A Comparative Analysis of Transcription Errors from Major Commercial Automatic Speech Recognition Systems on Speakers of Four Ethnic Backgrounds in the Pacific Northwest

Michael Kelly Scott

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2023

Committee:

Gina-Anne Levow

Alicia Beckford Wassink

Program Authorized to Offer Degree:

Department of Linguistics

University of Washington

Abstract

A Comparative Analysis of Transcription Errors from Major Commercial Automatic Speech Recognition Systems on Speakers of Four Ethnic Backgrounds in the Pacific Northwest

Michael Kelly Scott

Chair of the Supervisory Committee:
Gina-Anne Levow
Department of Linguistics

Major commercial ASR systems have demonstrated higher transcription error rates for non-white American English speakers, particularly for African American speakers, and there is evidence that sociophonetic features are highly associated with these errors (Koenecke et al., 2020; Wassink et al., 2022). In this thesis, I analyze the transcription results of four major commercial ASR systems—Apple Speech, Amazon Transcribe, Google Speech-to-text, and IBM Watson Speech-to-text—on recordings from the Pacific Northwest English (PNWE) corpus originally collected for Wassink (2015), and I attempt to answer two research questions:

1. Do sociophonetic markers typical of African American Language (AAL) correlate with higher inaccuracy rates in major commercial ASR systems for African American speakers than for speakers of different ethnic backgrounds?
2. Are there any phonological features representative of AAL which appear more frequently on incorrectly transcribed speech for African American speakers than for other co-regional speakers?

To do this, I ran automatic transcription on recordings of 16 speakers from four ethnic backgrounds—African American, Caucasian American, ChicanX, and Yakama—for all four ASR systems evaluated. I identified ten target linguistic variables which represent common sociophonetic markers of African American Language (AAL) and identified co-occurrences of these markers with tran-

scription errors for each ASR system in order to perform both a quantitative and heuristically informed qualitative analysis.

From this, I determined that the resistance to the low-back merger and the pre-nasal front merger (pen-pin merger) are both most strongly associated with errors for the African American speakers than for any other ethnic group, and that consonant cluster reduction is more strongly associated with errors for the Yakama and Caucasian American speakers than for the African American speakers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ABSTRACTION: The process in linguistics by which certain details of specific phenomena are ignored because they are irrelevant to the analysis being performed, e.g., speech sounds can be abstracted to "phonemes," which are units of sounds that distinguish words within a language

AAL: (African American Language) The group of languages spoken within African American communities

APPLICATION: Computer software designed to carry out a specific task

API: (Application Programming Interface) A type of shared boundary across separate software systems which provides the definitions necessary for one software system to use the functionality of another

ASR: (Automatic Speech Recognition) An application of computational linguistics that, among other tasks, is largely focused on identifying spoken language and translating it into text

AWS: (Amazon Web Services, Inc.) A subsidiary of Amazon.com, Inc. that provides cloud-computing services

CLI: (Command-line Interface) A text-only software application that interprets and executes specific commands typed by the user

CSV: (Comma-Separated Value) A file format that uses commas to separate values into a table-like structure

EBONICS: A term created by a group of African American scholars in 1973 to refer to the language of African American speech-communities; a portmanteau of "ebony" and "phonics"

GAE: (General American English) The umbrella accent/prestige accent of American English

iOS: The mobile operating system used by Apple's iPhones

LEVENSHTEIN DISTANCE: A measurement of the distance between two sequences, such as phonemes in a word or words in a sentence, determined by the minimum number of insertions, deletions, or substitutions required to turn one sequence into the other

macOS: The Unix operating system used by Apple's Mac computers

PERPLEXITY: A measurement in information theory of how well a distribution predicts a given sample; this measure is commonly used to evaluate transcribed text in ASR or text from generative models

PER: (Phonetic error rate) An evaluation metric for the performance of ASR systems derived from the Levenshtein distance of a word sequence on the phonemic-level, as opposed to the word-level

SDK: (Software Development Kit) A collection of software definitions and development tools for a specific platform to facilitate the creation of software applications

STT: Speech-to-text

WAVE: Waveform Audio File Format; also known as `WAV`

WER: (Word error rate) An evaluation metric for the performance of ASR systems derived from the Levenshtein distance of a word sequence on the word-level, as opposed to the phoneme-level

Chapter 1

INTRODUCTION

1.1 Problem Statement

Communities of color in the United States are demonstrably under-represented in Automatic Speech Recognition (ASR) technologies, particularly African Americans (Koenecke et al., 2020). For the purposes of this thesis, I am defining "racial bias in ASR" to refer to a higher rate of transcription errors for ethnolects of speakers who identify as "persons of color" than for Caucasian/white-identifying speakers. Specifically, Koenecke et al. (2020) found that, across five major commercial ASR systems, African American speakers experienced an average 0.35 error rate, compared to an average 0.19 error rate for Caucasian American speakers.

The most likely predominant causes for this disparity for African Americans in particular are

1. the lack of African American-speaker training data used to train the models behind these systems, and
2. the lack of consideration for sociophonetic differences in African American Language (AAL) varieties of English.

Koenecke et al. (2020) indicated that language models do not demonstrate similar bias for African American speakers, yielding lower perplexity, i.e., the measurement of how well the language model predicts utterances, for AAL utterances than for General American English utterances, despite higher word error rates (WERs). Koenecke et al. (2020) therefore concluded that the results "must be due" to phonetic factors. Evidence provided by Wassink et al. (2022) seems to corroborate this conclusion. Given these observations, it seems prudent to continue to focus future research in the "racial bias" ASR space on phonetic factors from a sociolinguistic perspective.

In this master's thesis work, I investigated the correlation of transcription errors with sociophonetic variables identified in Wassink et al. (2022) for four of the five major ASR systems evaluated in Koenecke et al. (2020), using speaker data from the Pacific Northwest English (PNWE) Corpus (Wassink, 2015). Those systems are the Apple Speech SDK, Amazon Transcribe from Amazon Web Services (AWS)[1], Google Speech-to-Text from Google Cloud, and Watson Speech to Text from IBM Cloud (Amazon, 2022; Apple, 2022; Google, 2022; IBM, 2022). I evaluated an heuristically determined phonetic error rate (PER) of these systems as well as word error rate (WER) for the purposes of comparison with Koenecke et al. (2020). In particular, I focused on sociophonetic markers most common in African American Language (AAL) to see whether those markers significantly contribute to higher error rates for African American speakers. I also examined speaker data for the Mexican American (ChicanX), Yakama Nation, and Caucasian American ethnic groups in the PNWE corpus to control for confounding variables.

1.2 Research Questions

1. Do sociophonetic markers typical of African American Language (AAL) correlate with higher inaccuracy rates in major commercial ASR systems for African American speakers than for speakers of different ethnic backgrounds?
2. Are there any phonological features representative of AAL which appear more frequently on incorrectly transcribed speech for African American speakers than for other co-regional speakers?

1.2.1 Linguistic Variables

Wassink et al. (2022) identified the following sociolinguistic variables as particularly significant for PER among African American speakers:

- consonant cluster reduction (CC),

[1]I was employed by Amazon for the entirety of this work, but I have never had any internal affiliation with Amazon Transcribe nor with AWS. I address this further in Section 5.6: Ethical Considerations.

- *th*-stopping (TH),
- the merger of /ɪ/ and /ɛ/ before the nasal consonants [m], [n], and [ŋ], otherwise known as the *pen-pin* merger (IN),
- the merger of /ʊ/ and /uː/ before /l/ as well as /u/ or /ʌ/ and /o/ before /l/, otherwise known as the pre-lateral back vowel merger (prel),
- "other vowel error[s]" denoted with the variable (V).

Of these, consonant cluster reduction (CC) (particularly in the final position), *th*-stopping (TH), and the *pen-pin* merger (IN) are particularly common and/or unique to AAL (Thomas, 2007).

With these, I have included in my examination the following sociolinguistic variables common in AAL:

- word-final devoicing (Dv),
- word-final debuccalization of /t/ and /d/ (Db),
- *th*-fronting (TH-f), as opposed to *th*-stopping, which I will now denote as (TH-s),
- the merger of /ɪ/ and /iː/ before /l/, otherwise known as pre-lateral front vowel merger (prel-i) or the *fill-feel* merger, as opposed to the pre-lateral back vowel merger, which I will now denote as (prel-o),
- the monophthongization of /aɪ/ to [aː~äː] (AY), and
- the absence of or resistance to the low-back merger of /ɔ/ to [a] or [ɑ], otherwise known as *caught-cot* merger (-AO).

In addition to being common markers of AAL (Farrington, 2018; Thomas, 2007), many of these variables have been called out in Prof. Wassink's Bias in ASR research group as of particular interest for future work. It is important to note that I have begun participation in this research group and plan to continue the work begun in this thesis in order to contribute to the research group's goals.

For the resistance to the low-back merger variable (-AO), it is important to clarify that in identifying this variable, I looked for words for which my chosen "canonical" pronunciation

dictionary had both /ɔ/ and /ɑ/ pronunciation entries, i.e., "caught", /kɔt/ or /kɑt/. I then identified an utterance as (-AO) if it met this criterion and the speaker uttered the /ɔ/ variant. While this may seem like I am only examining the merger in one direction, this is the easiest way to assert resistance to the merger. For this work, I am making the assumption that there are likely few if any words in which General American English (GAE) produces only with /ɔ/ but many speakers produce them with /ɑ/; similarly there are likely few if any words in which GAE produces only with /ɑ/ but many speakers produce them with /ɔ/. However, it is important to note that the reality is more complex, and in fact, as demonstrated by (Hall-Lew, 2013), when a merger is in progress, there are instances when vowel classes may flip-flop such that words historically produced only with /ɑ/ may be produced with /ɔ/, and vice-versa. I discuss my approach with the low-back merger in greater detail in Section 3.2.4

Table 1.1 and Table 1.2 respectively outline my target consonantal and vowel linguistic variables that I examine as markers of AAL varieties and likely to be correlative with automatic orthographic transcription errors. The tables also include examples of orthographic transcription errors.

1.3 Motivation

While it is likely that collecting and utilizing more training data from African African speakers will improve error rates in ASR systems, the gap in demographic distribution in the United States makes this exceedingly difficult. If the goal of creating a single language model that is applied to all populations requires more training data than is representative of minority populations, then minority inequities will likely persist, simply because the burden of data collection will be significantly higher on smaller populations. Even where communities of color are no longer significant minorities of the population, systemic inequities continue to exist, which still unfairly increase the burden of data collection on those communities, compared to non-marginalized communities. Further, given that Koenecke et al. (2020) found that ASR WERs were higher for AAL speakers, despite the fact that language model perplexities were *lower* for AAL speakers, it stands to reason that research should focus on phonetic factors, and that is my primary motivation in pursuing

Table 1.1: Targeted consonantal linguistic variables and example transcription errors

Word-final Devoicing (Dv)	**Word-final Debuccalization (Db)**	**Consonant Cluster Reduction (CC)**		**Th-fronting (TH-f)**	**Th-stopping (TH-d)**
/b/ → [p] /d/ → [t] /g/ → [k]	/d/ → ∅ or ʔ /t/ → ∅ or ʔ	/nt/ → [n] /nd/ → [n] /md/ → [m] /ft/ → [f]	/st/ → [s] /sp/ → [s] /sk/ → [s] /ts/ → [t]* /ts/ → [s]*	/θ/ → [f] [ð] → [v]	/θ/ → [t] [ð] → [d]
cub → cup had to → hat to bug → buck	side → sigh** exhibited → exhibit	pant → pan hand → hen named → name craft → carafe	test → Tess last → lass cats → cast	with → whiff	this → dis

* Clusters ending in /s/ or /z/ exhibit variation in whether the first or second element is deleted (Labov, 1972, pp. 17–18)

** In particular, Farrington (2018) indicates that vowels preceding word-final neutralization of /d/ are significantly longer than those preceding /t/ and states that this may be an AAL marker.

this work.

In the United States, according to the U.S. Census Bureau (2020), roughly 13.4% of people self-identified their race as Black or African American alone; 76.5% self-identified as white or Caucasian American alone. It is therefore a reasonable requirement than any universal ASR tool should perform equally well on all represented speakers having used training data representative of the population distribution, e.g., ~77% of training data from Caucasian American speakers, and ~13% from African American speakers. However, it is clear that systems trained with more data from a specific group perform better for that group (Dorn, 2019), suggesting that proportionate data collection under current conditions will continue to yield disproportionate results. This disparity argument further suggests that we are far more likely to see success by incorporating sociophonetic knowledge into the design of ASR systems than from simply increasing the amount

Table 1.2: Targeted vowel linguistic variables

Prenasal Front Merger (IN)	**Absence of Low Back Merger (-AO)**	**Prelateral Back Vowel Merger (prel-o)**	**Prelateral Front Vowel Merger (prel-i)**	**Monophthongization of /aɪ/ (AY)**
Before [m, n, ŋ]: /ɪ/ ↔ /ɛ/	/ɔ/ ↛ [a] or [ɑ]	Before /l/: /u/ ↔ /o/ /ʊ/ ↔ /u/ /ʌ/ ↔ /o/	Before /l/: /ɪ/ → [iː] /eɪ/ ↔ /ɛ/	/aɪ/ → [aː∼äː]
pen ↔ pin	caught ≠ cot	fool ↔ full hole ↔ hull	feel ↔ fill fail ↔ fell	mile → mall

* The ↛ symbol denotes that the merger does not occur

of representative training data.

Finally, I draw motivation from Prof. Geneva Smitherman, a University Distinguished Professor Emerita of English and co-founder of the African American and African Studies doctoral program at Michigan State University—and a personal hero of mine. In one of her many writings advocating for African American children's education, Prof. Smitherman stated, "today's negative pronouncements on Ebonics reveal a serious lack of knowledge about the scientific approach to language as well as galling ignorance about what Ebonics is . . . Most critically, these pronouncements reveal an appalling rejection of the language of everyday Black people" (Smitherman, 1998, p. 99). This lack of knowledge persists today and natural language technology fields are not immune to it. As ASR technologies rapidly become widespread, AAL varieties are still in danger of being further marginalized, left behind at best and ridiculed or demonized at worst.

In Chapter 2, I review the literature which inspired and informed my work, focusing on a general overview of ASR technology and how sociolinguistic knowledge has been applied to evaluation of ASR performance.

In Chapter 3, I summarize the subset of the PNWE corpus I used in my experiments and outline

the methodology of those experiments. From this, I propose a repeatable method by which to find and analyze the co-occurrence of phonetic markers with transcription errors across multiple automatic orthographic transcription systems.

In Chapter 4, I summarize the results of my experiments, and in Chapter 5, I analyze and discuss those results, which suggest that at least two of my target sociophonetic markers are more frequently associated with word errors for African American speakers than for other co-regional speakers.

Finally in Chapter 6, I suggest directions that future, expansive work should take.

