Modeling Gender and Dialect Bias in Automatic Speech Recognition

Camille Harris[1] , Chijioke Mgbahurike[2], Neha Kumar[1], Diyi Yang[2],
[1]Georgia Institute of Technology, [2]Stanford University,

Abstract

Dialect and gender-based biases have become an area of concern in language-dependent AI systems including around automatic speech recognition (ASR) which processes speech audio into text. These potential biases raise concern for discriminatory outcomes with AI systems depending on demographic- particularly gender discrimination against women, and racial discrimination against minorities with ethnic or cultural English dialects.As such we aim to evaluate the performance of ASR systems across different genders and across dialects of English. Concretely, we take a deep dive of the performance of ASR systems on men and women across four US-based English dialects: Standard American English (SAE), African American Vernacular English (AAVE), Chicano English, and Spanglish. To do this, we construct a labeled dataset of 13 hours of podcast audio, transcribed by speakers of the represented dialects. We then evaluate zero-shot performance of different automatic speech recognition models on our dataset, and further finetune models to better understand how fine-tuning can impact performance. Our work fills the gap of investigating possible gender disparities within underrepresented dialects.

1 Introduction

Multiple fields within computing such as Human-Computer Interaction, Natural Language Processing (NLP), Speech Processing, and Algorithmic fairness have identified the ways in which algorithmic systems produce disparate outcomes for minority groups, including on the basis of gender, race, and ethnicity. In language driven applications, such biases have been identified with respect to gender and with respect to dialects associated with minority groups (Sun et al., 2019; Sap et al., 2019). Automatic speech recognition (ASR), the machine learning application in which speech audio is transcribed to text, is among these applications which such biases have become an area for concern and further analysis (Koenecke et al., 2020; Tatman, 2017; Tatman and Kasten, 2017; Wassink et al., 2022).

As a result of such biases in ASR, minority dialect speakers and women may be more likely to struggle with accurate downstream applications using ASR models, such as captioning (Harris et al., 2023) and voice assistants (Cunningham, 2023; Harrington et al., 2022). Mitigating this discrepancy is an important step towards developing equitable technologies that work well regardless of a user's racial, ethic or gender background.

In this work, we specifically focus on four English variants with specific significance within a US context: Spanglish, Chicano English, African American Vernacular English, and Standard American English. Spanglish, also called Engañol (Ardila, 2005), is a language variety that broadly includes any combination English and Spanish features such as grammar structure, innotation and words in real world conversational contexts. As English and Spanish are the first and second most spoken languages in the US respectively this language variety has particular significance, especially as a community among the Latinos within the country (Casielles-Suárez, 2017). Chicano English, also known as Mexican American English, is a specific English dialect originating from the Mexican-American population in the US, most widely spoken in the South-west region of the country. Importantly, while distinct from one-another, there is often overlap between Spanglish and Chicano English, and many speakers may use both, which is reflected in our data. African American Vernacular English (AAVE), also called Ebonics, African American Language, African American English, or Black English, is an English variety that derives from the US Black population, originating from the Southern region of the country. These three languages varieties, which we refer to as minority

Findings of the Association for Computational Linguistics: EACL 2024, pages 15166–15184

dialects throughout this work, are widely spoken among marginalized oppressed populations in the United States. To offer a better understanding of these minority dialects, example sentences and explanation of some of the unique dialect features is offered in Table 1. Finally, we also explore Standard American English (SAE), also called White Mainstream English, which is the dialect of English most commonly spoken by white American English speakers, the dominant racial group in the US. SAE is also the most common dialect used in formal literature and writing in the US, hence it tends to be over-represented compared to other English dialects in many language datasets. We focus specifically on English minority dialects to better understand how ASR bias may or may not reflect wider systems of oppression which further marginalize speakers of these dialects.

We examine the overlap of gender and dialect by building upon prior research in algorithmic fairness that focuses on the intersection of gender and race or ethnicity (Kong, 2022; Wang et al., 2022), examining ethnicity using related dialects (Harris et al., 2022). Examining the overlap of multiple protected attributes can give a richer analysis on the impact of systems of oppression and the severity of potential downstream harms of intelligent systems (Foulds et al., 2020). Prior works have identified serious concerns about how intelligent systems in other domains impact women of color (Buolamwini and Gebru, 2018) finding that in many cases, gender and race based biases compound to most negatively impact women of color (Kong, 2022), but few works focusing on ASR specifically analyze the impact of both race and gender on performance. In this paper, we investigate whether these issues are also applicable to the modality of speech.

We investigate gender and dialect bias by sampling the existing Spotify Podcast Dataset (Clifton et al., 2020), and use a dialect-centered annotation process to transcribe with dialect and gender data. In this dialect-centered annotation process, we recruit minority dialect speakers to produce transcriptions capturing unique spellings, grammar patterns, and words that may not be accurately captured by those with less understanding of the specific English dialect. We then conduct zero shot evaluation of several state of the art automatic speech recognition models on the datasetwith respect to gender and dialect. We find that with respect to dialect, minority dialects consistently perform worse on speech models than Standard American English, while with respect to gender, women typically perform better than men, except in the case of SAE speakers, for which men perform better. Our findings indicate that there is a significant discrepancy for minority dialect speakers. Further, consistent with prior research showing gender discrepancies vary by language (Attanasio et al., 2024; Boito et al., 2022; Gody and Harwath, 2023; Tatman and Kasten, 2017), we find that these discrepancies also differ across English dialects. We also find that finetuning improves overall performance, and we conclude our study with qualitative analysis out the model outputs, giving insight to the areas where ASR models produce errors.

To summarize our contributions are four-fold: (1) We study the gender and dialect bias in ASR from an intersectional perspective. We (2) introduce a dataset annotated for dialect, gender, and other metadata. (3) We conduct extensive experiments to show such disparity with respect to dialect and gender, and (4) we highlight the challenges that ASR systems suffer with respect to these issues.

2 Related Works

Gender and dialect biases have been identified in a variety of NLP applications (Sun et al., 2019) including machine translation (Vanmassenhove et al., 2019), sentiment analysis (Thelwall, 2018), content moderation (Sap et al., 2019) and word embeddings (Zhao et al., 2019). This phenomenon has also been observed with speech models and speech technologies (Ngueajio and Washington, 2022). Here, we outline prior work exploring gender bias, dialect bias, and their implications in real world systems.

