

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D. & Zweig, G. (2016). Achieving human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. https://doi.org/10.1109/TASLP.2017.2756440

Zhou, L., Li, J., Sun, E. & Liu, S. (2022). A Configurable Multilingual Model is all you need to recognize all languages. *Proceedings of ICASSP*. https://doi.org/10.1109/ICASSP43922.2022.9747905