Chapter 2

LITERATURE REVIEW

To explain the approaches I take in this study to evaluate bias in major commercial ASR systems, I must review three bodies of literature:

1. the functions of and mechanisms behind automatic speech recognition and how ASR systems can be prone to racial bias;
2. sociolinguistics, particularly sociophonetics and the sociolect known as African American Language; and
3. efforts to apply sociolinguistic knowledge to the evaluation of ASR systems' performance.

Through this review, I contextualize the methods I outline in Chapter 3 and how I intend to answer my research questions.

2.1 Automatic Speech Recognition (ASR)

Automatic speech recognition (ASR) is an application of computational linguistics that, among other tasks, is largely focused on identifying spoken language and translating it into text. ASR is thriving in commercial applications, and as such, implementations can vary dramatically and few are transparent to the public, in order to protect trade secrets. However, I can make some generalizations about the theory of practical ASR implementation.

In general, most ASR systems have two major components: the acoustic model and the language model. The acoustic model represents a statistical mapping from an acoustic waveform to some representation of spoken language, e.g., phonemes or words. The language model represents a statistical prediction of the frequency and order of words, e.g., n-grams. The function of a typical ASR system is to use the acoustic model to generate some textual representation of speech and then use the language model to correct errors and make the result more natural (Lamere et al.,

2003).

One common approach to general-purpose ASR—particularly for small-vocabulary systems—is based on the hidden Markov model (HMM), approximating speech as a Markov process. These systems typically characterize speech as a sequence of states, e.g., words or phones. This results in a model that can produce a probability for each observed sequence of tokens, which the ASR systems can then maximize via some decoding method such as the Viterbi algorithm and then map from phonetic realizations to graphemes via a manually determined pronunciation dictionary (Lee, 1990; Rabiner & Juang, 1986).

Some systems that utilize HMMs train neural networks to generate a probability distribution for a sequence of acoustic frames and then use those distributions as emission probabilities for the HMM, which produces acoustic scores to be rectified by the language model. However, this approach has resulted in a disconnect between the objective functions used to train those neural networks, which are primarily phonetic, and the inputs used to train the speech recognition systems, which are primarily lexical. This disconnect often requires not only the labor-intensive generation of pronunciation dictionaries but also other labor-intensive efforts that require significant expertise (Graves & Jaitly, 2014). In their presentation, Tatman (2020) points out that as a result of requiring significant amounts of manually tailored data and dictionaries, HMMs that are trained on one or a few language varieties are prone to make transcription errors when they encounter other varieties.

In general, the HMM approach to ASR models language independently of acoustic productions, because the language models are trained on text corpora. As discussed above, this approach requires significant expert labor, and as such has motivated research into end-to-end ASR systems, which directly map acoustic sequences to orthographic representations and therefore require much less expertise in annotation. However, end-to-end ASR systems do require substantially more annotated training data. Perhaps the most popular end-to-end approach is the connectionist temporal classification (CTC)-based system, first described in Graves and Jaitly (2014). Essentially, the system comprises a single recurrent neural network (RNN) trained directly on transcribed speech corpora. This network processes spectrograms of speech and the

CTC output layer, which was chosen based on the assumption that speech is essentially a sequence with variable timing, outputs individual graphemes. While this type of system is capable of matching many uttered phones to their corresponding English graphemes, it naturally misspells many words and cannot properly handle digraphs such as "th" or "au", and the same is true for diphthongs which often map to a singular grapheme, such as "a" for /eɪ/ or "o" for /oʊ/. In general, CTC systems may naturally make mispellings whenever there is a one-to-many or many-to-one relationship between phones and graphemes. As a result, CTC-based systems typically utilize a separate language model to de-conflict and correct these mistakes, mapping the outputs to English words (Graves & Jaitly, 2014). One consequence of this end-to-end nature is that these systems can be prone to errors for language variants that do not closely resemble the texts used to train the language model. Because the system uses the language model to essentially correct spelling mistakes, phonetic markers such as th-stopping or consonant cluster reduction, which are rarely written, may be incorrectly transcribed as different words entirely, e.g. "whiff" for "with" or "lass" for "last."

2.1.1 "Racial Bias" in ASR

For the purposes of this work, I have defined "racial bias in ASR" to refer to a higher rate of transcription errors for ethnolects of speakers who identify as "persons of color" than for Caucasian/white-identifying speakers.

In particular for ASR systems, the struggle to interpret strong accents or English ethnolects other than General American or British English has become a staple of pop culture. In 2011, the BBC Scotland sketch comedy show *Burnistoun* aired a sketch about a voice-activated elevator that couldn't understand Scottish accents BBC Scotland (2011). In 2017, *Wired* magazine published an article by Sonia Paul in which she describes the comical yet frustrating mistakes Amazon's Alexa on the Echo Dot makes when interpreting the speech of her immigrant parents, both of whom have spoken American English fluently for over 50 years, albeit with strong accents (Paul, 2017). In general, it is well known in American society that major commercial ASR systems struggle to properly transcribe American English ethnolects other than the mainstream General American

spoken by most Caucasian Americans, as many mainstream media outlets such as the Verge and the New York Times reported on Koenecke et al. (2020).

More concretely, Dorn (2019) first described the disproportionately high error rates ASR systems demonstrated for AAL, citing how unrepresented the dialect is in major speech corpora, despite its being widely spoken. Furthermore, Dorn demonstrated that, unsurprisingly, systems with models trained on AAL data, or on a combination of AAL and General American English (GAE) data, experienced fewer errors than systems with models trained only on GAE, improving by over 16.6% (Dorn, 2019).

Koenecke et al. (2020) took this investigation one step further in analyzing word error rates (WERs) for five popular commercial ASR systems and comparing the results. The researchers determined not only that there was a unilateral major disparity in WER between Caucasian American speakers and African American speakers, but that major commercial language models such as GPT-2 on average perform better, i.e., with lower perplexity, for transcribed AAL speech than for transcribed GAE speech. This unexpected disparity between the performance of ASR systems and the performance of language models led the researchers to propose that phonetic variation was more likely the underlying cause for transcription errors, rather than other linguistic variations such as syntax or lexicon (Koenecke et al., 2020).

2.2 Sociolinguistics

The core focus of sociolinguistics is the description of the effects that culture and society have on language, i.e., the sociocultural significance and context of language variation (Chambers, 2009). Sociophonetics is a subfield of sociolinguistics that is concerned with how sociocultural variation manifests specifically in sounds (Di Paolo & Yaeger-Dror, 2011). The seminal work Labov (1970) describes the linguistic variable, which is a dependent variable that represents a linguistic feature with two or more variants which can be predicted with the independent variables of sociocultural aspects or demographics, such as geography, gender, age, ethnicity, and economic class (Labov, 1970). The linguistic variable is commonly denoted with some defining name in parentheses, such as Labov (1970)'s `(r)` for the variation of the English phoneme /r/ in New York City, i.e.,

"the presence or absence of consonantal [r] in postvocalic position in *car, card, four, fourth*, etc." The linguistic variable represents the core theoretical construct by which sociophonetic research describes the unique defining features of a particular lect (Chambers, 2009).

Koenecke et al. (2020) demonstrates—through the hypothesis that phonetic markers are the primary contributors to transcription errors—that it is highly likely that most modern commercial ASR systems do not consider sociophonetic information in design or in training models. Because of this, it is appropriate in this work for me to identify and investigate a set of linguistic variables that contribute to the phonetic description of AAL.

2.2.1 African American Language (AAL)

Variously referred to as Black English, Ebonics, African American English (AAE), African American Vernacular English (AAVE), and African American Language (AAL), AAL refers generally to variants of English spoken within the African American community (Kendall et al., 2018). Two major early works describing specific features of regional variants of AAL are Wolfram (1969) for Detroit and Labov (1972) for Harlem, New York City, both of which had a major focus on how age contributes to variation within AAL.

While there is a great diversity of variation across North America within AAL, there are also several phonological markers which frequently recur and thus act as unifying characteristics of AAL. Those markers, which I have chosen and outlined in Table 1.1 and Table 1.2, include:

- consonant cluster reduction (Green, 2002; Labov, 1972),
- *th*-stopping and *th*-fronting (Green, 2002),
- the merger of /ɪ/ and /ɛ/ before the nasal consonants [m], [n], and [ŋ], otherwise known as the *pen-pin* merger (Labov, 1972),
- word-final devoicing (Farrington, 2018; Green, 2002),
- word-final debuccalization of /t/ and /d/ (Farrington, 2018; Thomas, 2007),
- the merger of /ʊ/ and /uː/ before /l/ as well as /u/ or /ʌ/ and /o/ before /l/, otherwise known as the pre-lateral back vowel merger (Labov, 1972),

- the merger of /ɪ/ and /iː/ before /l/, otherwise known as pre-lateral front vowel merger or the *fill-feel* merger (Labov, 1972),
- the monophthongization of /aɪ/ to [aː~äː] (Labov, 1972), and
- the absence of or resistance to the low-back merger of /ɔ/ to [a] or [ɑ], otherwise known as the low-back merger or the *caught-cot* merger (Thomas, 2007).

2.3 Applying Sociolinguistic Understanding to ASR Evaluation

Surprisingly little research has been done that applies sociolinguistic understanding to the evaluation of ASR systems' performance (Wassink et al., 2022). Recently, there have been at least three significant efforts:

1. Tatman and Kasten (2017), which examines the effects of gender and race on the accuracy of Bing Speech and YouTube Automatic Captions,
2. Koenecke et al. (2020), which details a comparative analysis of WERs for five popular commercial ASR systems (Amazon, Apple, Google, Microsoft, and IBM) on both African American and Caucasian American speakers, and
3. Wassink et al. (2022), which expands both on Wassink (2015) and Koenecke et al. (2020) to examine the correlation between PER and sociophonetic features of four ethnolects of the Pacific Northwest (African American, Caucasian American, ChicanX, and Yakama).

2.3.1 Tatman and Kasten (2017)

Tatman and Kasten (2017) compared the WERs of Bing Speech and YouTube's automatic captions when transcribing four varieties of American English from recordings from the Dialects of English Archive, a dialect resource for actors created by Paul Meier. The varieties covered by the recordings included General American, Northern Cities, Southern, and Californian. They ran automatic transcriptions for Bing Speech by writing a custom Android app using the Bing Speech SDK, and for YouTube Automatic Captions by using the speech files as soundtracks to `mp4`-encoded videos that they uploaded to YouTube.

For evaluation, because many of the Bing Speech transcripts were partial, they discounted deletion errors, resulting in a modified version of WER. They then performed a comparative analysis of these results across dialect, speaker gender, and speaker race. With respect to dialect, they found that both systems performed best on General American and worst on Californian, though the disparities were far less significant for Bing Speech than they were for YouTube Automatic Captions. With respect to speaker race, they found that for both systems, WERs were higher for African American and mixed race speakers than for Caucasian American speakers, though the disparity was far more pronounced for YouTube Automatic Captions than for Bing Speech.

Tatman and Kasten (2017) represents one of the earliest significant academic evaluations of commercial ASR systems that considers sociolinguistic variation, specifically across dialect background, gender, and race. The work I have done in this thesis is similar in these aspects, but it focuses more narrowly on specific sociophonetic markers, in order to determine if there is association with these specific markers and word errors.

2.3.2 Koenecke et al. (2020)

The work that follows in this thesis was first inspired by the conclusions drawn in Koenecke et al. (2020), which ultimately proposed that future investigations into racial bias in ASR should be phonetic. In their work, they evaluated five major commercial ASR systems, identified as "state-of-the-art ASR systems—developed by Amazon, Apple, Google, IBM, and Microsoft" (Koenecke et al., 2020). They used two corpora for their speech data to be transcribed: the Corpus of Regional African American Language (CORAAL) for their African American speaker data, and the Voices of California (VOC), specifically the predominantly Caucasian American communities of Sacramento and Humboldt County.

For evaluation, they ran automatic transcriptions on this data for all five evaluated systems and then calculated standard WER for each of them. They then performed a comparative analysis of these results across systems and the identified race groups, as well as by speaker location. With respect to speaker race, they found that across all systems, African American speakers experienced significantly higher WERs than the Caucasian American speakers: 35% vs. 19%.

Koenecke et al. (2020) hypothesized that the language models likely accounted for the higher number of word errors for African American speakers, but ultimately found that African American speaker data tended to have lower perplexity than the Caucasian American speaker data. This discovery led them to propose that future work should focus on phonetic features unique to AAL, which they suspected must be the largest contributor to increased WER. This proposal for future work first inspired the work I have done in this thesis.

For my work, I have chosen 10 sociophonetic markers commonly found in AAL and have identified their realization in the speaker data and whether those markers co-occur with transcription errors from the ASR systems evaluated, four of the five systems evaluated in Koenecke et al. (2020). Moreover, the dataset I used, the Pacific Northwest English Corpus (Wassink, 2015) comprises a more narrow set of speaker locations, i.e., the Pacific Northwest, and also includes two additional two additional ethnic groups: ChicanX and Yakama. This co-regional variation in ethnic identity should help further control for confounds as I attempt to determine the association of markers with ethnic identity and transcription errors.

2.3.3 Wassink et al. (2022)

More recently, I was inspired by work on racial bias in ASR by a small research team led by Prof. Alicia Wassink, which I joined shortly before the time of writing, as part of the Pacific Northwest English Project (PNWE) (Wassink et al., 2022). The Bias in ASR group's work aims to "leverage sociolinguistic knowledge of the fine phonetic detail in dialect variation" in order to facilitate the aforementioned investigations.

Specifically, Wassink et al. (2022), examines the impact of a set of dialect features of Pacific Northwest English on transcription errors by Microsoft's Speech SDK. The corpus they used is roughly the same as the one I used in this thesis, except that, in addition to the reading passage task and word list task, they examined lexical game tasks and conversational speech. After performing the automatic transcriptions, they examined all transcription errors and determined whether any target markers were present, and if so, attributed the presence of those markers to those errors. In addition to confirming Koenecke et al. (2020)'s results of higher error rates

for African American speakers than for Caucasian American speakers, they also demonstrated higher PERs for ChicanX and Yakama speakers. Furthermore, they found that dialectal features accounted at least 20% of errors, a clear demonstration of the tremendous value of applying sociophonetic knowledge to the evaluation and ultimately training of ASR systems.

My work in this thesis was more narrowly focused on 10 sociophonetic markers common in AAL varieties. I also made the decision to discount errors of homophony, in order to reduce noise, particularly with the word list task. Furthermore, I chose to perform a comparative analysis across four major commercial ASR systems, similar to Koenecke et al. (2020), in order to establish whether there are systemic patterns in the technology itself, as opposed to with specific implementations.

2.4 Summary

ASR systems currently have a clear problem with racial bias in terms of transcription errors, particularly for African Americans. Sociolinguistics is a field of study concerned with how language varies across social strata, including race. Therefore, it seems logical that sociolinguistic understanding should be applied in the field of ASR in an attempt to mitigate this systemic bias issue.

Moreover, Thomas and Reaser (2004) makes it clear that sociolinguistic variation is systematic, as evidenced in part by their demonstration that many Americans can accurately identify African American speakers even without prominent AAL features. Ultimately, because of this systematic variation, it is vital to apply the knowledge of linguistic variation along sociocultural variables. As Tatman (2020) aptly puts, "Any automated system trained predominately on one variety will not work as well for other varieties." If the acoustic models or language models of ASR systems are trained without consideration for variation, ASR systems will continue to perform poorly on non-mainstream varieties.