2.1 Gender Bias in Speech Systems

Multiple works have explored biases of speech systems with respect to gender with varying results reported. While many systems find worse performance for women (Tatman, 2017), performance varies depending on the language of the speaker (Attanasio et al., 2024; Boito et al., 2022; Gody and Harwath, 2023; Tatman and Kasten, 2017), and in many instances women's speech performs better than men's speech (Fuckner et al., 2023). Boito et al., 2022 explores performance of gender specific wav2vec 2.0 models against models with varying gender balance in training data on downstream ASR and speech translation (ST) tasks in French. They find that gender specific compared to gender balanced models do does not produce a significant

Dialect	Feature	Use in dialect	Text Example
AAVE	Auxiliary verbs	Auxiliary verbs including 'be' 'done' 'been' have distinct use in AAVE.	She been told him she needed the money.
	Copula deletion	The omission of some form of the word 'be'.	She always doing that.
	Negative Concord/ Multiple negation	More than one negative element occurs in a sentence but the sentence only signifies one negation.	I ain't never scared.
Chicano English	Preposition	Prepositions are used before nouns or pronouns, in Chicano English shows substrate influence by using grammar patterns more consistent with Spanish.	We get out of here on June.
	Negative Concord/ Multiple negation	More than one negative element occurs in a sentence but the sentence only signifies one negation.	I didn't see nothing no more.
Spanglish	Intersentenial code-switching	Code switching across sentences between Spanish and English.	His cousin Pedro Pablo sucked his teeth with exaggerated disdain. Esto aqui es un maldito inferno.
	Intrasentenial code-switching	Code switching within sentences between Spanish and English.	These are not gente de calidad.
	Congruent lexicalization	When a sentence contains a shared grammar structure with Spanish and English it can be filled with lexically with elements from either language.	Bueno, in other words, el flight que sale de Chicago around three o clock.

Table 1: Features, uses and text examples of AAVE (Myhill, 1995), Chicano English (Kortmann and Lunkenheimer, 2012) and Spanglish (Casielles-Suárez, 2017). This is a limited list of examples, and does not represent all unique linguistic features of these dialects, nor are these features exclusive to these dialects. In fact we show some features that overlap in multiple of our dialects of study.

performance disparity. Gody and Harwath, 2023 investigates topic diversity, number of speakers, and gender of speakers in the fine-tuning subset for ASR performance on HUBERT. They find minimal impact of gender diversity, but find maximizing number of speakers and topic diversity improves performance. Liu et al., 2022 examines ASR model performance with respect to speaker age, gender, and skin tone. They evaluate multiple training configurations of recurrent neural network transducer (RNN-T) models, finding word error rate across speaker gender and skin tone. Fuckner et al., 2023 explores performance of whisper and wav2vec for Dutch speakers, and finds disparities for non-native, children and elderly speakers. Finally Attanasio et al., 2024 conduct analysis of multilingual ASR models across languages, exploring gaps across gender within languages. They find that models do not perform equally across men and women speakers, performing better for male or female speakers depending on the language and dataset. They also find that phonetic analysis showed no significant differences across gender. Liu et al., 2022 and Fuckner et al., 2023 adds to minimal studies that examine gender and race or ethnicity based characteristics simultaneously. Similarly, our work explores the overlap of minority dialect and gender.

2.2 Racial and Dialect Bias in Speech Systems

In addition to gender bias, several works have explored discrepancies in ASR performance between racial groups and dialects associated with racial groups. Throughout prior works, lower performance for minority groups and dialects used by minority groups is reported in analysis of ASR systems (Koenecke et al., 2020; Tatman and Kasten, 2017; Wassink et al., 2022; Radford et al., 2023). For instance, in an analysis of models of industry ASR models from IBM, Apple, Microsoft, Google, and Amazon across racial groups found all models to have significantly worse performance for Black speakers (Koenecke et al., 2020). Another bias analysis of Client Libraries Oxford captioning system found the highest error rates in Chicanx and African American speakers (Wassink et al., 2022). In the social media context, an evaluation of YouTube captions, analysis of YouTube across ethnic groups found the highest error rates for African Americans (Tatman and Kasten, 2017). Another work from Radford et al. (2023) studied the per-

	Keyword List
Women	women, girls, woman, ladies
Men	men, man, boys, boy, guys, male
Latino	hispanic, hispanic american, boricua, mexican american, latino, latina, lantinx, chicano,chicana, chicanx
Black	african american, black women, black woman, black men, black man, black people

Table 2: Seed keywords used to identify podcasts of different demographic groups.

Group	Duration (mins)	# Speakers
Gender		
Men only	318.0	36
Women only	390.6	35
Men and Women	96.6	21
Dialect		
SAE	623.7	52
AAVE	151.6	10
Chicano	146.8	9
Spanglish	157.8	9
Total	805.2	92

Table 3: Duration and number of speakers for demographic groups.

formance of wav2vec2, whisper, hubert and other models on several datasets, including CORAAL which represents African American speech. This comprehensive study identifies the word error rate of English transcription on several datasets using greedy decoding across model sizes, giving some understanding of how models perform on underrepresented dialects, but doesnt explicitly explore racial disparities. None of these prior studies explore differences within marginalized groups, such as how performance of ASR systems differs for men and women African American English speakers or men and women Spanglish speakers. Further nearly all studies on ASR that include African American English rely on the same dataset, the corpus of regional African American language (CORAAL), and have minimal analysis of Spanglish. We fill these gap in the research by exploring zero-shot performance of state of the art models with our novel dataset, labeled for minority dialect speech and gender.

3 Data and Analytical Methods

3.1 Data Collection

We take an approach of data annotation centered on representing the minority dialects and demographic groups among annotators that are represented in our data. Following is a description of how we collected and annotated data.

