The Language of Law: An Analysis of Gender and Turn-Taking in U.S. Supreme Court Oral Arguments

Haley Lepp

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

Master of Science

University of Washington

2020

Reading Committee:

Gina-Anne Levow

Richard Wright

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

The Language of Law: An Analysis of Gender and Turn-Taking in U.S. Supreme Court Oral Arguments

Haley Lepp

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

In this study, I present a corpus of short exchanges between speakers in U.S. Supreme Court Oral Arguments. Each exchange is labeled on a spectrum of "cooperative" to "competitive" by a human annotator with legal experience in the United States. To show the importance of this corpus, I analyze the relationship between speech features, the nature of exchanges, and the gender and role of the speakers. Finally, I train machine learning models with the corpus, and demonstrate that the models can be used to predict the label of an exchange with moderate success. The automatic classification of the nature of exchanges indicates that future studies of turn-taking in oral arguments can rely on larger, unlabeled corpora.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

TURN: An uninterrupted period of speech by one person taking part in a conversation.

TURN CHANGE: The period in which one person finishes a turn and another person starts a turn during a conversation.

COMPETITIVE TURN CHANGE: A label for a turn change, in which the second speaker competes with the first speaker for the chance to speak.

COOPERATIVE TURN CHANGE: A label for a turn change, in which the first speaker may expect a turn-change and give the floor to the second speaker; the second speaker may leave space for the first speaker to finish their turn; or a speaker may be giving backchannel feedback.

AUDIO SEGMENT: An audio clip with several seconds before and after a turn change, including the speech of at least two speakers.

CORPUS: The set of all audio segments and corresponding metadata and labels used in the experiments in this paper.

ACKNOWLEDGMENTS

I gratefully acknowledge the scholars who provided academic advice in this cross-disciplinary study: Gina-Anne Levow, Richard Wright, Keelan Evanini, Vikram Ramanarayanan, Rutuja Ubale, Shuju Shi, Victoria Zayats, Emily Bender, and the board, reviewers, and participants in Widening NLP at ACL 2019. In addition, I thank the legal professionals who supported data annotation and those who shared their networks so I could meet more annotators, and the friends and family who helped test the annotation survey. Finally, I express my appreciation for and inspiration from the public servants and activists who spend their daily lives fighting for equality and fairness in the U.S. Judicial System.

DEDICATION

To my grandparents:

Evie, who is ever committed to inclusion, equity, and public service,

Bud, who has represented his community in hearings and beyond for more than 50 years,

Anita, who has shown two generations of women how to get a word in edgewise, and

Bob, who taught me to read and inspires me to write.

Chapter 1

INTRODUCTION

Of the many institutions in which the U.S. government influences gender equality, none is as controversial and pressing as the Supreme Court. The Supreme Court plays a role in defining, identifying, and rooting out gender discrimination by hearing the cases that will determine the way gender-related rights are evaluated across the country. However, there are few checks on the presence of gender bias within the court itself. This study offers a novel corpus of annotated speech from Supreme Court Oral Arguments, and proposes a framework to analyze turn-changes according to gender in the oral arguments.

In statistics, the bias of an estimate is defined as "the difference between its mathematical expectation and the true value it estimates" (Bias, 2003). For example, if one estimates the amount of times an event can be expected to occur based on observations of a broad population, and then one observes the amount of times an event actually occurs on a subset of the population, the difference between the prediction and the actual occurrences of the event is considered the bias. This definition translates into the social sciences, in which a person can be said to express bias if the way they act, in general, toward some population of people is different from the way they could be predicted to act toward a random subset of people. For example, if speakers interrupt women more than they do men, then the use of an interruption as a speech act is biased toward women.

To determine whether the Supreme Court justices and the attorneys who argue cases demonstrate bias in the courtroom, one must look for ways that the distribution of their behavior toward a population is systematically different depending on a characteristic of that population. The most prevalent and influential behavior of the actors in an oral argument is conversation. In conversation, people interact with one another by taking turns speaking.

By grouping these turns into categories, it is possible to examine the distribution of those categories across the genders of the speakers. Turns can be categorized by sociolinguistic perception, and syntactic, phonetic, and acoustic features of turns in certain categories can be extracted automatically as a way to define those categories. For example, if listeners hear turns and label those turns as "competitive," and most or all turns labeled as "competitive" are generally louder than other categories of turns, loudness could be used as a defining feature of "competitive" turns.