My work aims to combine the methods of both Koenecke et al. (2020) and (Wassink et al., 2022) in order to add weight to the observations that sociophonetic variation in AAL strongly correlates with disproportionately higher error rates. In Chapter 3, I describe the methodology I

developed to that end.

Chapter 3

METHODOLOGY

3.1 Data

In this work, I used speaker data examined in Wassink et al. (2022) from the corpus generated by the Social Networks of Ethnic Minority Members in Washington State as part of the Pacific Northwest English (PNWE) Study (Wassink, 2015). The subset of the corpus I examined contains recorded speech of 16 speakers from four different ethnic groups, identified by the speakers themselves, their social network membership, and how long they've been in the speech community. Table 3.1 describes the breakdown of the examined speakers by ethnic group and sex. Each recording is between 45 and 90 minutes, with each speaker having a minimum of 20 minutes of speech, for a total of 13 hours of recorded speech.

Each speaker participated in the following tasks:

1. Free-flowing speech in a casual dyadic style using generated common topics,
2. Lexical word games including making lists, reading minimal pairs, and answering questions on semantic differentials, and
3. Passage reading from Aesop's Fable "The Cat and the Mouse" (Wassink, 2015).

Each speaker ID encodes demographic data which includes ethnicity and a unique identifier code, among other items. This enabled me to filter my results by ethnicity, using regular expressions to extract the relevant demographic data.

All of the speech data in the PNWE corpus were recorded in unsegmented, stereo `wave` files with speakers and facilitators on the same channels. These files were then downsampled and extracted to single-channel for use with their transcription system (Wassink, 2015). Along with these recordings, I used Praat TextGrids (Boersma & Weenink, 2022), created in Wassink (2015),

that provided time-aligned narrow phonetic transcriptions and orthographic transcriptions for all of the speech. These TextGrids are time-aligned for separate tasks and provide a means for separating out individual tasks from the larger combined `wave` files. The phonetic transcription tier for these TextGrids is done using the ARPABET transcription code format, except in cases where phones are not represented, such as /ɣ/; in these cases, the corresponding IPA character was used. At the time of writing, only the Reading Passage (RP) and Word List (WL) tasks have TextGrids for all examined speakers. As a result, I examined only these two tasks, which cover roughly 4:30 hours of recorded speech. Both tasks naturally result in more heavily attended speech than some of the lexical tasks or the free-flowing conversation, and this represents a significant limitation of the work herein.

Finally, all recordings include speech from both the speakers and the facilitators of the recording tasks. However, the included Praat TextGrids annotate the facilitator speech with special silence markers to identify them for removal.

Ethnic Group	# Male Speakers	# Female Speakers	Total Speakers	Total Time	Total Utterances
Yakama	3	2	5	2:11:32	11,297
Mexican American	2	2	4	1:26:59	10,456
African American	1	2	3	1:23:00	7719
Caucasian American	1	3	4	1:15:39	8375
Total	8	7	16	6:17:13	37,847

Table 3.1: Summary of speaker data used from the PNWE corpus

3.2 Experimental Approach

Koenecke et al. (2020) evaluated five major commercial speech-to-text (STT) systems: Apple Speech, Amazon Transcribe from Amazon Web Services (AWS), Google Speech-to-Text from Google Cloud, IBM Watson Speech-to-Text[1] from IBM Cloud, and Microsoft Azure's Speech SDK.

[1]Koenecke et al. (2020) would have had access only to what is, at the time of writing, referred to as the "prevgen" IBM Watson models. This prevgen model was deprecated in 2021 and scheduled for removal on 15 September 2022.

I chose to evaluate the same systems from a phonological perspective rather than an orthographic perspective, and I ran all of my transcription jobs on each system in July and August 2022 (Amazon, 2022; Apple, 2022; Google, 2022; IBM, 2022). However, I elected not to evaluate Microsoft Azure's SDK, because the work in Wassink et al. (2022) evaluated Microsoft's STT system exclusively. Whereas Koenecke et al. (2020) evaluated these systems strictly by comparing WERs across systems for each target ethnic group, I evaluated the systems by comparing the rates of word errors associated with targeted sociophonetic markers, for each evaluated ethnic group. For example, I compared how often the marker (CC) coincided with a word error, for each ethnic group, across evaluated systems.

It is important to note that because these systems are "black boxes," I do not have access to the phonetic determinations the systems make when generating orthographic transcriptions. In fact, it is unlikely that these systems make phonetic determinations at all[2]. Therefore, I decided to compare manual narrow transcriptions created in Wassink (2015)'s work with "canonical" phonemic representations of the STT systems' hypotheses, using a superset of CMUdict (Carnegie Mellon University, 2015) created for the PNWE corpus (hereafter referred to as PNWEdict), to generate an heuristically determined PER. As this phonemic analysis is inherently speculative, it is important to note that it does not represent literal phonetic transcription errors—after all, these systems do not make phonetic transcriptions of any kind[3]—but rather provides a controlling mechanism for errors caused by incorrect word boundary decisions, which do not typically result from phonetic realizations.

My experimental procedure was as follows:

1. properly format speaker data for each evaluated STT system, upload data, and collect or-

---

I chose to take advantage of the timing to include separate evaluations for the prevgen and nextgen IBM Watson models in order to more accurately compare to the work in Koenecke et al. (2020)

[2]This is discussed in more detail in Section 2.1

[3]While the technology presumed to be used by these commercial systems do not make phonetic transcriptions per se, they do make classification outputs which are graphemic, which may or may not correlate with phonetic categories

thographic transcriptions;

2. perform standardization tasks on manual and automatic transcriptions;
3. determine "canonical" pronunciations of all words using the PNWEdict on the orthographic transcriptions from the STT systems and compare them with manual narrow transcriptions in the TextGrids from Wassink (2015) to provide an heuristically determined phonetic error rate (PER);
4. automatically identify candidate utterances for targeted sociophonetic markers, using regular expressions;
5. from these candidates, manually verify markers which also coincide with errors; and
6. calculate marker-aligned error rates from those correlative errors.

Figure 3.1 visually outlines the flow of this procedure.

3.2.1 Pre-processing

In order to prepare the data for transcription, I first had to extract the reading passage and word-list tasks for which I had TextGrids. Using PraatIO (Mahrt, 2016) to extract the time-aligned boundaries, I created new `wave` files for only the RP and WL tasks, respectively for each speaker. Once the files were segmented by task, I used PraatIO to find all intervals annotated as silence and used Pydub (Robert, Webbie, et al., 2018) to silence that interval in the `wave` file. This ensured that no facilitator speech would be transcribed by the ASR systems tested. Finally, each ASR system has its own restriction for the length of an audio file it can transcribe; Google and Apple are the most restrictive at one minute (Apple, 2022; Google, 2022). To account for this, I modified an existing script developed for Wassink (2015) that searches given `wave` files for low amplitude sub-intervals within a 59-second interval and segments the `wave` into new files that are less than 59 seconds. Each segmented file was annotated with the original filename and the timestamp at which the segment fit in the original file, in order to facilitate later re-stitching.

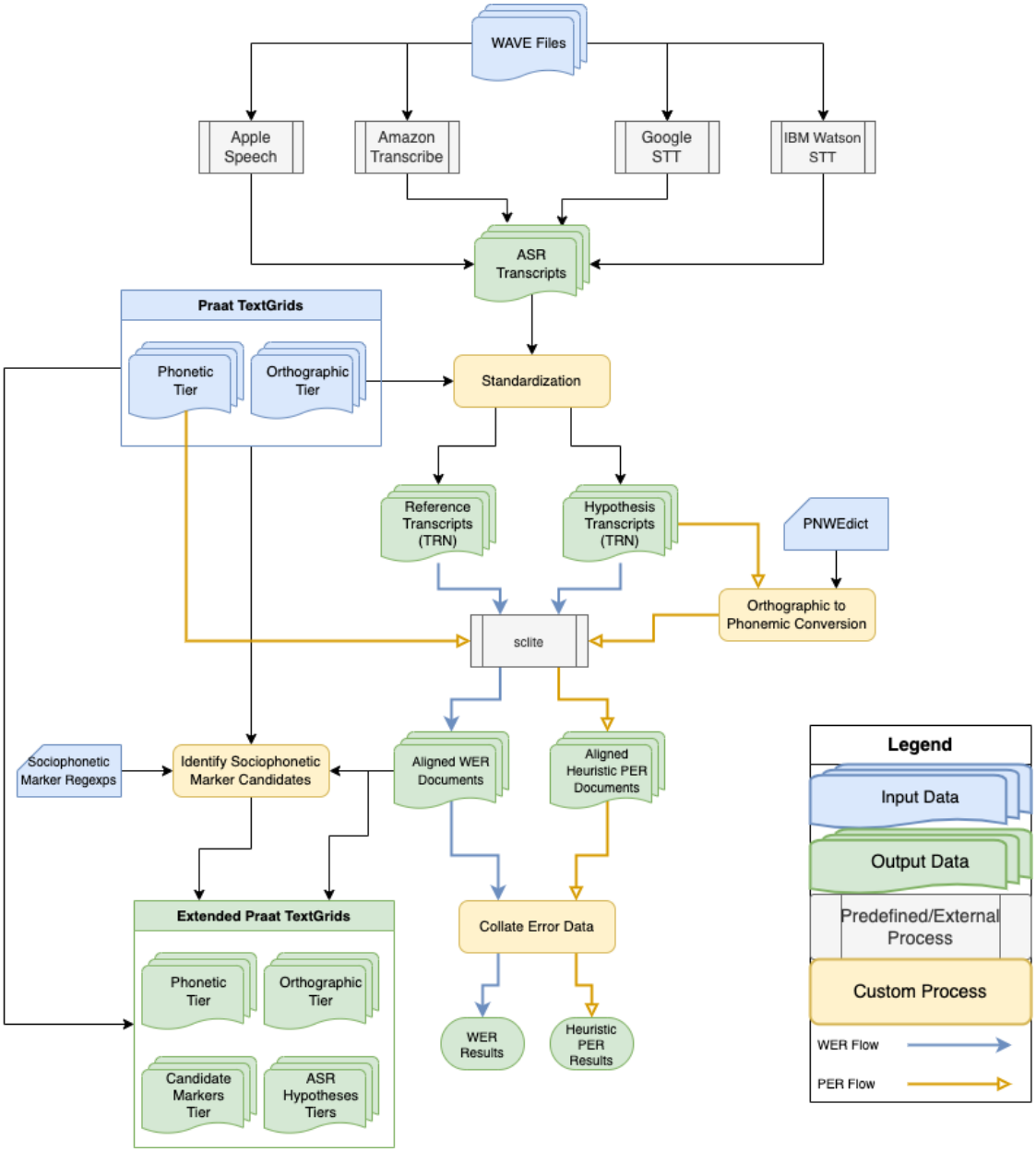


Figure 3.1: Flowchart of experimental procedure

3.2.2 Transcription

Each of the four services evaluated has its own software development kit (SDK) for interacting with the speech-to-text API. AWS, Google Cloud, and IBM Cloud each require the creation of a

commercial account to access their APIs, which have fee-based usage, but also have some form of free-tier service or new account credits (Amazon, 2022; Google, 2022; IBM, 2022). Apple's STT service is free, but requires the creation of an iOS or macOS application and has a limited number of requests (Apple, 2022). It is important to note that the purpose of Apple's STT service differs from the others: it is intended to be used with iOS or macOS applications for digital assistant-type use cases exclusively (Apple, 2022), whereas AWS, Google Cloud, and IBM Cloud's services have broader applications, such as automatic transcription of conversations or the creation of captions for videos.

For each service, I wrote an application with a command-line interface (CLI) that queued each of my segmented `wave` files for transcription and collated the results into a standardized comma-separated value (CSV) format. I then stitched the resulting transcriptions together into one combined CSV file organized by system, speaker, and task.

Finally, in order to facilitate evaluation, I needed to create a "gold standard" transcript for each recording. The Praat TextGrids mentioned in Section 3.1 contain both a narrow phonetic transcription tier and an orthographic transcription tier, manually aligned and verified for Wassink (2015). Using PraatIO, I extracted each word from the orthographic transcription tiers for each TextGrid and added the full text to a "gold-standard" entry for each file to the master transcription CSV mentioned above.

3.2.3 Standardization & Evaluation

In order to evaluate word error rate (WER), I used the "Score Lite", or `sclite`, utility from the National Institute of Standards and Technology (NIST) Scoring Toolkit (SCTK) (NIST, 2021). One of the methods `sclite` uses to align reference transcripts with hypothesis transcripts requires the transcripts to be formatted in a specific transcript format called `trn`. I converted each entry in my master transcript CSV to the `trn` format as outlined by the `sclite` documentation and annotated each file with its speaker ID, task type, transcription system, and whether it was a reference or hypothesis file.

Standardization

I then executed several text standardization tasks, which included the following:

- remove[4] any special silence, pause, or disfluency markers, such as the %HESITATION marker added by IBM Watson's prevgen model;
- lowercase all words, to ensure instances like "Word" vs. "word" aren't incorrectly annotated as errors;
- reduce multiple spaces to one space, to ensure uniformity in spacing;
- remove all punctuation, including for contractions such as "can't" and "won't"[5];
- convert all numerals, e.g., "21" or "21st" to their corresponding words, e.g., "twenty one" or "twenty first," using `num2words` (Dupras, 2019), to eliminate false positives between different transcription conventions;
- replace dashes (-) with spaces, to avoid false positives in situations like "twenty-one" vs "twenty one";
- remove all remaining punctuation, again to eliminate false positives between different conventions; and
- Americanize any British spellings such as "colour" or "favourite" to "color" or "favorite," respectively, using the `hyperreality/American-British-English-Translator` dictionary (hyperreality, 2022).

I applied these standardization tasks to both the "reference" and "hypothesis" sets of orthographic transcriptions to ensure the calculated orthographic transcription errors weren't false positives.

---

[4]NB: These markers can optionally be converted to @ for the reference transcripts, according to `sclite`'s documentation

[5]NB: The punctuation removal means that there are potentially missed errors, such as for the homographous "wont" with "won't', but these are exceedingly rare and are accounted for later with the heuristically determined phonetic error rate

Accounting for homophones

After standardization, I then created a homophone dictionary of all words that existed in all reference and hypothesis transcripts, using CMUdict (or a customized dictionary in the same format, such as the aforementioned PNWEdict) to determine pronunciation of each word. Some words—particularly inflections for plural nouns, third person verbs, gerunds/participles, etc.—are not in CMUdict/PNWEdict. In those cases, I applied the following rules to add the appropriate phonemes, according to regular pronunciation rules in English:

1. If the word ends in `'s`, drop those characters and see if the resulting word is in the pronouncing dictionary
    (a) If so, for each phonemic representation, add the correct possessive phoneme(s) (i.e., `S`, `Z`, or `IH0 Z`), and add that pronunciation
2. If not, and the word ends in `n'` or `in`, drop the apostrophe if present and add a `g` and see if that word is in the pronouncing dictionary
    (a) If so, add that resulting pronunciation
3. If not, add the resulting raw word, because there is no phonemic representation, so it's unlikely to have a collision anyway.