We collect data starting with the Spotify podcast dataset (Clifton et al., 2020) which comes with podcast audio and metadata such as podcast title, description, publisher name, etc. We collect data for our specific demographic groups of focus by using demographic related keyword searches (see Table 2 for the list of keywords). We rely on a prior study that used keyword searches to identify keywords (Richard and Kafai, 2016), then build on the keyword lists further. For each keyword within a category, we identify an audio sample as a potential match if it contains that keyword in the podcast title or podcast description. For each dialect group, we then have multiple podcast shows with multiple episodes each that are potential matches. We randomly sample only one episode within each show to prevent one speaker or group of speakers from over-representing any group of interest.

After sampling, annotators document the demographic and speaker metadata of the podcast including number of speakers, gender, dialect use, etc. and transcribe audios with accurate ground truth transcriptions. Annotators determine speaker gender by either relying on how the speakers refer to themselves/refer to one another in the audio, using the podcast description, searching the podcast online or if none of these methods reveal the gender information, infer using their best judgement for the initial annotation, and schedule a follow up conversation with the authors to discuss the annotation further and determine a final annotation. Note that annotator identified gender is a limitation of the work. Podcasts represent a wide range of topics, including finance, children's stories, music, skincare, advice, religion, and lifestyle. To provide further information on the minority dialects of interest in this study, we display examples of features within the dialects that are not present within SAE and example sentences in Table 1.

3.1.1 Dialect-Centered Transcription

We recruit data annotators who have experience speaking and being in community with speakers of non-standard English dialects to annotate our data. Annotators listen to audio and transcribe the audio samples, using automatically generated tran-

scripts from whisper-tiny as a base. Annotators are instructed to pay special attention to properly transcribing words, grammar patterns, and phrases that are unique to dialects of interest. These linguistic differentiation are often the source of automatic speech recognition errors.

Our annotators were three crowd-workers recruited and paid their requested rate through the platform Upwork. As a requirement for the project, annotators had to have experience with speaking one of the three minority dialects, Spanglish, Chicano English, or African American English, with a preference for those that identified as speakers of at least one dialect. Annotator A was a self-identified speaker of African American English, Annotator B was a self identified speaker of both African American English, Spanish, and Spanglish with experience with Chicano English, and Annotator C was an English and Spanish speaker with experience with Spanglish and Chicano English. More details on annotation can be found in the Appendix.

Ultimately this process resulted in 13 hours of audio data across various groups of interest. We report the details about the dataset in Table 3.

3.2 Performance Evaluation

Zero-Shot Performance We evaluate zero-shot model performance across the different demographic groups in our corpus with the following models: wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and Whisper (Radford et al., 2023). Here, we choose zero-shot performance as it is straightforward to assess how well existing speech models perform on our dataset which we evaluate with word error rate (WER). For Wav2vec 2.0 we use *wav2vec2-base-960h*, a 94.4 million parameter model trained on 960 hours of data from the Librispeech dataset, and *wav2vec2-conformer-large* with relative position embeddings, also trained on the same data. For HuBERT we use *hubert-large-ls960-ft* which is a finetuend version of HuBERT large finetuend on the same data. Finally for Whisper we use *whisper-tiny multilingual*, a 39 million parameter model trained on 680k hours of labeled speech data which infers language, *whisper-tiny-en* which has the same training parameters but is trained on English only data, *whisper-base-en* which is 74 million parameters, *whisper-small-en* 224 million parameters, and *whisper-medium-en* which is 769 million parameters.

In addition to measuring performance on our own dialect-centered dataset, we also evaluate these models on samples of other datasets for robustness including CORAAL[1] (Farringon and Kendall, 2024) and voxpopuli [2] (Wang et al., 2021) datasets. CORAAL is a public corpus of AAVE data, compiled with recordings of sociolinguistic interviews conducted with African American interview participants from various regions of the U.S, born between 1888-2005 (Farringon and Kendall, 2024). The sample contains 800 audio samples from CORAAL, with 100 samples each of the 8 location based components of CORAAL, with 404 samples representing men's speech and 396 samples representing women's speech. We analyze zero-shot results across this dataset to understand if similar patterns arise across this data and the AAVE portion of our dataset. Voxpopuli is a dataset of audio recordings from European Parliment event recordings from 2009-2020. We use the test set of the English language portion of the dataset which includes 1842 audio samples, including 511 samples from women speakers and 1331 from men speakers.

Finetuning Performance We use vanilla finetuing to train selected models (Whisper tiny, Hubert large, wav2vev base, and wav2vec2 conformer rope) on 80% of our dataset and evaluate the performance of the finetuned model on held out data balanced for demographic representation to ensure the training data has the same proportion of each demographic group as the overall data. Hyperparameters used to train these models is reported in the Appendix.

4 Results

4.1 Zero-Shot Performance Evaluation

We evaluate zero-shot performance of state of the art models on overall, on gender only, on dialect only, and within gender-dialect combined groups in our corpus, which are described in detail below.

Overall Model Performance The mean Word Error Rate (WER) overall across the three models is displayed in Table 4, Figure 1 depicts these WER results. We see the lowest overall WER with whisper-med which gives 24.8%.

Results on Gender Results for gender performance are shown in Table 4. We exclude podcasts

[1] huggingface.co/datasets/DynamicSuperb/AAVESpeechRecognition_CORAAL

[2] huggingface.co/datasets/facebook/voxpopuli

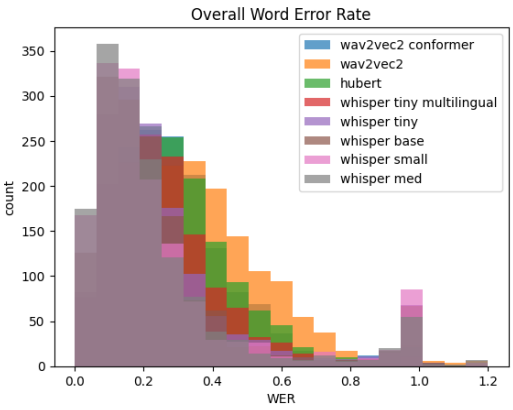

Figure 1: Distribution of Word Error Rate Counts Overall by Model.