There are hundreds of conversational turns in each oral argument, but only a limited number of speakers, and an even smaller number of women who speak. As such, to explore whether bias is present in turn-taking, an extensive number of conversational turns need to be labeled by category. Furthermore, the intricacy of speech acts and the presence of implicit biases among annotators make the standardization of manual labeling of turns unlikely to be accurate on a large corpus of turns. The scale of necessary analysis and the standardization required for accuracy suggest that machine learning models could both improve the definitions of the categories of turns, and use those definitions to predict what categories unlabeled turns fall into. Once a machine learning model has been trained to predict a turn label, a larger corpus of turns can be labeled automatically. Using that corpus, it will be possible to see whether certain categories of turns correlate with the genders of speakers. In addition, the extraction of speech features correlating with biased speech could provide tangible remediation techniques to train speakers to avoid bias in conversational discourse.

The first part of this paper describes a perception study in which legal professionals categorize turns from Supreme Court oral arguments. The second section outlines the extraction phonetic, acoustic, and syntactic linguistic features from these turns, and determines whether the features align with the sociolinguistic categories from the first part of the study. The end of this paper explores to what degree machine learning models can predict the competitiveness or cooperativeness of a turn based on speech features. All data created in this study is publicly available, and the framework makes use of open-source tools, with the hope that other researchers can build upon the results in the future.

Chapter 2

BACKGROUND

2.1 Categorization of Speech Acts

For decades, scientists have argued that women are interrupted more than men in professional settings, indicating that this speech act could be an indicator of gender bias.[1] Jacobi and Schweers (2017) find that interruptions correlating with gender within Supreme Court Oral Arguments have occurred consistently over time, and are not necessarily due to political polarization or personalities of certain justices. However, in conversational turn-taking, an interruption is not inherently a negative act. As demonstrated by Tannen (1994) in her seminal research on gender and language, interruptions cannot be defined categorically as acts of rudeness or dominance. Interruptions can be part of regular discourse depending on the context of a conversation, and are especially common among speakers of certain social groups in the United States.

Furthermore, the term "interruption" is not a clear-cut linguistic term. Some studies consider an interruption to be an overlap in speech between two speakers. Others find the term to be more complex, and consider backchannels[2] to be an exception to this rule (Laskowski, 2010; Yang, 2003). Goldberg (1990) indicates categories of interruptions; "power type" interruptions "wrest the discourse from the speaker" and can include a "topic change attempt", while "rapport type" interruptions "bolster the interruptee's positive face." Wichmann and Caspers (2001) explore not just interruptions, but any exchange, or "turn-change" between speakers, and create two binary categories to define the turn-taking: cooperative turns and

[1]The New York Times has described "being interrupted, talked over, shut down or penalized for speaking out" as "nearly a universal experience for women when they are outnumbered by men." (Chira, 2017)

[2]Such as when a speaker talks over another speaker with an "mhmm" or "uh-huh".

not-cooperative turns. If the first speaker in an exchange intends to cede the turn to the next speaker, then the turn is "cooperative." The opposite can be inferred as an instance in which the first speaker does not intend to cede the turn.

2.2 Feature Extraction from Turns

A number of studies aim to identify the linguistic features that define competitive exchanges as opposed to cooperative exchanges. Wichmann and Caspers (2001) find that the syntactic completion of a phrase is the most defining predictor of listeners expecting a turn to be ceded; for example, if a person has not finished their phrase yet, and another person starts talking, the listener would consider this exchange to be competitive.

Phonetic and acoustic features have also been shown to define or predict the nature of an exchange. Studies of interruptions in British and American English indicate that pitch contours bolster such syntactic cues to demonstrate whether an interruption is cooperative or competitive (Gorisch et al., 2012; Truong, 2013; Yang, 2003; Wichmann and Caspers, 2001). Yang (1996) discovers a correlation between high pitch and competitive interruptions, hypothesizing that raised pitch has to do with the speaker seeking additional attention. The duration of speech overlaps is also commonly cited as a feature used to distinguish between when a turn change is competitive or cooperative (Kurtić et al., 2010, 2013). A short overlap indicates a more cooperative exchange, while longer overlaps have speakers competing for the chance to speak. These studies have also found that a higher relative amplitude in a speech signal demonstrates a more competitive exchange.

In addition to the use of individual features to find patterns in turn-taking, scientists have used large groupings of features for speech recognition tasks. Certain features of speech have been documented by researchers in psychology, linguistics, and machine learning to correlate with emotion: high intensity and fundamental frequency mean correlate with expressions of stress, anger, and sadness. Certain groupings of these features are provided by the open source software `openSMILE` for use in emotion recognition research Eyben et al. (2016).