Using that homophone dictionary, I created a Global Mapping (GLM) rule file, which is an SCTK file format that defines tokens which are interchangeable. In this case, I created a homophone dictionary entry for each ARPABET-coded phonemic output, e.g.,

`'R-AY1-T': ['RIGHT', 'WRITE', 'RITE', 'WRIGHT']`.

It is important to note that there are some limitations with this approach, particularly with respect to homographs, e.g.,

- `'R-IY1-D': ['READ', 'REED']`
- `'R-EH1-D': ['READ', 'RED']`

where "read" as in "let's read" will get scored as correct if it is transcribed as "red," as in the color. The SCTK script `csrfilt` uses this GLM rule file to make substitutions in the transcript

files so that `sclite` will score, for example, the hypothesis token "right" as correct even if the reference token is "write." This process enabled me to ignore errors caused by homophony, which are not interesting to the phonetic analysis and also not strictly "fair" for the word-list task, which has no meaningful context for speech-to-text systems to use in determining the correct transcription for the carrier phrase "write ____ today."

It is important to note that this method relies on the provided pronouncing dictionary. If the dictionary accounts for mergers such as the "caught-cot" merger, i.e., it includes both /kɔt/ and /kɑt/ pronunciations for "caught", then it will mark "caught" and "cot" as homophones and not count them as transcription errors for one another. However, if the pronouncing dictionary does *not* account for a merger, such as the "pen-pin" merger, i.e., it includes only the /pɛn/ pronunciation for "pen" and not /pɪn/, it will mark those mergers as substitution errors. Both of these examples are the case for CMUdict.

I applied this homophony accounting method to both the WER and heuristically-determined PER calculations. This means that, in the "pen-pin" example above, if the speaker uttered /pɪn/ for "pen", and the ASR system correctly transcribed the word as "pen", it would be marked as correct in the WER calculation, but it would be considered a phone error, even though it is not an orthographic error This effectively means the method would penalize the language model for correctly identifying the situation. While these situations are likely fairly rare, future work may see some benefit from ensuring more mergers are accounted for in the pronunciation dictionary.

Heuristically determined phonetic hypothesis transcripts

Finally, for the phonetic transcripts, I simply extracted the narrow phonetic transcriptions from the Praat TextGrids to generate the reference transcripts and converted the orthographic hypothesis transcript tokens to individual phonemes using the pronouncing dictionary to generate the "heuristically determined" phonetic hypothesis transcripts.

Evaluation

For both the orthographic transcripts and the phonetic transcripts, I used `sclite` to output transcript alignments for each file and report the numbers of correct tokens, inserted tokens, substituted tokens, and deleted tokens, as well as the overall word/phonetic error rates. From this data from all the raw output files, I created a combined result summary `csv` that contains for each system:

- errors by type for a speaker and one task,
- total errors for a speaker across all tasks,
- total errors for a task type across all speakers,
- total errors within an ethnic group, and
- total errors across all speakers for the system.

3.2.4 Sociophonetic Marker Identification & Error Evaluation

I used the aligned transcripts output by `sclite` to add new tiers to the TextGrids for each system, which included intervals for the hypothesis tokens. Because these aligned transcripts have the same number of tokens, I was able to ensure each hypothesis tier is correctly aligned with the word tier. Correct tokens are written in all lowercase, and substitutions are written in all upper case. Deleted tokens are marked with asterisks (*), and inserted tokens are marked with `INS:` followed by all inserted tokens in brackets, and then the next token, which may be correct, a substitution, or a deletion, e.g., `INS: [CAN] calmly`. Figure 3.2 illustrates a sample error-aligned segment output from `sclite`.

```
Scores: (#C #S #D #I) 8 2 1 1
REF: at first the mice ** AVOIDED the cat like the PLAGUE
HYP: at first the mice TO VOID    the cat **** the PLAY
Eval:                  I  S               D        S
```

Figure 3.2: Sample segment of `sclite` alignment output with error coding

I then identified potential sociophonetic markers using these new TextGrids to identify whether they correlated with transcription errors, using regular expressions that I defined for each of the target sociophonetic markers to identify candidate markers. For each marker, I wrote an exhaustive list of pairs of regular expressions, representing the "standard" form, e.g., /d/ for devoicing `(Dv)` and the marker form, e.g., /t/ for `(Dv)`. These lists included separate pairs for initial, medial, and final positions, as well as separate pairs for different realizations, such as /b,p/ and /d,t/, for `(Dv)`, to reduce complexity and increase accuracy.

I then used these regular expressions on the `phone` and `word` tiers of the TextGrid to identify target markers. I created a new `markers` tier for each TextGrid, then counted all markers in that tier that coincide with errors in the `hypothesis` tiers for each system. Finally, I generated a summary of results for all speakers, for each ethnic group, and for each system. These final modified TextGrids enabled me to manually verify and validate identified sociophonetic markers easily. Figure 3.3 demonstrates what intervals in these expanded TextGrids look like when viewed in Praat (Boersma & Weenink, 2022).

All code for these tasks will be posted on GitHub, with except of CLOx-Preprocessor, which will be on the Bias in ASR website and/or GitHub[6].

There is an important distinction to make regarding how my error-coding method differs from the method used in Wassink et al., 2022. Wassink et al., 2022 looked at all errors and classified them into one of 17 possible sociophonetic error types and several general error types. I looked at all errors where a marker *could* be present. For example, if the speaker uttered "caught" /kɔt/, there could be the presence of the `(-AO)` marker, because the pronouncing dictionary has both /kɑt/ and /kɔt/ as canonical pronunciations. Therefore, if an error coincided with this, I counted it as an error potentially caused by `(-AO)`. This differs from Wassink et al., 2022, in which they looked at both the utterance and the error and made a determination of whether the error was caused by the marker.

[6]https://github.com/scottmk/bias-in-asr

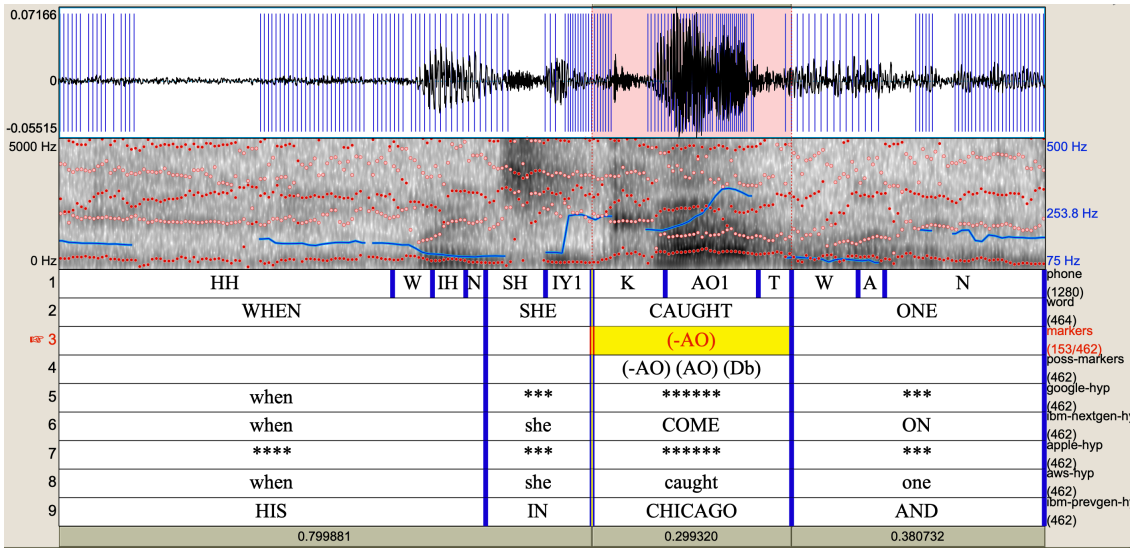


Figure 3.3: Screenshot of intervals in expanded Praat TextGrid, showing tiers for realized sociophonetic markers, possible markers, and ASR hypotheses

3.3 Summary

Utilizing the orthographic "word" tiers in the Praat TextGrids provided with the `wave` file recordings of speech data in the PNWE corpus (Wassink, 2015; Wassink et al., 2022), I generated reference transcripts by which to evaluate WER for each system's hypothesis transcripts using `sclite`. I was able to do the same evaluation for PER by utilizing the TextGrids' phonetic "phone" tiers for the reference transcripts and comparing them with phonetic hypothesis transcripts "heuristically determined" by means of a custom pronunciation dictionary based on CMUdict (Carnegie Mellon University, 2015; Wassink et al., 2022).

Finally, I evaluated the association of word errors with the presence of my target linguistic variables from Table 1.1 and Table 1.2 using regular expressions to identify the markers from the "phone" tiers of the TextGrids and the "canonical" pronunciations of the tokens in the corresponding "word" tiers. This enabled me to align the presence of target variables with word errors, in order to make comparisons across speaker ethnic groups and ASR systems. I then selected several interesting or pertinent errors and listened to the relevant sections of the original recordings

to manually audit the transcriptions and inform my qualitative analysis. I have detailed these results in Chapter 4.

Chapter 4

RESULTS

In this chapter, I summarize the results of all of my experiments. The focus of these experiments was threefold:

1. Determine word error rates (WER) and heuristic phonetic error rates (PER) for each evaluated system, by ethnicity;
2. Determine the number of contexts where target sociophonetic markers might occur and where they were actually realized; and
3. Determine the number of co-occurrences between target sociophonetic markers and word errors.

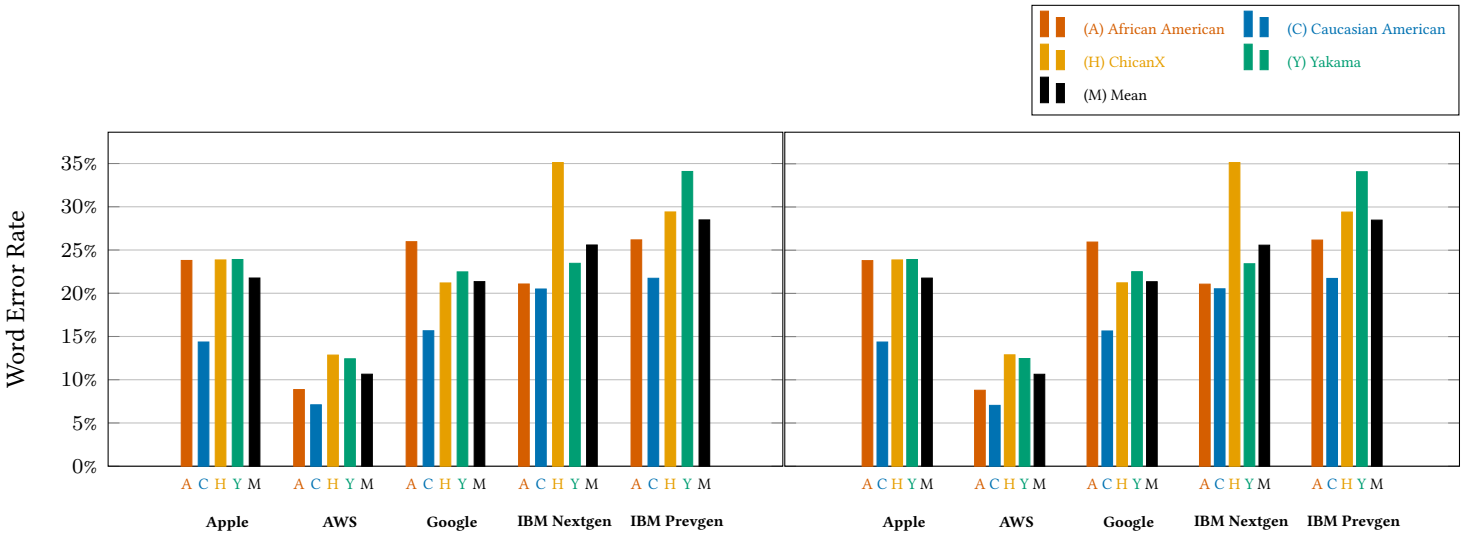
These three components were critical to answering my research questions:

1. Do sociophonetic markers typical of African American Language (AAL) correlate with higher inaccuracy rates in major commercial ASR systems for African American speakers than for speakers of different ethnic backgrounds?
2. Are there any phonological features representative of AAL which appear more frequently on incorrectly transcribed speech for African American speakers than for other co-regional speakers?

This chapter contains a summary of my experimental results. In Chapter 5, I discuss the implications of these results and possible answers to the above questions.

4.1 Word Error Rates (WER)

Figure 4.1 summarizes WER for each speaker ethnic group across all evaluated systems, as well as the mean WER for each system. Across all systems, Caucasian American speakers experienced the lowest WER at 14% for Apple, 7% for AWS, 16% for Google, 21% for IBM Nextgen, and 22% for