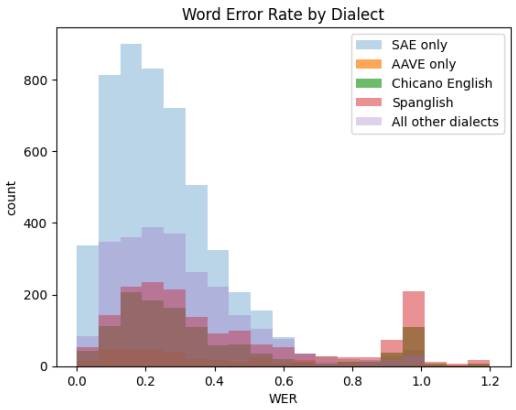

Figure 3: Distribution of Zero-shot WER Result Count by Dialect.

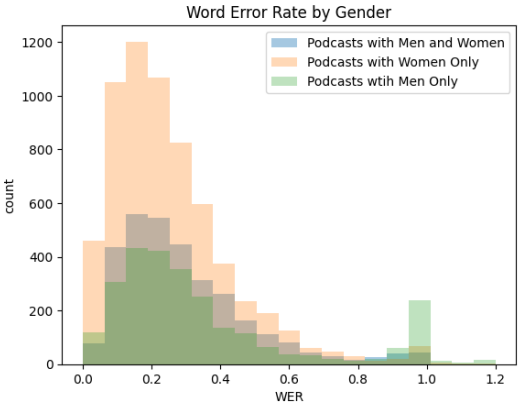

Figure 2: Distribution of Zero-shot WER Result Count by Gender.

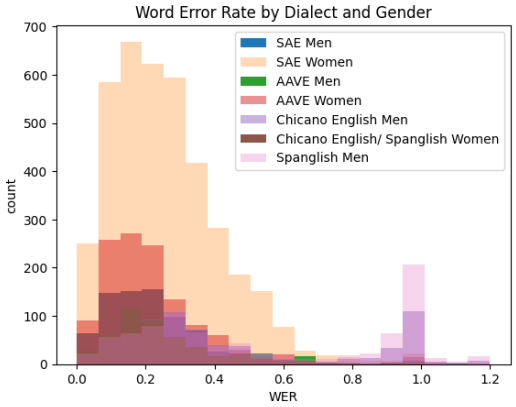

Figure 4: Distribution of Zero-shot WER Result Count by Gender and Dialect.

that have both women and men speakers from this result, using only those with all women speakers or all men speakers. We provide a visualization of these results in Figure 2. Pairwise t-test results with the exception of whisper-tiny-multilingual showed statistically significant difference between men and women (full statistical significance results are reported in the Appendix). We see across all models with the exception of wav2vec2 conformer, women's speech consistently outperform men's speech. Whisper-med yielded the best results on both subgroups with WER of 32.8% and 21.5% for men and women respectively. Among statistically significant results, Wav2vec2 showed the smallest difference in men and women's performance with a difference of 5.1% in WER. We observe the largest error difference for whisper-tiny with a difference of 43.6%.

Results on Dialect Results for dialect performance are shown in Table 5 and displayed in Figure 4. Our results show SAE consistently outperforms all other dialect groups across models. We see the best results across all four dialects with whisper-med. We observe the largest performance discrepancies between SAE and minority dialects with whisper-tiny-en, with WER differences of 57.3%, 77.3% and 16.2% for AAVE, Chicano English, and Spanglish respectively.

Results on Gender-Dialect Combination Results for the intersection of gender-dialect combined categories are shown in Table 6. Statistical significance of the results was evaluated with pairwise t-test and are reported in full in the appendix. Our results show a discrepancy between minority dialects and Standard American English, with men SAE speakers outperforming men minority dialect speakers across all models. Further results frequently show a discrepancy between women and men within dialect groups, with overall results showing lower error rate for women than men within minoirty dialect subgroups. However within SAE, men outperform women in across all models.

Model	Overall	Men	Women
Whisper tiny multilingual	0.523	0.479	0.485
Whisper tiny	0.311	0.680	0.244
Whisper base	0.296	0.562	0.237
Whisper small	0.271	0.481	0.225
Whisper med	0.248	0.328	0.215
Hubert large	0.296	0.385	0.302
Wav2vec2	0.423	0.467	0.416
Wav2vec2 Conformer	0.3412	0.383	0.416

Table 4: Mean word error rate on our full dataset with respect to gender. Pairwise statistical significance test reveals statistically significant results for all pairs of means within gender groups.

Model	AAVE	Chicano English	Spanglish	SAE
Whisper tiny	0.559	0.548	0.488	0.447*
Whisper tiny en	**0.816***	**1.016***	0.405*	0.243*
Whisper base	0.374*	**0.660***	**0.830***	0.232*
Whisper small	0.208*	**0.581**	**0.655**	0.224*
Whisper med	0.224	0.382*	0.422*	0.205*
HuBERT large	0.396*	0.350*	0.461*	0.297*
Wav2vec2	0.473*	0.429*	**0.549***	0.368*
Wav2vec2 Conformer	0.388*	0.342*	0.455*	0.291*

Table 5: Mean word Error Rate on our full dataset with respect to dialect.* Denotes statistically significant results in pairwise t-test when compared with SAE results.

These results suggest that men of color could potentially be more vulnerable to lower speech model performance compared to women of color, contrary to studies in other domains within algorithmic fairness that examine race and gender that typically find worse performance for women of color. Prior works that examine gender bias across multiple languages also find that gender bias may vary to favor men speakers or women speakers depending on the language (Attanasio et al., 2024; Boito et al., 2022; Gody and Harwath, 2023; Tatman and Kasten, 2017). Our results show that the minority dialects of English studied do not follow the same gender trends as SAE, similar to other languages outside of English.

Results on Other Datasets For robustness, we also evaluate how these models perform on other ASR datasets. The results for a evaluation on a sample of the CORAAL dataset which includes men and women African American Language speakers is shown in Table 7. Similar to the patterns observed with the AAVE data in our own dataset, we find that AAVE speaking men have consistently lower performance than AAVE speaking women, typically by around 0.05. Furthermore, while the models performed better on the CORAAL data than on AAVE data in our sample, nearly all still have higher error rate for CORAAL data than for Standard American English data in our sample, especially when comparing to men SAE speakers.

Both results on this dataset and our own provide interesting insights on minority dialect performance. Prior studies which study gender disparities in ASR tend to focus on Standard American English and European dialects, in these instances models tend to perform worse on women's speech (Tatman, 2017). The results on CORAAL further confirm men minority dialect speakers may be a greater risk for ASR failures, despite opposite gender patterns in other dialects.