2.3 Machine Learning and Deep Learning for Automatic Labeling of Turn Categories

The use of machines to automatically label, or score speech, instead of or in addition to the use of human annotators, is a well-documented phenomenon in the assessment industry. Human reviewers are fallible to inconsistent reviewing, can drift in their standards as they review more examples, and are prone to making mistakes (Wang et al., 2018; Ling et al., 2014). Computers can complete reviews near-instantly, without susceptibility to inconsistencies, drift, or mistakes. Therefore, the design of a predictive model allows for the creation of larger and more accurately annotated corpora (Zhang, 2013).

While there are few examples of predicting distributions of bias and power in spoken language by studying non-lexical features, many researchers have explored bias using patterns in text conversations or transcribed speech conversations. Prabhakaran and Rambow (2014) use a binary support vector machine (SVM) classifier to detect overt displays of power in written dialogs, despite such displays being relatively rare within the corpus. Danescu-Niculescu-Mizil et al. (2013) use an SVM to detect human-annotated elements of politeness, and analyze the results of the classification to extrapolate the distribution of acts of politeness and power in broader contexts. Bramsen et al. (2011) use a similar method to extract social power relationships. Recent studies have used other models for emotion recognition in speech with relative success; Noroozi et al. (2017) use Decision Trees and Random Forests trained with speech features to predict six emotional categories, reaching a recognition rate of 45% accuracy averaged across emotions.

The Multilayer Perceptron neural network is also a common classifier of speech features for emotional recognition, and has been compared to deep learning methods in performance (Albornoz et al., 2011). State-of-the-art emotion recognition using deep learning algorithms do not show significantly higher results than do these classifiers, reaching an average of 59-64% accuracy across different emotions and models (Jiang et al., 2019).

Chapter 3

CORPUS CREATION

3.1 Motivation

To explore the distribution of turn categories in Supreme Court oral argument speech, I create a corpus of short audio recordings of turns labeled on a spectrum between two categories: competitive turns and cooperative turns. Neither of these categories is associated with a value judgement, as terms such as "interruption" can be, and so can be annotated with less potential for bias. The following sections describe the development of the turn-change corpus, the recruitment of annotators, the creation of an online survey for annotations, and the results of the annotation process.

3.2 Audio and Transcription Retrieval

3.2.1 Audio Files and Transcriptions

The transcriptions and audio recordings of all U.S. Supreme Court oral arguments since October 2006 are publicly available online. The oral argument recordings are available in mp3 format. Unfortunately, this format is lossy, losing details in compression of the original recording. The transcriptions, written by court stenographers, include the name of the speaker followed by the transcribed speech. The transcriptions include some disfluencies[1] and speech that ends mid-sentence or word[2] (The Supreme Court of the United States, 2019).

---

[1]such as "in the – in – in the case" (Mitchell v. Wisconsin, 2019)

[2]such as "And so we're certainly asking for this Court's –" (Mitchell v. Wisconsin, 2019)

3.2.2 Time-Aligned Transcripts

The transcriptions do not include the time at which each statement is said, so I retrieve publicly available time-aligned transcripts from The Oyez Project.[3] The Oyez research team synchronizes the transcripts with the audio recording by using `Aeneas` and `HTK` forced alignment (Pettarin, 2017; Young et al., 2015). Oyez displays an interactive animation which allows a user to listen to recordings of oral arguments while following along with the transcription from the court stenographer. I retrieve the time-stamps used in this animation by extracting the HTML from the animation.

3.2.3 Turn Extraction

I define a turn as each event in which a speaker speaks; or, any time a speaker name occurs in the transcript. For example, the following is considered a single turn:

> Andrew R. Hinkel: The Wisconsin statute at issue here doesn't lead to that result. (Mitchell v. Wisconsin, 2019)

I consider a turn-change to occur when one speaker stops speaking and a second speaker starts speaking, according to the transcript. For example, the following would be a turn-change:

> Ruth Bader Ginsburg: He's uncapable – he's incapable of hearing what he's told, but, in the – in – in the case of the unconscious driver, could his license be revoked?
> Andrew R. Hinkel: The Wisconsin statute at issue here doesn't lead to that result. (Mitchell v. Wisconsin, 2019)

[3]Oyez, a multi-media archive curated by Cornell's Legal Information Institute (LII), Justia, and Chicago-Kent College of Law, is "the most complete and authoritative source for all of the [U.S. Supreme] Court's audio since the installation of a recording system in October 1955" (The Oyez Project, 2019).