(a) CMUdict-derived homophones

(b) PNWEdict-derived homophones

Figure 4.1: Word Error Rates (WER) by ethnicity for each transcription system

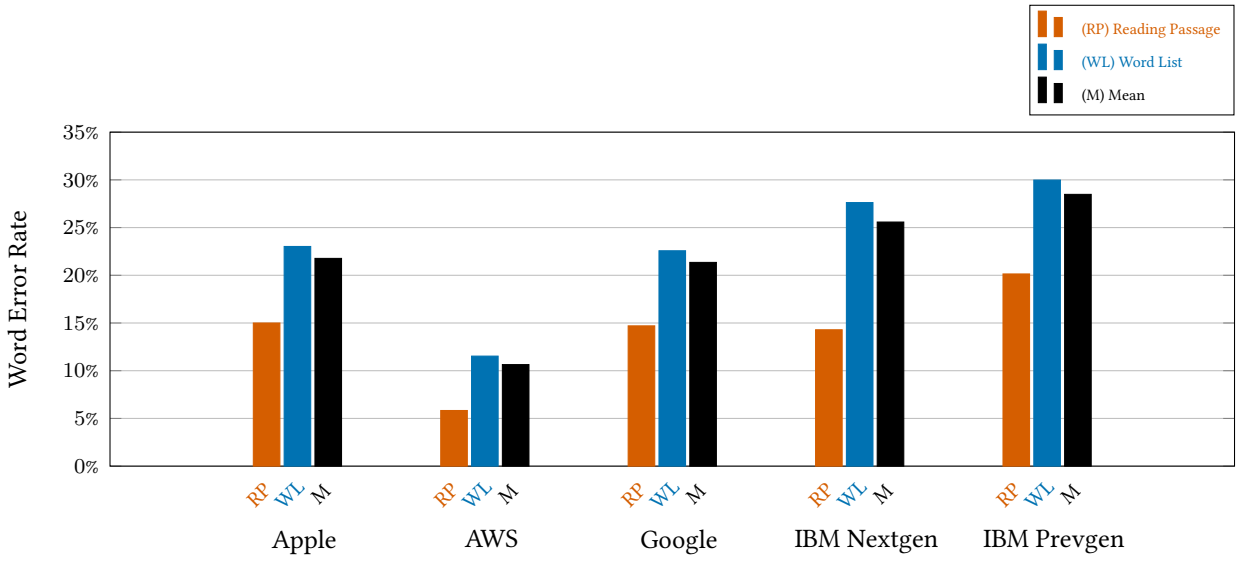


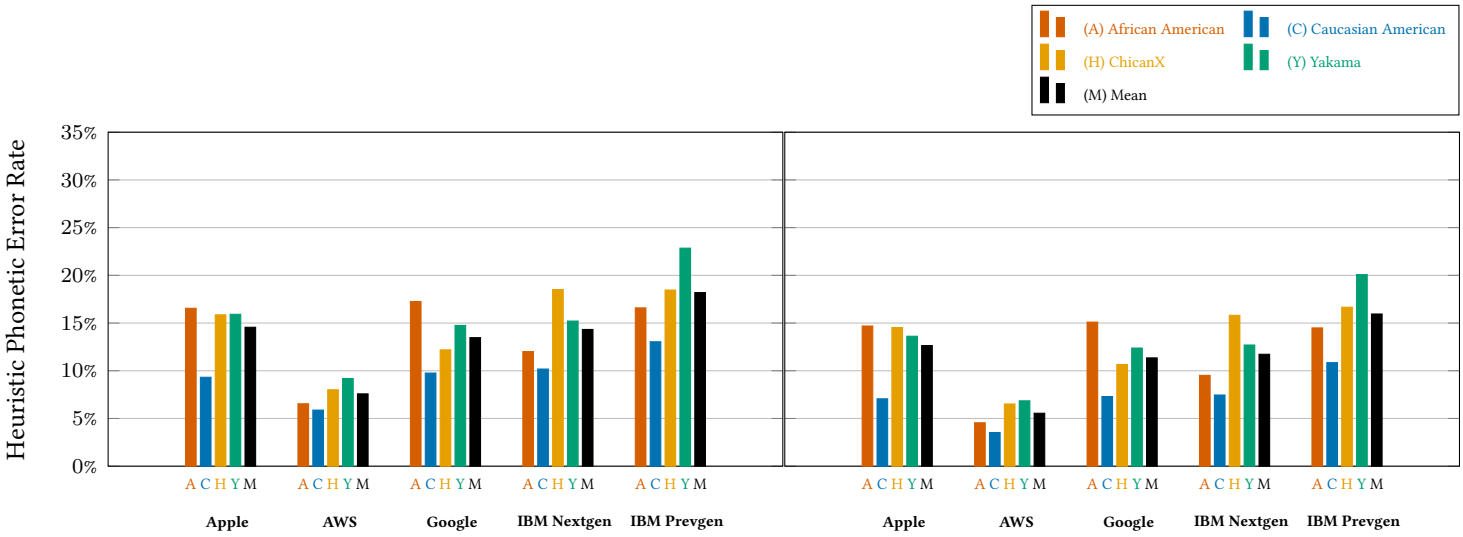
Figure 4.2: Word Error Rates (WER) using PNWEdict, by task for each transcription system

IBM Prevgen. African American speakers experienced the highest WER for Google (26%), but the second lowest for both IBM models (21% and 26% respectively), and roughly equal with ChicanX

and Yakama speakers for Apple (24%). Across all ethnic groups, AWS exhibited significantly lower WERs than the other systems evaluated.

As discussed in Chapter 3, I used pronouncing dictionaries to create a homophone dictionary, in order to ignore errors caused by homophony, e.g., "right" vs. "write." Figure 4.1a represents word error rates (WER) where homophones were determined using "canonical" pronunciations from CMUdict (Carnegie Mellon University, 2015). Figure 4.1b represents WER where homophones were determined using pronunciations from a superset of CMUdict generated for Wassink et al. (2022), which I have designated as "PNWEdict." Both figures contain five clusters of bar graphs, where each bar in a graph represents the WER for a specific ethnic group, color-coded and denoted by a letter. The purpose of the PNWEdict was to cover both the lexical gaps and pronunciation gaps in CMUdict for words present in the PNWE speech recordings, and in some cases, alternate pronunciations were removed for PNWEdict. From these figures, it is clear that using PNWEdict to determine homophones did not have a significant impact on the WER, suggesting that the data evaluated for this work did not have a substantial number of pronunciations not found in CMUdict but included in PNWEdict. This comports with the fact that the RP and WL passages contained only common words and therfore indicates that evaluations using the conversational speech data may perhaps yield a more substational difference in results between the two dictionaries.

Further, Figure 4.2 represents WER differentiated by the Reading Passage (RP) and Word-List (WL) tasks. There is a clear trend of higher error rates for the WL task, suggesting that these ASR systems benefit greatly from semantic context. Here, I did not include a figure demonstrating the difference between using CMUdict and PNWEdict for determining homophones, simply because the differences are negligible–less than 0.01% in all cases. Therefore, these results come from using the more robust PNWEdict. All subsequent figures in this chapter which do not include the CMUdict vs. PNWEdict side-by-side comparison can be assumed to have used PNWEdict as the source for canonical pronunciations.



(a) CMUdict-derived hypothesis phones

(b) PNWEdict-derived hypothesis phones

Figure 4.3: Phonetic Error Rates (PER) with heuristically determined hypothesis phones using PNWEdict, by ethnicity for each transcription system

4.2 Heuristically Determined Phonetic Error Rates (PER)

As discussed in Chapter 3, I determined a stand-in for PER by comparing the phones uttered by speakers and narrowly transcribed in the accompanying Praat TextGrids with the phonemes from the "canonical" pronunciations of the hypothesis words emitted by each ASR system. Figure 4.3 details this "heuristically determined PER", as I have chosen to call it. As with Figure 4.1, Figure 4.3a represents PER where I determined the hypothesis phones using "canonical" pronunciations from CMUdict, and Figure 4.3b represents PER where I determined the hypothesis phones using PNWEdict. Additionally, Figure 4.4 represents PER differentiated by the RP and WL tasks, in the same fashion as Figure 4.2.

It is important to note immediately that for heuristically-determined PER, there is a substantial difference between the results for CMUdict and the results for PNWEdict. On average, across all systems and ethnic groups, there was a 2 - 3% reduction in PER. This is most likely accounted for

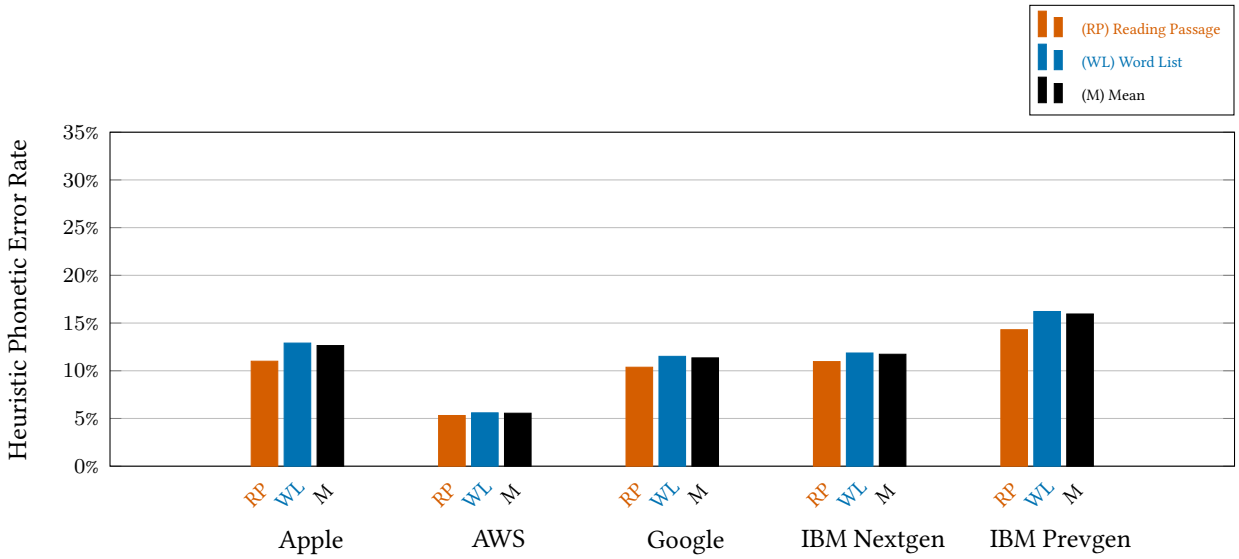


Figure 4.4: Phonetic Error Rates (PER) with heuristically determined hypothesis phones, by task for each transcription system

by the enrichment of PNWEdict to account for pronunciation gaps in CMUdict. In light of this, the following explanation of the heuristically-determined PER results focuses exclusively on the PNWEdict-derived results.

Across the board, all PERs are proportional to the WERs and reduced by roughly 6% - 14%, on average. The most notable outliers occurred for Google, at 8% reduction for Caucasian American speakers, and for the IBM Nextgen model, at 19% reduction for for ChicanX speakers. While it is difficult to speculate on the causes of these outliers, the standard deviation across all systems and ethnic groups is about 3.5% , and much lower for each system individually except for the IBM Nextgen model: 1.1% for Apple, 1.1% for AWS, 1% for Google, 3.4% for IBM Nextgen, and 1.2% for IBM Prevgen; IBM Nextgen accounts for almost all of the deviation.

IBM Watson speech-to-text tends to a output higher degree of out-of-vocabulary (OOV) words in its transcriptions, many of which are nonsense words, suggesting that it may have a higher reliance on phonetics than the other systems. Given that IBM Nextgen performs worst for the ChicanX speakers by a substantial margin in terms of WER, I'm inclined to attribute this anomaly

	# Speakers	(Dv)	(Db)	(CC)	(TH-f) / (TH-s)	(IN)	(-AO)	(prel-o)	(prel-i)	(AY)
African American	3	342	2744	246	329	381	56	180	173	2397
Caucasian American	4	384	2990	292	365	408	59	187	180	2588
ChicanX	4	502	4008	308	385	411	74	206	170	3422
Yakama	5	540	4246	341	432	465	84	229	182	3637

Table 4.1: Raw distribution of possible contexts for targeted markers in reference transcriptions, by speaker ethnicity

	# Speakers	(-AO)		(CC)		(Db)		(IN)	
		Total	**Mean**	**Total**	**Mean**	**Total**	**Mean**	**Total**	**Mean**
African American	3	13	4.33	24	8	0	0	16	5.33
Caucasian American	4	16	4	35	8.75	1	0.25	14	3.5
ChicanX	4	7	1.75	21	5.25	0	0	12	3
Yakama	5	0	0	28	5.6	1	0.2	15	3

Table 4.2: Raw distribution of realized targeted markers, by speaker ethnicity

to that difference. For whatever reason, it seems that, for the ChicanX speakers, the IBM Nextgen model's phonetic model has more influence than the language model.

Regardless, the relative differences between error rates for each ethnic group remain the same. Overall, the otherwise consistent reduction in error rate for these results is consistent with my expectations, given that the phoneme-level error rate determination method would eliminate word boundary errors, e.g., "into day" vs. "in today" and would give "partial credit" for nearly correct transcriptions, e.g., "record" vs. "recording."

4.3 Error Rates With Target Sociophonetic Markers

Table 4.1 summarizes the total number of contexts in which each target sociophonetic marker *could* have occurred, by each speaker ethnic group, based on the reading passage and word list reference orthographic transcriptions from the Praat TextGrids. This breakdown is important because, while each speaker read the same elicitation script for the reading passage task, there were

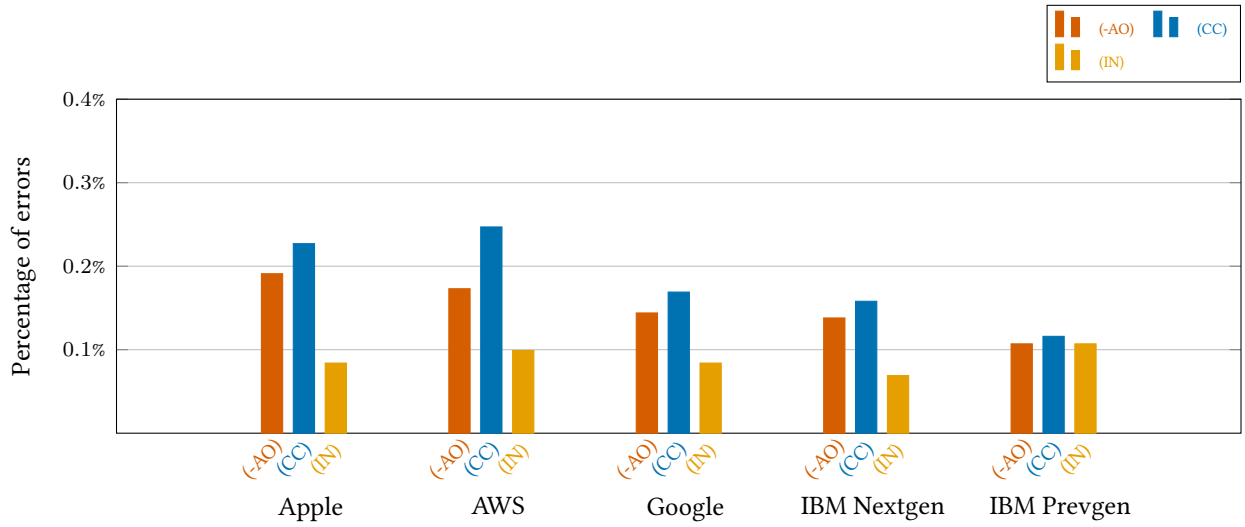


Figure 4.5: Percentage of errors containing target markers, for each transcription system

two possible elicitations for the word list task. The African American and Caucasian American speakers each read one set of word list elicitations, while the ChicanX and Yakama speakers each read another set of elicitations, different from the former set. Further, there were some discrepancies in what each speaker actually read, such as one Yakama speaker who read only 150 or so words from the roughly 750 words in the elicitation. These counts are otherwise similar in most respects to the counts from the original elicitation scripts.