The results on the vaxpopuli data sample show consistently low error rates across models compared to the other data samples evaluated. We observe no statistically significant difference in performance across gender in this sample. Despite low error rates across wav2vec2, HuBERT, and all other Whisper models, we observe the highest error rate with multilingual whisper tiny across all data with the voxpopuli sample. This is likely due to the representation of multiple English speakers across Europe including different accents. The multilingual whisper then incorrectly classifies the language and hallucinates words from other European langauges.

4.2 Challenges in ASR for Gender and Dialect

Following determining zero-shot results, we conduct a thorough qualitative examination of the results. Below we summarize common themes of errors across models for 100 samples of data and analysis of the results overall. Results for different error types are reported in Table 9.

4.2.1 Proper Nouns

One common theme was failure on proper nouns, wav2vec models and hubert would frequently fail to properly transcribe names of public figures and other individuals, by misspelling them. Whisper models more often would accurately capture the proper nouns, but in some cases misspelled or omitted them, also causing error, as shown in Figure 9. For example, one podcast mentioned the celebrity "*Traji P Henson*" and Whisper-base-en transcribed "*trotty begins*".

4.2.2 Dialect Specific Terms

Another theme was failures of dialect specific terms and words, especially words that are unique to the

	AAVE		Chicano English		Spanglish		SAE	
Model	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
Whisper tiny mul	**0.972**	**0.525***	**0.560**	**0.465**	**0.542**	0.433	0.330	**0.469***
Whisper tiny en	0.380*	0.207*	**1.127***	0.304*	**1.709***	0.301*	0.189	0.231*
Whisper base	0.330	0.218*	**0.881**	0.296	1.283	0.363	0.200	0.222*
Whisper small	0.310*	0.189*	**0.753**	0.299	**1.018**	0.281	0.173	0.202*
Whisper med	0.353*	0.166*	0.447	0.275	**0.580***	0.260	0.179	0.200*
HuBERT	**0.575***	0.318*	0.466*	0.159*	**0.597***	0.321	0.225	0.323*
Wav2Vec2	**0.655***	0.441*	**0.562***	0.212*	**0.702***	0.391	0.294	0.395*
Wav2Vec2 Conformer	**0.564***	0.310*	0.457*	0.154*	**0.594***	0.321	0.228	0.315*

Table 6: Mean word Error Rate on our full dataset of models on gender and dialect combined categories. * Denotes statistically significant pairwise t test result when comparing with SAE men. Full statistical significance results can be found in the appendix.

Model	**Overall**	**Men**	**Women**
Whisper tiny multilingual	0.406	0.451	0.360
Whisper tiny	0.343	0.389	0.296
Whisper base	0.290	0.316	0.264
Whisper small	0.189	0.216	0.161
Whisper med	0.178	0.203	0.152
Hubert large	0.378	0.425	0.329
Wav2vec2	0.466	0.517	0.412
Wav2vec2 Conformer	0.391	0.437	0.344

Table 7: Mean word error rate on CORAAL sample.

Model	**Overall**	**Men**	**Women**
Whisper tiny multilingual	1.174	1.264	0.938
Whisper tiny	0.142	0.144	0.136
Whisper base	0.116	0.115	0.119
Whisper small	0.105	0.104	0.110
Whisper med	0.099	0.100	0.102
Hubert large	0.160	0.156	0.165
Wav2vec2	0.225	0.226	0.222
Wav2vec2 Conformer	0.158	0.155	0.166

Table 8: Mean word error rate on voxpopuli sample.

Model	**Proper Nouns**	**Hallucination**	**Dialect Terms**
whisper tiny multilinugal	5%	12%	4%
whisper tiny	5%	12%	4%
whisper base	6%	2%	2%
whisper small	0%	1%	3%
whisper med	0%	0%	1%
hubert large	11%	0%	2%
wav2vec2	14%	0%	2%
wav2vec2 conformer	13%	0%	4%

Table 9: Percent of samples with each type of error for the 100 samples qualitatively studied for each model.

dialect or have a different spelling depending on the dialect. These words would be represented by the model either using a spelling more common in standard English than the original dialect or completely misrepresenting the word.

As shown in the column of "Dialect Terms" in Figure 9, we find that Whisper more frequently replaces these words with another phonetically similar word from SAE, while wav2vec 2.0 and HuBERT were more likely to produce a phonetically similar output but fall short of correctly capturing it. For instance "*I'm gonna*" becomes "*I'm going to*" with Whisper in one example, while it becomes "*I'm ging*" with HuBERT and "*I'm gon*" with wav2vec. In another example the term "*cholos*" was transcribed as "*children*" by Whisper tiny.

4.2.3 Hallucinations with Repetition

As shown in the column of "Hallucinations" in Figure 9, whisper in particular was susceptible to hallucinations, errors in which the transcriptions are coherent but mostly unrelated to the speech in the audio. Often these include the same word or repeating two words several times, usually the last few words successfully processed by the model. The repetition leads to significantly higher word error due to generating more words than were in the ground truth transcription. This frequently occurred with minority dialect speech, especially Chicano English. This error was found in 12/100 samples generated by Whisper tiny models, whereas these errors were not observed with hubert or wav2vec models. For instance, "*It would have been all the way around yeah yeah, your mom is ballsy though for giving you the option though...*" was transcribed as "*It would have been all the way around. Yeah, yeah....*" by whisper-tiny-en, with "*Yeah, yeah.*"

Model	WER	CER
Whisper tiny Zero Shot	0.521	0.416
Whisper tiny Finetuned	0.437	0.335
Hubert large Zero Shot	0.361	0.191
Hubert large Finetuned	0.245	0.146
Wav2Vec2 Base Zero Shot	0.223	0.128
Wav2Vec2 Base Finetuned	0.204	0.119
Wav2Vec2 Conformer Rope Zero Shot	0.351	0.19
Wav2Vec2 Conformer Rope Finetuned	0.178	0.108

Table 10: Zero-shot and finetuned model results on a subset of held out data. Whisper tiny multilinugal is used for this evaluation.

repeated over 30 times.