If a speaker starts speaking, and the stenographer records a second speaker begin to speak, even if the first speaker has not finished their sentence, I consider this turn-change. For example:

> Hannah S. Jurss: And so we're certainly asking for this Court's –
> John G. Roberts, Jr.: But I'm not faulting them for that.(Mitchell v. Wisconsin, 2019)

I extract the start and end time-stamp of each speaker's turn using the Python library BeautifulSoup to parse the HTML files from Oyez.com (Richardson, 2007). I devise a rule-based formula, described below, to determine the length of each audio file, which produces the names of the speakers involved in the turn exchange and the time stamps for the turn exchange. I segment the audio recordings into short clips using the sound processing software `SoX` (Sound eXchange, 2015).

3.2.4 File Segmentation

In the process described in the following section, annotators listen to audio files of turn-changes and sort those turn-changes into categories. The turn-change recordings that the annotators listen to include speech from the end of the turn of the first speaker and speech from the beginning of the turn of the next speaker.

The audio segments given to annotators are all less than six seconds long. This length is appropriate for several reasons. First, the annotators in this study are busy professionals. This condition provides a constraint as they have limited time to listen to and analyze full oral arguments. Second, multiple studies have demonstrated that listeners can perceive significant social and emotional information from a short slice of an audio, despite not knowing the greater context of a conversation (Ambady et al., 2006; Ambady and Rosenthal, 1993). In addition, because of the controversial nature of the oral arguments, and because this study aims to find patterns in speech without regard to the subject matter of the case, limiting the content which an annotator can listen to provides a control to avoid annotator bias.

Initially, I extracted turn-change segments of a total of four seconds in length; two seconds on either side of the end of the first speakers turn. Several test-annotators, both legally-trained and lay-person, listened to these segments and reported that they were too short to be coherent enough to label. I extended the default length to be six seconds long, or two seconds before the end-label of the first speaker and four seconds after the start label of the next speaker. The end-label of the first speaker and the start label of the next speaker are the same, and include silence. 56% of turn changes either had a preceding turn less than two seconds long, or a following turn less than four seconds long. I altered the time segments in these instances to ensure clarity for the annotators:

- If the turn of the first speaker is less than two seconds long, then I use the start of the first speaker's turn as the start of the turn change, instead of a full two seconds of audio.

- If the turn of the second speaker is less than four seconds long, then I use the end of that speaker's turn as the end of the turn change, instead of the full four seconds.

After segmenting the audio files according to the timestamps produced by this formula, I manually listen to every segment. I make the following changes, with the aim to delete or change as few audio segments as possible:

- I remove all segments in which at least one speaker is inaudible. This effect is usually due to a faint backchannel or an instance in which the stenographer heard something in person that the microphone did not record.

- I trim recordings if another turn occurs that makes it unclear what an annotator might be documenting. These instances are usually less than a second at the end of an audio file, and may be due to slight inaccuracies in time stamps.

- I extend recordings by no more than one second if the change becomes more clear with extension.

- I switch names if the ordering seems wrong or names were incorrect. For example, if a number of turns occur in quick succession or there are two or more speakers talking at the same time, I change the label so that the first speaker heard is the first speaker listed, and the second speaker heard is the second person listed.

- I remove turns that are scripted, such as "Mr. Chief Justice, and may it please the court."

- I remove turns that are listed as separate in the transcripts but are actually the same person with a pause.

An example segment is shown in Figure 3.1.