Conversely, Table 4.2 summarizes the number of *realized* sociophonetic markers, by speaker ethnic group. Moreover, the table includes mean values for each marker, i.e., the relative proportions of markers within each group, to account for skew caused by differing numbers of speakers. From this data, African American speakers realized (IN) significantly more than speakers of other ethnic groups at 5.33 instances per speaker, compared to 3.5 for Caucasian American speakers, 3 for ChicanX speakers, and 3 for Yakama speakers. Similarly, African American speakers realized (-AO) slightly more than Caucasian American speakers (4.33 vs. 4) and significantly more than ChicanX speakers (4.33 vs 1.75). Yakama speakers did not realize (-AO) at all. African American speakers also realized (CC) slightly less than Caucasian American speakers, at 8 in-

	Apple				AWS				Google			
	(-AO)	(CC)	(IN)	*Tot. Err.*	(-AO)	(CC)	(IN)	*Tot. Err.*	(-AO)	(CC)	(IN)	*Tot. Err.*
African American	6	3	4	1854	1	0	2	681	5	2	4	2075
Caucasian American	5	4	1	1218	0	1	0	593	1	5	1	1340
ChicanX	5	4	0	2533	6	0	0	1361	6	0	0	2267
Yakama	0	8	2	2753	0	9	2	1419	0	7	2	2623

	IBM Nextgen				IBM Prevgen			
	(-AO)	(CC)	(IN)	*Tot. Err.*	(-AO)	(CC)	(IN)	*Tot. Err.*
African American	6	1	2	1695	3	3	5	2096
Caucasian American	3	2	3	1800	2	2	4	1882
ChicanX	5	6	0	3889	7	1	0	3165
Yakama	0	7	2	2735	0	7	3	4077

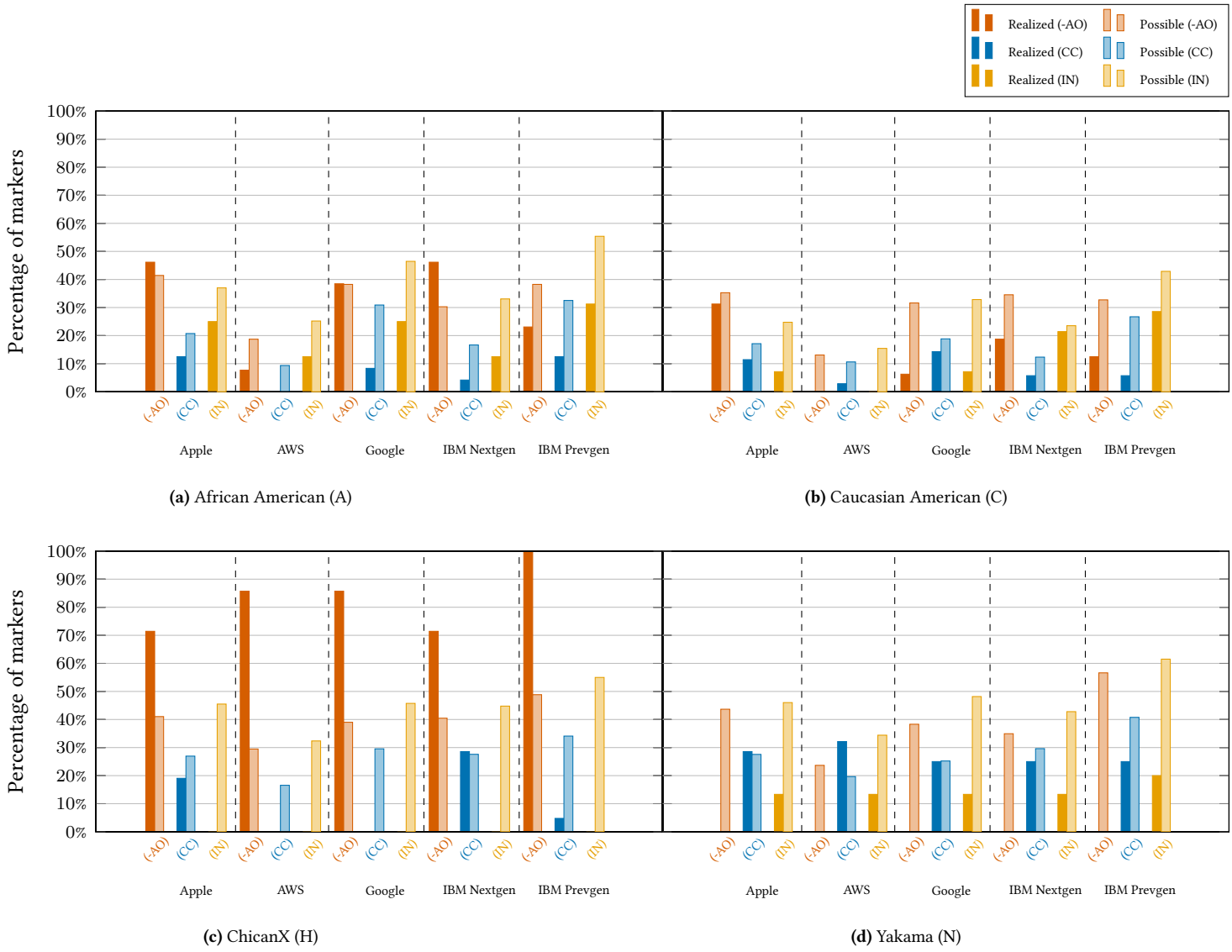
Table 4.3: Raw distribution of realized targeted markers coinciding with transcription errors, by speaker ethnicity

stances per speaker vs. 8.75, but significantly more than ChicanX and Yakama speakers, at 5.25 and 5.6 respectively. Only two speakers, who identified as Caucasian American and Yakama, respectively, realized the (Db) marker, and even then only once each. None of the speakers realized any of the other targeted markers.

Figure 4.5 represents the percentage of total errors which correlate with the target sociophonetic markers, for each system, across all speakers. All three of the realized markers that co-occurred with word errors did so roughly equally across all systems. From this, it is clear that (CC) most commonly co-occurred with word errors with a mean of $0.32 \pm 0.04\%$, followed by (-AO), with a mean of $0.15 \pm 0.03\%$. Finally, (IN) co-occurred with word errors with a mean of $0.09 \pm 0.01\%$.

4.3.1 Marker and Transcription Error Co-occurrences

Finally, Table 4.3 summarizes the raw distribution of these realized markers as they co-occurred with word errors from the ASR system transcriptions, split by speaker ethnic group. Similarly, Figure 4.6 summarizes this data proportionally and is summarized in detail in the next few subsections.



(a) African American (A)

(b) Caucasian American (C)

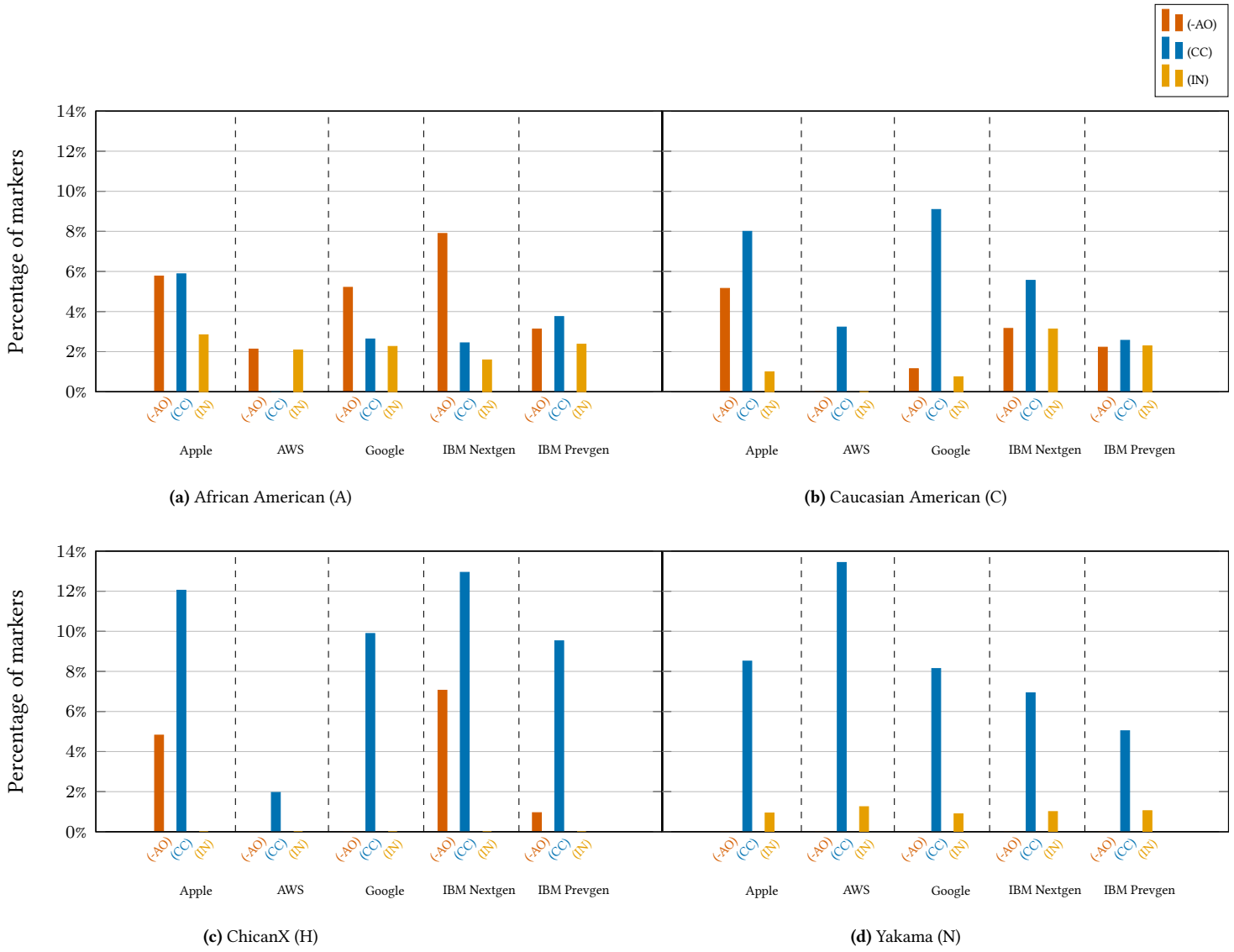
(c) ChicanX (H)

(d) Yakama (N)

Figure 4.6: Percentage of targeted markers co-occurring with errors, by ethnicity for each transcription system Solid bars represent realized markers and lighter bars represent possible contexts for markers

(-AO) occurrences

From the data in Figure 4.6 , it is clear that (-AO) more frequently co-occurred with errors for the African American speakers than for the Caucasian American speakers across all systems: 46%



(a) African American (A)

(b) Caucasian American (C)

(c) ChicanX (H)

(d) Yakama (N)

Figure 4.7: Normalized percentages of targeted markers co-occurring with errors, by ethnicity for each transcription system

vs. 31% for Apple, 8% vs. 0% for AWS, 38% vs. 6% for Google, 46% vs. 19% for the IBM Nextgen model, and 23% vs. 13% for the IBM Prevgen model. Further, the African American speakers had significantly more possible contexts for (-AO) that co-occurred with errors than the Caucasian American speakers across all systems: 41% vs. 17%, 19% vs. 11% for AWS, 38% vs. 19% for Google,

30% vs. 12% for IBM Nextgen, and 38% vs. 26% for IBM Prevgen.

To help contextualize this disparity, I can determine for each ethnic group the percentage of possible contexts that co-occur with transcription errors and are also actually realized markers. After this normalization, the African American speakers still demonstrate a substantially higher association of (-AO) with transcription errors for all systems: 6% vs. 5% for Apple, 2% vs. 0% for AWS, 5% vs. 1% for Google, 8% vs. 3% for IBM Nextgen, and 3% vs. 2% for IBM Prevgen.

Therefore, it is reasonable to conclude that overall, (-AO) is more strongly associated with transcription errors for the African American Speakers than for the Caucasian American speakers.

However, (-AO) co-occurred with errors significantly less frequently for the African American speakers than for the ChicanX speakers, again across all systems: 46% vs. 71% for Apple, 8% vs. 86% for AWS, 38% vs. 86% for Google, 46% vs. 71% for the IBM Nextgen model, and 23% vs. 100% for the IBM Prevgen model. Possible contexts for (-AO) that co-occurred with errors were generally higher for the ChicanX speakers than for the African American speakers for AWS and both IBM models, by 11%, 10%, and 11% respectively, and roughly equal for the Apple and Google systems. Similarly, possible contexts for (-AO) that co-occurred with errors were generally higher for the ChicanX speakers than for the Caucasian American speakers across all systems: 6% for Apple, 16% for AWS, 7% for Google, 6% for IBM Nextgen, and 16% for IBM Prevgen.

After normalization using the same technique described above, the African American speakers experienced slightly more co-occurrences than the ChicanX speakers across all systems except AWS and the IBM Prevgen Model: 6% vs. 4% for Apple, 2% vs. 6% for AWS, 5% vs. 4% for Google, 8% vs. 4% for IBM Nextgen, and 3% vs. 4% for IBM Prevgen.

After accounting for the disparity in possible contexts, it appears that even though the African American speakers realized the (-AO) marker more than twice as often as the ChicanX speakers, the marker is more strongly associated with transcription errors for the ChicanX speakers for two systems, and only slightly less strongly associated for three systems. Given the relatively low counts, it seems reasonable to conclude that the marker is roughly equally associated with transcription errors for both of these speaker ethnic groups.

Yakama speakers showed no realizations for (-AO) at all.

(CC) occurrences

(CC) co-occurred with transcription errors more frequently for the African American speakers than for the Caucasian American speakers for Apple and the IBM Prevgen model, but less frequently for all other systems: 13% vs. 11% for Apple, 0% vs. 3% for AWS, 8% vs. 14% for Google, 4% vs. 6% for IBM Nextgen, and 13% vs. 6% for IBM Prevgen. The African American speakers had generally higher possible numbers of contexts for (CC) that co-occurred with errors as the Caucasian American speakers, across all systems except AWS: 21% vs. 17% for Apple, 9% vs. 11% for AWS, 31% vs. 19% for Google, 17% vs. 12% for IBM Nextgen, and 33% vs. 27% for IBM Prevgen.

After normalizing the results, the African American speakers still demonstrated a significantly lower percentage of realized markers co-occurring with transcription errors than the Caucasian American speakers, across all systems except the IBM Prevgen model: 6% vs. 8% for Apple, 0% vs. 3% for AWS, 3% vs. 9% for Google, 2% vs. 6% for IBM Nextgen, and 4% vs 3% for IBM Prevgen.

This suggests that, overall, (CC) slightly more strongly associates with errors for the Caucasian American speakers than for the African American speakers.

Similarly, (CC) co-occurred with transcription errors less frequently for the African American speakers than for the ChicanX speakers across all systems except for Google and the IBM Prevgen model: 13% vs. 19% for Apple, 0% vs. 0% for AWS, 8% vs. 0% for Google, 4% vs. 29% for IBM Nextgen, and 13% vs. 5% for IBM Prevgen. The African American speakers had fewer possible contexts for (CC) that co-occurred with errors than the ChicanX speakers had, across all systems except Google: 21% vs. 27% for Apple, 9% vs. 16% for AWS, 31% vs. 30% for Google, 17% vs. 28% for IBM Nextgen, and 33% vs. 34% for IBM Prevgen.

After normalization, the African American speakers demonstrate a slightly higher percentage of realized markers co-occurring with transcription errors across all systems than the ChicanX speakers, across all systems except for the IBM Nextgen model: 6% vs. 5% for Apple, 0% vs. 0% for AWS, 3% vs. 0% for Google, 2% vs. 7% for IBM Nextgen, and 4% vs. 1% for IBM Prevgen.

This suggests that (CC) is associated with transcription errors more strongly for the African

American speakers than for the ChicanX speakers.