4.2.4 Other Challenges

We also observe other types of challenges that are relatively infrequent but are crucial for understanding the ASR performances. The first one is **accented speech** (Hinsvark et al., 2021). Within Spanglish and Chicano English samples, transcription often fails even when dialect specific terms or Spanish terms aren't included in the text. We speculate this is due to the accented pronunciation, this can result in small errors like leaving off some characters or larger errors like generating unrelated text. Another challenge is around **multiple speakers**. Consistent with prior work (Li et al., 2023), across models, the error rate is higher on audios that have multiple speakers compared to those that have only one speaker. Within audios with multiple speakers, minority dialects are more represented in our data, however we find that when controlling for multiple speakers in an audio, minority dialect data still consistently performs worse than SAE.

4.3 Fine-tuning

We present results for vanilla fine-tuning on a set of models compared to the zero-shot performance. We finetune models on on a subset of the held out data in Table 10. Models were fine-tuend on a subset of the dataset balanced for gender and dialect categories and tested on a held out set of balanced data. We find that the Wav2vec2 Conformer rope finetuned model had the largest improvement with our finetuning, showing the largest growth in performance on WER (17.3%) and achieving the lowest word error rate overall on the held out dataset.

4.4 Recommendations for Improving Performance for Minority Dialects

Based on our results, we conclude with two recommendations for improving performance of ASR models on minority dialects.

1. Diversity of training data including but not limited to dialect and gender diversity. We observe lower performance of minority dialect speech, especially when including multiple minority dialects at a time, for men minority dialect speakers, and when there are multiple speakers in general. Prior studies have shown that spontaneous and casual speech have less accurate performance on ASR than scripted reading (Butzberger et al., 1992; Riviere et al., 2021) and single speaker audio has better performance than multiple speakers (Chang et al., 2020). With our work, we observe the worst minority dialect performance when there are multiple speakers using minority dialects, showing that these two issues with data diversity may compound.

2. Language model dictionary expansion for dialect specific words. As expansive as dictionaries for large models like Whisper compared to smaller models like wav2vec, this dictionary excludes many dialect specific terms. This likely contributes to the behavior we observe of dialect specific terms being more likely to be captured correctly by wav2vec and HuBERT, whereas Whisper models replace these terms with SAE terms. In contexts where dialect aware systems are more desirable, expanded dictionaries including dialect specific terms should be included. In downstream applications such as for transcription, custom models with user-provided terms could also improve performance.

5 Conclusion

In our work we create a small dataset with diverse gender and dialect representations by using the Spotify Podcast Dataset (Clifton et al., 2020) and our dialect-centered annotation process. We determine the word error rate of state of the art ASR models on our dataset. We find that across all models, Standard American English outperforms minority dialects. Further we find that within minority dialects women speakers perform better, but within SAE men speakers perform better, suggesting men of color, particularly minority dialect users, may be at the highest risk for inaccurate transcription within English.

6 Limitations

This study is subject to several limitations. One limitation with our data was the level of audio annotation; annotators labeled metadata by podcast episode, rather labeling individual audio snippets. While this was more cost effective and efficient for data labeling, it limits our analysis. Further, with gender labels we limit our study to binary gender. This is due in part to the lack of data on non-binary identities that could be easily identified in our starting dataset. As discussed in the methods, we instruct annotators to assess gender by a combination of self-identification of the speaker in the audio, in the podcast description, online searches of the podcast, or if they aren't able to assess gender to have a followup conversation with the research team about the sample to determine the best annotation. We recognize the use of binary gender labels and leaving gender assessment to be determined by annotators as a limitation that can often lead to mis-categorization and erasure of non-binary and transgender individuals (Scheuerman et al., 2020). Another limitation is the distribution of codeswitching in our data. We find that by gender, more podcast samples with women minority dialect speakers contain codeswitching than those with men minority dialect speakers. Of podcasts with AAVE speakers, 87% of samples with women speakers were indicated as having some amount of codeswitching, compared to none of those containing men's speech. Similarly all podcasts with Chicano English and Spanglish speaking women in our dataset were indicated as having some codeswitching compared to none for Chiacno English and Spanglish speaking men. We observe that AAVE women speech that is not indicated as codeswitching still yields much lower WER than AAVE speaking men in our dataset, and observe that women AAVE speakers have better performance with the CORAAL data, which shows that the gender-dialect trend of lower performance for men may still hold beyond codeswitching. But further analysis of how codeswitching and density of dialect features (Demszky et al., 2020) would offer an even richer analysis of how dialect impacts ASR performance. Finally, another limitation of this work was limited access to the source dataset. As of December 2023, Spotify ceased maintaining the dataset and limited access. While we were able to complete experiments on the data sampled using the sampling methods described in the Methods section, additional data would have allowed for more robust analyses of gender-dialect groups.

7 Ethical Consideration

There are multiple ethical considerations with respect to this work. Firstly, while our work shows that future iterations or versions of the models studied may be improved for minority dialect speaker with training on diverse dialect data, we limit encouraging this as training large models comes with a high environmental impact (Tokayev, 2023). Secondly, minority dialect speech should be protected from cultural appropriation and malicious use. Scholars in African American Studies and related fields have written at length about the misuse of AAVE by non-AAVE speakers for profit, social capital, and other benefits, while authentic AAVE speakers continue to be discriminated against (Roth-Gordon et al., 2020). Collecting AAVE and other minority dialect data as done in this study could encourage this type of misuse, and training models on such resources could lead to downstream applications that perpetuate this same problematic dialect misuse for AAVE or other minority dialects studied. Finally, we consider the downstream application of ASR, voice assistants. Prior work has shown many users of voice assistants fear their data being collected by their smart speaker or phone without explicit their consent for targeted ads or other unwanted uses (Seymour et al., 2023; Voit et al., 2020). Existing inaccuracies in ASR systems could serve as a barrier to the privacy issue; less accurate processing of the speech data could make it less likely to be used effectively in unwanted ways.

8 Acknowledgements

We would like to thank members of the SALT who gave feedback on the work, especially Caleb Ziems for his support as research mentor. We would also like to thank the reviewers for their thoughtful feedback on the work.