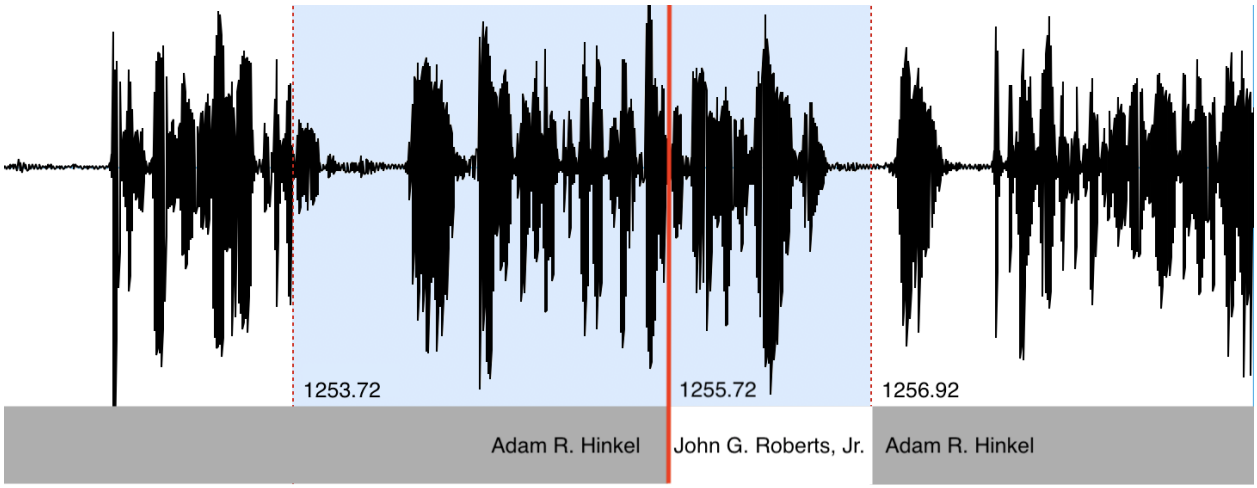


Figure 3.1: An example segment, highlighted in blue, which starts two seconds before the end of Mr. Hinkel's turn at 1253.72, and ends at the end of Chief Justice John Roberts' turn, because Justice Roberts' turn has a length of less than four seconds.

In total, I alter 21 segments and remove 101 segments from the corpus out of 813. At the time of writing, 711 segments are annotated by two different annotators.

3.3 Corpus Annotation

3.3.1 Motivation for a Survey

The conversational rules of speech within a courtroom setting are not the same as those in informal conversational speech; power-relationships, formal rules and procedures, and field-specific argument strategies are among the many factors that influence the ways that speakers interact within an oral argument. Turn-taking that may be considered competitive or disrespectful in another domain could be considered differently within the context of an oral argument. As such, the annotation of conversational turns from a courtroom setting should be completed by people who have experience working with that type of speech. In the annotation process, all annotators are required to be U.S.-based, and to identify as an attorney, judge, legal scholar, or law student in their second year or above.

Traditionally, linguistic studies requiring mass annotations make use of Amazon's Mechanical Turk, an online marketplace for participants to look for studies that will pay them for participation. As this is an unfunded study, and because all annotators are legal professionals unlikely to be searching for work with Mechanical Turk, I design an independent online survey. The survey is anonymous, does not record personally-identifiable information about participants, and only makes use of passive observation, and therefore is approved by the University of Washington Review Board as exempt from Human Subjects Review.

3.3.2 Survey Design

I design the survey to be easy to use and quick to complete. To ensure ease of participation, the survey uses the JavaScript library JSPsych (de Leeuw, 2015). The library uses a client-side framework to run experiments in a browser, allowing participants to participate in the study on a computer or mobile device at their convenience. I host the survey on the University of Washington's servers. The survey uses brief instructions and simple events to ensure participants can complete the annotations quickly.

The survey consists of a series of slides through which a participant proceeds by pressing

a "continue" button. A progress bar indicates to users how long the survey is, to help participants keep track of how long the survey will take to complete.

3.3.3 Demographic Questionnaire

When the user starts the survey, they are given a brief description of the purpose of the survey as well as a warning that the survey requires sound. The browser then prompts the user for an ID. During recruitment, which is described below, I give IDs to interested parties to share among their networks. The IDs serve to provide insight into recruitment methods led to the most hits, and also create a check on participant participation to avoid ineligible users from taking the survey.

The next page of the survey asks the user a series of demographic questions. As the way listeners perceive speech differs depending on dialect, culture, and many other factors, I collect demographic data to document whether the age, gender, ethnic, political, and linguistic diversity of listeners was relatively representative of the diversity of the United States. The questionnaire includes the following questions:

- What is your profession? (If you are a student, please list your year in law school.)
- What is your age?
- What is your gender?
- What is your race?
- Do you identify with a certain U.S. political party? If so, which?
- What is your best guess at the variant(s) of English you speak?
- When the users complete the questionnaire, they click "continue".

The responses for all but the last question are open-ended, so that users may self-identify. All questions are optional, so that users can choose to abstain from sharing.

3.3.4 Instructions

Following this data collection, the user is given instructions for the activity. The instructions are brief and define a clear task: to categorize short clips on a spectrum of "cooperative" to "competitive." The instructions present a brief description of the activity:

> In this survey, you will listen to a series of short clips from recordings of Supreme Court oral arguments.
>
> Using a sliding scale, you will listen to one person stop speaking, and another person start speaking. You will identify whether you consider the first and second speakers to be engaging in a cooperative or competitive conversational exchange.
>
> If you hear multiple exchanges in the recording, please only focus on the SECOND speaker in their response to the FIRST speaker.