Finally, (CC) co-occurred with transcription errors substantially less frequently for the African American speakers than for the Yakama speakers, across all systems: 13% vs. 29% for Apple, 0% vs. 24% for AWS, 8% vs. 27% for Google, 4% vs. 24% for IBM Nextgen, and 13% vs. 32% for IBM Prevgen. However, the African American speakers had significantly fewer possible contexts for the marker that co-occurred with errors than the Yakama speakers did, across all systems: 21% vs. 46% for Apple, 9% vs. 34% for AWS, 31% vs. 48% for Google, 17% vs. 43% for IBM Nextgen, and 33% vs. 61% for IBM Prevgen.

After normalizing the results, the African American speakers still demonstrate substantially lower percentages of realized markers co-occuring with transcription errors, across all systems: 6% vs. 9% for Apple, 0% vs. 13% for AWS, 3% vs. 8% for Google, 2% vs. 7% for IBM Nextgen, 4% vs. 5% for IBM Prevgen.

This clearly demonstrates that (CC) is associated with transcription errors much more strongly for the Yakama speakers than for the African American speakers.

(IN) occurrences

(IN) co-occurred with errors significantly more frequently for the African American speakers than for the Caucasian American speakers across all systems except the IBM Nextgen model: 25% vs 7% for Apple, 13% for 0% for AWS, 25% vs 7% for Google, 13% vs. 21% for IBM Nextgen, and 31% vs. 29% for IBM Prevgen. However, it is important to note that the African American speakers had significantly more possible contexts for (IN) that co-occurred with errors: 37% vs. 25% for Apple, 25% vs. 16% for AWS, 46% vs. 33% for Google, 33% vs. 24% for IBM Nextgen, and 55% vs. 43% for IBM Prevgen.

After normalizing the results, the African American speakers demonstrate slightly higher percentages of realized markers co-occurring with transcription errors for all systems except the IBM models: 3% vs. 1% for Apple, 2% vs. 0% for AWS, 2% vs. 1% for Google, 2% vs. 3% for IBM Nextgen, and 2% vs. 2% for IBM Prevgen.

It seems reasonable to conclude that (IN) is more strongly associated with errors for African

American speakers than for Caucasian American speakers.

Similarly, (IN) co-occurred with errors more frequently for the African American speakers than for the ChicanX speakers, who had no co-occurrences across all systems. Of the four ChicanX speakers, only one made an utterance demonstrating (IN), suggesting that this is not a commonly realized marker among the ChicanX speakers interviewed.

Finally, (IN) co-occurred with errors more frequently for the African American speakers than for the Yakama speakers for Apple, Google, and the IBM Prevgen model (25% vs. 13% for Apple, 25% vs. 13% for Google, and 31% vs. 20% for IBM Prevgen), but slightly less frequently for AWS and for the IBM Nextgen model (12.5% vs. 13% for both). However, the African American speakers had substantially fewer possible contexts for (IN) that co-occurred with errors than for the Yakama speakers across all systems: 37% vs. 46% for Apple, 25% vs. 34% for AWS, 47% vs. 48% for Google, 33% vs. 43% for IBM Nextgen, and 55% vs. 62% for IBM Prevgen.

When accounting for this discrepancy and normalizing the realized results, the African American speakers demonstrate slightly more co-occurrences than the Yakama speakers across all systems: 3% vs. 1% for Apple, 2% vs. 1% for AWS, 2% vs. 1% for Google, 2% vs. 1% for IBM Nextgen, and 2% vs. 1% for IBM Prevgen.

These results are consistent with my intuition, because the African American speakers realized the marker substantially more per speaker than the Yakama speakers (5.33 instances per speaker vs. 3 instances per speaker). In short, this demonstrates that the marker is equally problematic for both groups across all systems, but more frequently realized by the the African Americans speakers than the the Yakama speakers.

It is also important to note that, given the relatively high number of possible marker contexts that coincide with transcription errors, this marker may be particularly problematic for ASR systems in general.

(Db) occurrences

There were only two realized occurrences of (Db), which were each uttered by one Caucasian American speaker and one Yakama speaker. Neither utterance co-occurred with word transcrip-

tion errors.

4.4 Summary

None of the four realized markers co-occurred with errors most frequently for the Caucasian American speakers, as expected, and each marker co-occurred with errors more frequently for the African American speakers than for the Caucasian American speakers. This is in line with the results from Wassink et al. (2022).

Moreover, (-AO) appears to be associated with errors most strongly for the African American speakers, followed closely by the ChicanX speakers. (CC) also appears to be associated with errors most strongly for the Yakama speakers, followed by the Caucasian American speakers. For the African American speakers (CC) appears to be slightly more associated with errors than for the ChicanX speakers. Finally, (IN) appears to be associated with errors slightly more strongly for the African American speakers than for all the other speaker ethnic groups, followed closely by the Caucasian American speakers. In the next chapter, I discuss these results and how they apply to the answers of my two research questions.

Chapter 5

DISCUSSION

The results outlined in Chapter 4 cover three of the ten sociolinguistic variables outlined in Section 1.2.1 and demonstrate that those three markers are strongly associated with orthographic transcription errors across all systems analyzed, particularly for the African American speakers. While the number of instances analyzed in this work were relatively few, there are some noticeable correlations worth discussing, all of which point to extremely promising results in future work with the expanded corpus including conversational speech recordings.

5.1 Declaration of Competing Interest

It is important to disclose that at the time of writing, I am an employee of Amazon.com, Inc. However, I have never worked on Amazon Transcribe nor have I worked on any Amazon Web Services (AWS) team, and I have no knowledge of how Transcribe works. Moreover, I have no incentives, financial or otherwise, to misrepresent the results in this work. The opinions expressed in this work are my own and do not reflect the opinions of Amazon.com, Inc.

5.2 RQ1: Do AAL markers correlate with disproportionately higher word error rates for African American speakers?

For the speaker recordings analyzed in this work, three AAL markers co-occurred with orthographic transcription errors across all four systems analyzed:

1. resistance to the low-back vowel merger, (-AO),
2. the pre-nasal merger of /ɛ/ and /ɪ/, otherwise known as the *pen-pin* merger, and (IN), and
3. consonant cluster reduction, (CC).

While these markers each, on average, accounted for only 0.04%, 0.03%, and 0.02%, respec-

tively, of the orthographic transcription errors for the African American speakers, each instance of the markers co-occurred with errors 32%, 21%, and 7.5% of the time, respectively. This suggests that, of the targeted markers outlined in Chapter 1, (-AO) and (IN) likely correlate with disproportionately higher error rates for the African American speakers analyzed.

5.3 RQ2: Are there phonological features representative of AAL which appear more frequently on poorly transcribed speech for African American speakers than for other co-regional speakers?

None of the four realized markers co-occurred with errors most frequently for the Caucasian American speakers, as expected, and each marker co-occurred with errors more frequently for the African American speakers than for the Caucasian American speakers. This is in line with the results from Wassink et al. (2022).

Moreover, for the speaker recordings analyzed in this work, two targeted AAL markers were more strongly associated with transcription errors for the African American speakers than for the three other speaker ethnic groups analyzed:

1. resistance to the low-back vowel merger, (-AO) and
2. the pre-nasal merger of /ɛ/ and /ɪ/, otherwise known as the *pen-pin* merger, (IN).

A third targeted marker, consonant cluster reduction, (CC), was more strongly associated for the African American speakers than for the ChicanX speakers, but not more strongly than for the Caucasian American or Yakama speakers.

5.4 Discussion of Individual Errors

While any analysis of individual transcription errors is inherently speculative, it is nonetheless a useful exercise in order to gain some insight into the potential flaws of these ASR systems. Of the orthographic transcription errors which co-occurred with the targeted markers, there are three types which recurred and are therefore interesting to discuss. I have reproduced the full list of transcription errors in Appendix A.

The first type of error co-occurred with (-AO), in which CAUGHT was transcribed as CUT

or CARTE. This error type occurred for Speaker 1, who is African American, by Apple, Google, and both IBM Watson models, for Speaker 2, who is African American, by the IBM Nextgen model, for Speaker 3, who is African American by Apple, and for Speaker 4, who is Caucasian American, by Apple and IBM Prevgen. In all of these examples, the speaker uttered /kɔt/ and the system transcribed it as CUT, which has a pronunciation of /kʌt/ in PNWEdict, or CARTE, which has a pronunciation of /kɑɹt/ in PNWEdict. From these examples, it seems possible that the /ɔ/ pronunciation may have contributed to the error, particularly when considering the fact that most of these systems likely use CTC models, which emit graphemes from input waveforms, as discussed in Chapter 2. Given that these models can exhibit unusual behavior for digraphs and trigraphs, and given the fact that /kɑt/ is the more common pronunciation of CAUGHT in American English, it stands to reason that the presence of (-AO) may have directly contributed to these transcription errors, most of which occurred for the African American speakers.

The second type of error co-occurred with (IN), in which WHEN was transcribed as AND for Speaker 1, who is African American, by AWS, Google, and IBM Nextgen and in which WHEN was transcribed as WIND for Speaker 12, who is Yakama, by AWS. It is interesting that in both cases, words were chosen which end with /d/ and in both cases, the speakers uttered /wɪn/.

The third type of error co-occurred with (CC), in which NEXT was transcribed as MIX for Speaker 2, who is African American, by Apple, Google, and both IBM Watson models and in which OFTEN was transcribed variously as OFF AND, OFF OF, OFF ON, and OFF IN for Speaker 6, who is Caucasian American, by Apple and Google and for Speaker 15, who is Yakama, by Apple, AWS, and Google. It seems fairly obvious in all of these examples that the absence of the /t/ from the /st/ and /ft/ clusters is what accounts for the transcription errors, given that the results are often very close to what one would expect when the /t/ sounds are omitted.

5.5 Limitations

The most obvious limitation to discuss is the type of data examined. The RP and WL tasks both consist of inherently attended speech. I was unable to use the lexical games (LEX) and conversational speech (CS) data because they lacked, at time of writing, Praat TextGrids with time-aligned

narrow phonetic transcriptions and orthographic transcriptions. There are, technically, orthographic transcriptions available for the LEX and CS data, but they are not time-aligned and do not include narrow phonetic descriptions, and I decided that the effort to analyze these transcripts was too great for this work. The extraction and analysis methods detailed in this work are still applicable to that data, and I suspect would be very successful for future work. I speak more on how I would approach this in Section 6.2.

As a result of this decision, very few of the targeted sociophonetic markers were present. Given the results in Wassink et al. (2022), it's clear that most of the linguistic variables are present in the LEX and CS tasks, and so applying the methods in this work to those recordings would probably yield more interesting results.

Another limitation worth mentioning is the structure of the WL task. As discussed in Section 2.1, most ASR systems rely on some sort of language model to correct the outputs of their acoustic models. Those language models typically benefit from context and attempt to reduce perplexity to get an optimal emission. The lack of linguistic context in a list of words within the carrier phrase "write ____ today" could possibly have negatively impacted orthographic transcription results in terms of the influence of the language models, meaning that the acoustic models would have much more influence on the graphemes emitted. The results in Figure 4.2 seem to corroborate this claim.

5.6 Ethical Considerations

5.6.1 Data Collection

I did not perform the data collection. The PNWE Corpus was recorded and annotated for the NSF grant-funded project Pacific Northwest English Study (BCS-0643374, BCS-1147678). All participants provided informed consent to the recordings and were compensated for their time (Wassink, 2015). Further, I am contributing this work and any future derived work to the research goals of Prof. Wassink's Bias in ASR research group.

5.6.2 Cost of Research

None of the work herein was cost-prohibitive for me, personally, and was self-funded to a total of roughly $40 USD. While funding was technically available to me through the University of Washington, I decided that the bureaucratic hurdles to overcome in order to acquire the funds were not worth my time. AWS, Google Cloud, and IBM Cloud each required the creation of a commercial account to access their APIs, which have fee-based usage, but also have some form of free-tier service or new account credits (Amazon, 2022; Google, 2022; IBM, 2022). Apple's STT service is entirely free, but requires the creation of an iOS or macOS application and has a limited number of requests (Apple, 2022). However, I transcribed only about four hours of speech data across these services. Had I included the conversational speech and lexical game data from the PNWE corpus, I would have added an additional nine hours of speech data, more than tripling the cost.

Chapter 6

CONCLUSION

6.1 Summary

In this thesis, I analyzed the transcription results of four major commercial ASR systems—Apple Speech, Amazon Transcribe, Google Speech-to-text, and IBM Watson Speech-to-text—on recordings from the Pacific Northwest English (PNWE) corpus originally collected for Wassink (2015). I was motivated to do this analysis primarily by two seminal papers which attempted to apply sociolinguistic knowledge to the evaluation of ASR systems:

1. Koenecke et al. (2020), which evaluates five major commercial ASR systems on word error rate (WER) and found higher error rates for speakers who identify as African American than for speakers who identify as Caucasian American and also hypothesized that these increased errors were likely due to phonetic markers and not syntactic or semantic markers; and
2. Wassink et al. (2022), which evaluates one major commercial ASR systems on speaker in the Pacific Northwest from four different ethnic backgrounds—the PNWE corpus—and found that specific, targeted sociophonetic markers strongly correlate with increased errors for non-Caucasian American speakers.

I attempted to answer two research questions:

1. Do sociophonetic markers typical of African American Language (AAL) correlate with higher inaccuracy rates in major commercial ASR systems for African American speakers than for speakers of different ethnic backgrounds?
2. Are there any phonological features representative of AAL which appear more frequently on incorrectly transcribed speech for African American speakers than for other co-regional

speakers?

To do this, I ran automatic transcription on recordings of 16 speakers from four ethnic backgrounds—African American, Caucasian American, ChicanX, and Yakama—for all four ASR systems evaluated. The PNWE corpus includes orthographic and phonetic transcriptions, time-aligned as Praat TextGrids, which I used in conjunction with `sclite` (NIST, 2021) to evaluate WER. I also generated an heuristically determined phonetic error rate (PER) using a custom superset of CMUdict (Carnegie Mellon University, 2015) created specifically for the PNWE corpus, in order to further speculate on the potential effects of acoustic models on errors. Finally, I identified ten target linguistic variables which represent common sociophonetic markers of African American Language (AAL) and wrote custom software to identify possible contexts where those markers could occur as well as contexts where they did occur. I added these contexts to the Praat TextGrids and used these to identify co-occurrences of markers with transcription errors for each ASR system and performed both a quantitative and speculative analysis.

From this, I determined that the resistance to the low-back merger, `(-AO)`, and the prenasal front merger (pen-pin merger), `(IN)`, were both most strongly associated with errors for the African American speakers than for any other ethnic group, and that consonant cluster reduction, `(CC)`, was more strongly associated with errors for the Yakama and Caucasian American speakers than for the African American speakers.

6.2 Future Work

Perhaps the most promising avenue for future work is to reproduce the experiments outlined above using the conversational speech (CS) data from the PNWE corpus. Ultimately, it would be best to manually make narrow phonetic transcriptions for each CS recording, in order to faithfully reproduce the experiments, though such an endeavor would be labor and time intensive. Regardless, I believe it is possible to closely approximate the experiments with the CS recordings and orthographic transcriptions as they currently exist.