References

Alfredo Ardila. 2005. Spanglish: an anglicized spanish dialect. *Hispanic Journal of Behavioral Sciences*, 27(1):60–81.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Multilingual speech models for automatic speech recognition exhibit gender performance gaps. *arXiv preprint arXiv:2402.17954*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

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*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

John Butzberger, Hy Murveit, Elizabeth Shriberg, and Patti Price. 1992. Spontaneous speech effects in large vocabulary speech recognition applications. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Eugenia Casielles-Suárez. 2017. Spanglish: The hybrid voice of latinos in the united states. *Atlantis*, pages 147–168.

Xuankai Chang, Wangyou Zhang, Yanmin Qian, Jonathan Le Roux, and Shinji Watanabe. 2020. End-to-end multi-speaker speech recognition with transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6134–6138. IEEE.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jay L. Cunningham. 2023. Collaboratively mitigating racial disparities in automated speech recognition and language technologies with african american english speakers: Community-collaborative and equity-centered approaches toward designing inclusive natural language systems. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA. Association for Computing Machinery.

Dorottya Demszky, Devyani Sharma, Jonathan H Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2020. Learning to recognize dialect features. *arXiv preprint arXiv:2010.12707*.

Charlie Farringon and Tyler Kendall. 2024. Home. `https://oraal.github.io`. Accessed: 2024-10-4.

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151.

Reem Gody and David Harwath. 2023. Unsupervised fine-tuning data selection for asr using self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Camille Harris, Matan Halevy, Ayanna Howard, Amy Bruckman, and Diyi Yang. 2022. Exploring the role of grammar and word choice in bias toward african american english (aae) in hate speech classification. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 789–798.

Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. "honestly, i think tiktok has a vendetta against black creators": Understanding black content creator experiences on tiktok. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, et al. 2021. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

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.

Youjin Kong. 2022. Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 485–494, New York, NY, USA. Association for Computing Machinery.

Bernd Kortmann and Kerstin Lunkenheimer. 2012. *The Mouton world atlas of variation in English*. de Gruyter.

Chenda Li, Yao Qian, Zhuo Chen, Naoyuki Kanda, Dongmei Wang, Takuya Yoshioka, Yanmin Qian, and Michael Zeng. 2023. Adapting multi-lingual asr models for handling multiple talkers. *arXiv preprint arXiv:2305.18747*.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2022. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6162–6166. IEEE.

John Myhill. 1995. The use of features of present-day aave in the ex-slave recordings. *American Speech*, 70(2):115–147.

Mikel K Ngueajio and Gloria Washington. 2022. Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review. In *International Conference on Human-Computer Interaction*, pages 421–440. Springer.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Gabriela T Richard and Yasmin B Kafai. 2016. Blind spots in youth diy programming: Examining diversity in creators, content, and comments within the scratch online community. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*, pages 1473–1485.

Morgane Riviere, Jade Copet, and Gabriel Synnaeve. 2021. Asr4real: An extended benchmark for speech models. *arXiv preprint arXiv:2110.08583*.

Jennifer Roth-Gordon, Jessica Harris, and Stephanie Zamora. 2020. Producing white comfort through "corporate cool": Linguistic appropriation, social media, and@ brandssayingbae. *International Journal of the Sociology of Language*, 2020(265):107–128.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW1):1–35.

William Seymour, Xiao Zhan, Mark Coté, and Jose Such. 2023. A systematic review of ethical concerns with voice assistants. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 131–145, New York, NY, USA. Association for Computing Machinery.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

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

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Information Review*.

Kassym-Jomart Tokayev. 2023. Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*, 8(9):17–33.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088*.

Alexandra Voit, Jasmin Niess, Caroline Eckerth, Maike Ernst, Henrike Weingärtner, and Paweł W. Woźniak. 2020. 'it's not a romantic relationship': Stories of adoption and abandonment of smart speakers at home. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*, MUM '20, page 71–82, New York, NY, USA. Association for Computing Machinery.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 336–349, New York, NY, USA. Association for Computing Machinery.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

A Appendix

Code used for this project can be found at https://github.com/camille2019/asr_modeling, the dataset can be found at https://https://huggingface.co/datasets/SALT-NLP/spotify_podcast_ASR

A.1 Annotation Instructions

Annotators were given a short PowerPoint presentation explaining the task and a guide explaining what each column of metadata they would record in addition to the transcription. To begin annotators started out transcribing audios on the utterance level, recording the start and end time of each utterance, but we quickly pivoted to 30 second intervals which was easier for transcribers and more appropriate for ASR training.

A.2 Statistical Significance Tests

Statistical significance tests were conducted using the list of generated WERs for each group within each model. Table 12 depicts statistical significance results for men and women. Table 11 depicts the statistical significance results for each minority dialect compared to SAE.

A.3 Hyperparameters

Hyperparameters for the best performing finetuned models reported in Table 10 shown in Table 16

Dialect Pair	Model	P Value	Statistic
AAVE and SAE	Whisper tiny multilingual	0.608	0.517
	Whisper tiny en	0.098	-1.67
	Whisper base en	0.030	-2.19
	Whisper small	0.034	-2.150
	Whisper med	0.563	-0.583
	hubert large	0.038	2.159
	wave2vec2	0.030	2.260
	wav2vec2 conformer	0.038	2.160
Chicano English and SAE	Whisper tiny multilingual	0.758	0.310
	Whisper tiny en	0.034	2.182
	Whisper base en	0.061	1.919
	Whisper small	0.117	1.594
	Whisper med	9.554e-7	5.572
	hubert large	3.769e-7	5.820
	wave2vec2	1.261e-7	6.124
	wav2vec2 conformer	4.991e-7	5.745
Spanglish and SAE	Whisper tiny multilingual	0.783	0.277
	Whisper tiny en	0.026	2.299
	Whisper base en	0.047	2.034
	Whisper small	0.101	1.670
	Whisper med	3.417e-7	5.863
	hubert large	5.6337e-8	5.863
	wave2vec2	1.77e-8	6.669
	wav2vec2 conformer	7.731e-8	6.267

Table 11: Pairwise t-test results for minority dialects compared to SAE across models.

Model	P value	Statistic
whisper tiny multilingual	0.921	0.099
whisper tiny en	0.002	-3.192
whisper base en	0.003	-2.975
whisper small en	0.010	-2.590
whisper med en	3.055e-8	-5.626
hubert large	2.768e-7	-5.196
wav2vec2	4.561e-11	-6.710
wav2vec2 conformer	4.169e-8	-5.196

Table 12: Gender result pairwise t-test comparing results on podcasts with only men speakers and podcasts with only women speakers across models.

Pair	Model	P value	Statistic
AAVE women and SAE Women	whisper tiny multilingual	1.578e-5	-5.066
	whisper tiny en	5.551e-8	-6.618
	whisper base en	1.309e-05	-5.053
	whisper small en	3.787e-7	-6.283
	whisper med en	1.726e-7	-6.465
	hubert large	0.0006	-3.904
	wav2vec2	0.028	-2.353
	wav2vec2 conformer	0.008	-2.940
AAVE Men and SAE Women	whisper tiny multilingual	0.189	-1.376
	whisper tiny en	0.083	-1.853
	whisper base en	0.060	-2.010
	whisper small en	0.216	-1.289
	whisper med en	0.113	-1.679
	hubert large	1.069e-05	-6.299
	wav2vec2	9.270e-08	-8.822
	wav2vec2 conformer	4.991e-06	-6.700
AAVE Men and SAE Men	whisper tiny multilingual	0.0936	-1.780
	whisper tiny en	0.0117	-2.820
	whisper base en	0.0984	-2.797
	whisper small en	0.0268	-2.413
	whisper med en	0.0327	-2.329
	hubert large	3.645e-7	-8.133
	wav2vec2	1.510e-9	-10.992
	wav2vec2 conformer	2.330e-7	-1.3273
AAVE Women and SAE Men	whisper tiny multilingual	0.0007	-3.5952
	whisper tiny	0.0003	-3.6404
	whisper base	0.00012	-3.8722
	whisper small	6.732e-5	-4.0309
	whisper med	4.958e-11	-8.019
	hubert large	1.004e-5	5.091
	wav2vec2	0.0002	-4.389
	wav2vec2 conformer	0.0002	-4.311

Table 13: Gender-dialect pairwise significance tests on AAVE gender groups and SAE groups.

Pair	Model	P value	Statistic
Chicano English Women and SAE Women	whisper tiny multilingual	0.958	0.053
	whisper tiny en	0.030	**-2.218**
	whisper base en	-2.315	0.023
	whisper small en	0.003	-3.040
	whisper med en	0.013	-2.527
	hubert large	2.892e-23	12.01064
	wav2vec2	3.475e-23	12.026
	wav2vec2 conformer	1.441e-26	12.772
Chicano English Men and SAE Women	whisper tiny multilingual	0.437	0.780
	whisper tiny	0.0258	-2.271
	whisper base	0.041	-2.078
	whisper small	0.0667	-1.858
	whisper med	1.627e-06	-5.118
	hubert large	7.505e-6	4.749
	wav2vec2	7.053e-07	-5.327
	wav2vec2 conformer	1.203e-5	-4.632
Chicano English Men and SAE Men	whisper tiny multilingual	0.372	-0.900
	whisper tiny en	0.02011	-2.402
	whisper base en	0.0382	-2.129
	whisper small en	0.0730	-1.832
	whisper med en	3.858	-6.347
	hubert large	1.1359e-10	-7.995
	wav2vec2	2.631e-11	-8.367
	wav2vec2 conformer	4.194e-10	-7.642
Chicano English Women and SAE Men	whisper tiny multilingual	0.864	0.171
	whisper tiny en	0.0468	-1.9971
	whisper base en	0.058	-1.901
	whisper small en	0.132	-1.509
	whisper med en	0.358	-0.922
	hubert large	4.053e-21	-10.236
	wav2vec2	1.958e-23	**-10.967**
	wav2vec2 conformer	1.393e-22	-10.632

Table 14: Gender-dialect pairwise significance tests on Chicano English gender groups and SAE groups.

Pair	Model	P value	Statistic
Spanglish Women and SAE Women	whisper tiny multilingual	0.573	0.565
	whisper tiny en	0.004	**-2.895**
	whisper base en	0.022	-2.333
	whisper small en	0.002	-3.216
	whisper med en	0.010	-2.582
	hubert large	0.559	-0.586
	wav2vec2	0.553	-0.595
	wav2vec2 conformer	0.581	-0.553
Spanglish Men and SAE Women	whisper tiny multilingual	0.963	0.0470
	whisper tiny en	0.001	-3.342
	whisper base en	0.002	-3.167
	whisper small en	0.011	-2.567
	whisper med en	1.578e-13	-7.982
	hubert large	3.546e-16	-8.871
	wav2vec2	4.189e-18	-9.559
	wav2vec2 conformer	4.204e-16	-8.863
Spanglish Men and SAE Men	whisper tiny multilingual	0.714	0.367
	whisper tiny en	0.039	2.083
	whisper base en	0.058	1.907
	whisper small en	0.170	1.378
	whisper med en	8.250e-06	4.558
	hubert large	4.858e-13	7.617
	wav2vec2	3.904e14	7.996
	wav2vec2 conformer	1.118e-12	7.490
Spanglish Women and SAE Men	whisper tiny multilingual	0.827	0.2184
	whisper tiny	0.043	-2.036
	whisper base	0.246	-1.161
	whisper small	0.084	-1.732
	whisper med	0.088	-1.713
	hubert large	0.407	0.831
	wav2vec2	0.246	1.164
	wav2vec2 conformer	0.313	-1.011

Table 15: Gender-dialect pairwise significance tests on Spanglish gender groups and SAE groups.

Model	Learning Rate	Train Batch Size	Seed	Optimizer	LR Scheduler Type	LR Warmup Steps	Num Epochs
Wav2vec	0.0001	16	42	Adam with beta=(0,9, 0.999) epsilon =1e-8	linear	500	20
Wav2vec conformer rope	0.0001	8	42	Adam with beta=(0,9, 0.999) epsilon =1e-8	Linear	500	20
Hubert Large	0.0001	16	42	Adam with beta=(0,9, 0.999) epsilon =1e-8	linear	100	40
Whisper tiny	0.00001	16	42	Adam with beta=(0,9, 0.999) epsilon =1e-8	linear	500	3

Table 16: Hyperparameters for best performing finetuned models.