At the time of writing, not every CS recording has a Praat TextGrid with phonetic or ortho-

graphic transcriptions. However, there is a spreadsheet with manual corrections to orthographic transcriptions produced by CLOx, the system evaluated in Wassink et al. (2022). These transcriptions and corrections are time-aligned by sentences and also annotated with speaker diarization using rich text. Therefore, one could use these formatted transcriptions to generate ground truth reference transcripts for each speaker in each conversation and then calculate per-speaker WER with `sclite`.

Because the CS recordings currently lack narrow phonetic transcriptions, one would not be able to accurately reproduce the heuristically determined PER methods outlined in Section 3.2, but one could easily generate the correlative analysis of sociophonetic markers with transcription errors, which I believe would generate extremely useful and interesting results for very little effort.

Moreover, further analysis with different datasets and other sociolects will likely provide valuable insight into the impact of sociophonetic markers on transcription error rates and suggest how sociophonetic knowledge might be applied to mitigate accuracy issues.

Finally, as this work is only evaluative, it will be important for the future development of ASR systems and models to incorporate and apply sociolinguistic understanding in order to reduce the gap in transcription error rates. It stands to reason that, because sociophonetic features are strongly associated with transcription errors, the application of this knowledge will inform how we can reduce racial bias in ASR systems.

BIBLIOGRAPHY

Amazon. (2022). Amazon Transcribe [Computer Software]. https://aws.amazon.com/transcribe/

Apple. (2022). Apple Speech API [Computer Software]. https://developer.apple.com/documentation/speech

BBC Scotland. (2011, June). Burnistoun, series 1, episode 1, voice activated elevator. https://www.bbc.co.uk/programmes/p00hbfjw

Boersma, P., & Weenink, D. (2022, July). Praat: Doing phonetics by computer [Computer Program]. http://www.praat.org/

Carnegie Mellon University. (2015). Carnegie Mellon University pronouncing dictionary (CMUdict) [Computer Software]. https://github.com/Alexir/CMUdict/blob/master/cmudict-0.7b

Chambers, J. K. (2009). *Sociolinguistic theory : Linguistic variation and its social significance* (Rev. ed.). Blackwell.

Di Paolo, M., & Yaeger-Dror, M. (Eds.). (2011). Sociophonetics : A student's guide.

Dorn, R. (2019). Dialect-specific models for automatic speech recognition of African American Vernacular English. *Proceedings of the Student Research Workshop Associated with RANLP 2019*. https://doi.org/10.26615/issn.2603-2821.2019_003

Dupras, V. (2019). Num2words library - convert numbers to words in multiple languages [Computer Software] (Savoir-faire Linux, Ed.). https://github.com/savoirfairelinux/num2words

Eddington, D., & Taylor, M. (2009). T-Glottalization in American English. *American Speech*, *84*(3), 298–314. https://doi.org/10.1215/00031283-2009-023

Farrington, C. (2018). Incomplete neutralization in African American English: The case of final consonant voicing. *Language Variation and Change*, *30*(3), 361–383. https://doi.org/10.1017/S0954394518000145

Google. (2022). Google Cloud Speech-to-Text [Computer Software]. https://cloud.google.com/speech-to-text

Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (pp. 1764–1772). PMLR. https://proceedings.mlr.press/v32/graves14.html

Green, L. J. (2002). *African American English: A linguistic introduction* (Paperback). Cambridge University Press.

Hall-Lew, L. (2013). 'Flip-flop' and mergers-in-progress. *English Language and Linguistics*, *17*(2), 359–390. https://doi.org/10.1017/s1360674313000063

hyperreality. (2022). American British English translator [Computer Software]. https://github.com/hyperreality/American-British-English-Translator

IBM. (2022). Watson Speech to Text [Computer Software]. https://www.ibm.com/cloud/watson-speech-to-text

Kendall, T., McLarty, J., & Josler, B. (2018). ORAAL: Online resources for African American Language: AAL facts. https://oraal.uoregon.edu/facts

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*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Labov, W. (1970). The study of language in its social context. *Studium Generale*, *23*, 30.

Labov, W. (1972). Language in the inner city: Studies in the Black English vernacular. U of Pennsylvania P.

Ladefoged, P. (2001). *Vowels and consonants : An introduction to the sounds of languages*. Blackwell.

Lamere, P., Kwok, P., Walker, W., Gouvêa, E., Singh, R., Raj, B., & Wolf, P. (2003). Design of the cmu sphinx-4 decoder. https://doi.org/10.21437/Eurospeech.2003-382

Lee, K.-f. (1990). Context-dependent phonetic hidden markov-models for speaker-independent continuous speech recognition. *IEEE transactions on acoustics, speech, and signal processing*, *38*(4), 599–609.

Mahrt, T. (2016). PraatIO [Computer Software]. https://github.com/timmahrt/praatIO

Morris, A., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. https://doi.org/10.21437/Interspeech.2004-668

NIST. (2021). SCTK, the NIST scoring toolkit [Computer Software]. https://github.com/usnistgov/SCTK

Paul, S. (2017, March). Voice is the next big platform, unless you have an accent. https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Rabiner, L., & Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP magazine*, *3*(1), 4–16.

Robert, J., Webbie, M., et al. (2018). Pydub [Computer Software]. http://pydub.com/

Smitherman, G. (1998). Ebonics, King, and Oakland. *Journal of English Linguistics*, *26*(2), 97–107.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. https://doi.org/10.18653/v1/W17-1606

Tatman, R. (2020). Sociolinguistic variation and automatic speech recognition: Challenges and approaches. *Presented at the Annual Meeting of the American Academy for the Advancement of Science. Seattle.*

Tatman, R., & Kasten, C. (2017). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. *Proc. Interspeech 2017*, 934–938. https://doi.org/10.21437/Interspeech.2017-1746

Thomas, E. (2007). Phonological and phonetic characteristics of AAVE. *Language and Linguistics Compass*, *1*, 450–475. https://doi.org/10.1111/j.1749-818X.2007.00029.x

Thomas, E., & Reaser, J. (2004). Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of sociolinguistics*, *8*(1), 54–87.

U.S. Census Bureau QuickFacts: United States. (2020). https://www.census.gov/quickfacts/fact/table/US/PST045219

Wassink, A. (2015). Sociolinguistic patterns in seattle english. *Language Variation and Change*, *27*(1), 31–58. https://doi.org/10.1017/S0954394514000234

Wassink, A., Gansen, C., & Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. *Speech Communication*, *140*, 50–70. https://doi.org/10.1016/j.specom.2022.03.009

Wolfram, W. (1969). A sociolinguistic description of Detroit Negro speech. urban language series, no. 5.

Wolfram, W. (1994). The phonology of a sociocultural variety: The case of African American Vernacular English. *Child phonology: Characteristics, assessment, and intervention with special populations*, 227–244.

Appendix A

ERRORS WITH TARGETED MARKERS

Speaker ID	Ethnicity	Phones	Markers	Reference Word	Apple	AWS	Google	IBM Nextgen	IBM Prevgen
Speaker 1	African American	K AO1 T	(-AO)	CAUGHT				CUT	
Speaker 1	African American	K AO1 T	(-AO)	CAUGHT	CUT		CUT	CUT	
Speaker 1	African American	K AO1 T	(-AO)	CAUGHT	******	TALK	TALK	TALK	TALK
Speaker 1	African American	K AO1 T	(-AO)	CAUGHT				CUT	CUT
Speaker 1	African American	D IH1 D AH0 N	(CC)	DIDN'T			ins: [AND] didn't		
Speaker 1	African American	K UH1 D AH0 N	(CC)	COULDN'T	RAKING				
Speaker 1	African American	W IH1 N	(IN)	WHEN			AND		
Speaker 1	African American	W IH1 N	(IN)	WHEN		AND	AND	AND	
Speaker 1	African American	AH0 G EH1 N S T	(IN)	AGAINST	INTO				
Speaker 1	African American	AH0 G EH1 N	(IN)	AGAIN	KEEGAN	IN	CAN	GAME	KEEGAN
Speaker 2	African American	K AO1 T	(-AO)	CAUGHT	KOK				
Speaker 2	African American	K AO1 T	(-AO)	CAUGHT	KOK		BUY	CUT	
Speaker 2	African American	N EH1 K S	(CC)	NEXT	MIX		MIX	MIX	MIX
Speaker 2	African American	D IH1 D AH0 N	(CC)	DIDN'T					ins: [I] didn't
Speaker 2	African American	K UH1 D AH0 N	(CC)	COULDN'T					COULD
Speaker 2	African American	AH0 G EH1 N	(IN)	AGAIN					AS
Speaker 3	African American	K AO1 T	(-AO)	CAUGHT	******		******	COME	CHICAGO
Speaker 3	African American	K AO1 T	(-AO)	CAUGHT	CUT				
Speaker 3	African American	K AO1 T	(-AO)	CAUGHT			SCOTT		
Speaker 3	African American	K UH1 D AH0 N	(CC)	COULDN'T	QUICK				
Speaker 3	African American	HH W IH1 N	(IN)	WHEN	****				HIS
Speaker 3	African American	AH0 G EH1 N S T	(IN)	AGAINST					GUESS
Speaker 3	African American	W EH1 N	(IN)	WHEN	FOR		FOR		FOR
Speaker 4	Caucasian American	K AO1 T	(-AO)	CAUGHT					CUT
Speaker 4	Caucasian American	K AO1 T	(-AO)	CAUGHT	CUT			CARTE	
Speaker 4	Caucasian American	K AO1 T	(-AO)	CAUGHT	CUT				
Speaker 4	Caucasian American	AH0 G EY1 N	(IN)	AGAIN					*****
Speaker 5	Caucasian American	K AO1 T	(-AO)	CAUGHT	CARD				
Speaker 5	Caucasian American	K AO1 T	(-AO)	CAUGHT	CARD				
Speaker 5	Caucasian American	K UH1 D AH0 N	(CC)	COULDN'T	GOOD				
Speaker 5	Caucasian American	D IH1 D AH0 N	(CC)	DIDN'T	DON'T	DON'T	DON'T	DON'T	DON'T
Speaker 5	Caucasian American	K UH1 D AH0 N	(CC)	COULDN'T				DEMISE	
Speaker 5	Caucasian American	N EH1 K S	(CC)	NEXT					ins: [TILL] next
Speaker 5	Caucasian American	W EH1 N	(IN)	WHEN				ins: [OH K] when	ins: [HAD] when
Speaker 6	Caucasian American	K AO1 F	(-AO)	COUGH				COFFEE	
Speaker 6	Caucasian American	K AO1 T	(-AO)	CAUGHT				CARTE	
Speaker 6	Caucasian American	AO1 F AH0 N	(CC)	OFTEN			ins: [OFF] IN		
Speaker 6	Caucasian American	AO1 F AH0 N	(CC)	OFTEN	ins: [OFF] AND		ins: [OFF] IN		
Speaker 7	Caucasian American	K AO1 T	(-AO)	CAUGHT	GOT		GOT		GOT
Speaker 7	Caucasian American	W IH1 N	(IN)	WHEN					WHICH
Speaker 7	Caucasian American	W EH1 N	(IN)	WHEN	ins: [THAT] when		ins: [THAT] when	ins: [THAT] when	ins: [THAT] when
Speaker 8	ChicanX	K UH1 D AH0 N	(CC)	COULDN'T	RAKING			PANTED	
Speaker 8	ChicanX	D IH1 D AH0 N	(CC)	DIDN'T				ins: [THE RATE] didn't	
Speaker 9	ChicanX	K AO1 F	(-AO)	COUGH		*****	CLUB		BECAUSE

Speaker ID	Ethnicity	Phones	Markers	Reference Word	Apple	AWS	Google	IBM Nextgen	IBM Prevgen
Speaker 9	ChicanX	K AO1 T	(-AO)	CAUGHT	HOT		HOT	HOT	HOT
Speaker 9	ChicanX	K AO1 T	(-AO)	CAUGHT		CALL		CARTE	
Speaker 9	ChicanX	K AO1 T	(-AO)	CAUGHT	CARD	******	CLOCK	RATE	******
Speaker 9	ChicanX	K AO1 F	(-AO)	COUGH	CALL	CALL	CALL	CALL	CALL
Speaker 9	ChicanX	K AO1 T	(-AO)	CAUGHT	'CAUSE	CALL	CALL	******	******
Speaker 9	ChicanX	K AO1 F	(-AO)	COUGH	*****	CALL	CALL		ins: [OFF] TO
Speaker 9	ChicanX	K UH1 D AH0 N	(CC)	COULDN'T	********			CAN	CREAM
Speaker 9	ChicanX	K UH1 D AH0 N	(CC)	COULDN'T	********				
Speaker 10	ChicanX	D IH1 D AH0 N	(CC)	DIDN'T	DINNER			ins: [DID] IN	
Speaker 11	Yakama	D IH1 D AH0 N	(CC)	DIDN'T		******		DITTEN	INTO
Speaker 11	Yakama	K UH1 D AH0 N	(CC)	COULDN'T		CUTTING			********
Speaker 11	Yakama	K UH1 D AH0 N	(CC)	COULDN'T	CURTAIN				
Speaker 11	Yakama	K UH1 D AH0 N	(CC)	COULDN'T			********	********	
Speaker 11	Yakama	D IH1 D AH0 N	(CC)	DIDN'T	RIDING	******	IN		DID
Speaker 11	Yakama	K UH1 D AH0 N	(CC)	COULDN'T	SKIN		TO	MY	SKIN
Speaker 12	Yakama	D IH1 D AH0 N	(CC)	DIDN'T					******
Speaker 12	Yakama	K UH1 D AH0 N	(CC)	COULDN'T				********	COULD
Speaker 12	Yakama	IH1 N AH0 S AH0 N T S	(CC)	INNOCENTS		INNOCENT		INNOCENCE	
Speaker 12	Yakama	W EH1 N	(IN)	WHEN		WIND			
Speaker 13	Yakama	G R AE1 N P AA2	(CC)	GRANDPA				BUY	TO
Speaker 13	Yakama	K UH1 D AH0 N	(CC)	COULDN'T				********	
Speaker 13	Yakama	AH0 G EH1 N S T	(IN)	AGAINST					MIGRANT
Speaker 13	Yakama	HH W IH1 N	(IN)	WHEN	COSTLY		SPORTING	STORY	DO
Speaker 14	Yakama	HH W EH1 N	(IN)	WHEN	ins: [SAW] when	ins: [SAW] when	ins: [SAW] when	ins: [SAW] when	ins: [SAW] when
Speaker 15	Yakama	AO1 F AH0 N	(CC)	OFTEN	ins: [OFF] AND	ins: [OFF] ON	ins: [OFF] AND		
Speaker 15	Yakama	K UH1 D AH0 N	(CC)	COULDN'T		********			
Speaker 15	Yakama	AO1 F AH0 N	(CC)	OFTEN	ins: [OFF] OF	ins: [OFF] IN	ins: [OFF] IN		
Speaker 15	Yakama	D IH1 D AH0 N	(CC)	DIDN'T	RIDING				