Detailed descriptions of the terms are then described in a prominent, purple box. The descriptions define the category and provide tangible cases in which a turn-change would fit into a category. These descriptions demonstrate a division between the two categories of turn-changes, but make clear that the participant should use their training and experience to evaluate turn-changes with nuance.

> By cooperative, we mean that to your ears, the first speaker expects a turn-change and gives the floor to the second speaker. The second speaker might leave space for the first speaker to finish their turn. Or, the second speaker might talk at the same time as the first speaker, providing short spurts of feedback, for example saying "mhmm" or "yes".
>
> By competitive, we mean that to your ears, the second speaker competes with the first speaker for the chance to speak. You, the first speaker, or listeners might perceive this as a disruption to the previous speaker's speech. The second

> speaker may cause the first speaker to stop speaking, or talk over the first speaker to compete to be heard.

Further instructions are separated to another page to emphasize the importance of comprehension:

> The recordings are short, so you may not understand the content of the conversation; instead, please use cues like tone and speech speed to evaluate the exchange. If you are unsure how to classify the exchange, simply leave the slider in place.
>
> Please spend no more than a few seconds listening to and evaluating each recording.

3.3.5 Annotation

To ensure clarity of the task, the instructions precede two slides that show the interface that the participant will interact with during the annotation process. This interface includes an example recording and an explanation of why the turn change in that recording would be annotated in a certain way, as seen in Figure 3.2.

After the two examples, the survey alerts the user that the tasks will begin. Each task includes the prompt "How competitive or cooperative do you perceive this exchange to be?", which emphasizes to the participant that the annotation should be solely from their perception. Below this prompt is an audio element which the user can control, and a slider showing a spectrum from Competitive to Cooperative with Likert-style category labels .

A probe study of turns annotated by the author demonstrates that the vast majority of turns in a typical oral argument are neither competitive nor cooperative, or only slightly in either category. To accommodate this distribution, this study makes use of an annotation scheme implemented on a single spectrum, with a cooperative exchange on one end, and a competitive exchange on the other. This spectrum, represented on the page with the slider, gives the flexibility of allowing a granular distribution of results and leaves room for the creation of discrete categories for later analysis (Figure 3.2).

Figure 3.2: Instructional Slide in Annotation Survey

All audio segments collected in the corpus are listed in a directory on the server, twice. Every time a user opens the survey, recordings are selected from that directory using a JavaScript's Math.random() function. When the participant presses the "continue" button after reviewing an audio segment, that audio segment is removed from the directory on the server, so that when the next participant takes the survey, only unannotated audio segments

remain. On a limited number of occasions, one annotator heard the same two segments in one survey. In these cases, I remove that annotation and put the file back into the unannotated pool for another review.

The first two audio segments in the task, unbeknownst to the participant, are included for normalization purposes. These audio recordings should fall clearly on each side of the spectrum. The two segments clearly follow the cases listed in the instructions for each category. The example that follows the "cooperative" case includes a Speaker One who finishes their sentence, a pause, and a Speaker Two who replies after the pause. The example of the "competitive" case includes a Speaker One who does not finish their sentence, and a Speaker Two who loudly enters the conversation, causing the first speaker to stop speaking. The labels provided by the annotators for these examples show how the annotator would respond to a sample that could be labeled on an extreme of the spectrum.

After the normalization segments, the speaker is given up to 26 more segments to listen to. If the speaker leaves the survey before completion, all results are still saved. If the speaker continues past the final recording, they will see a slide thanking them for their time. Any time the survey is opened on a web browser, the results of answers to any response are saved to a file once the user clicks the "continue" button.

3.4 Annotator Recruitment

As described in the *Motivation for a Survey* section, participants are required to be either attorneys, judges, legal scholars, or law students in at least their second year of law school; and to have the United States be the primary country of practice or intended practice. To recruit annotators from this group, I shared a public call for participants through my personal Facebook, Twitter, NextDoor, and Instagram accounts, as well as with legal counsel staff at Educational Testing Service. Many people responded or shared my posts, saying that either they were eligible and would gladly share the survey with their network, or that they knew someone who was that they could share the survey with. I also contacted scholars or administrators of various law schools via email, but to my knowledge these emails did not

result in any participants.

I only shared the link to the survey through private messages, to prevent curious onlookers or ineligible participants from accessing the survey.

3.4.1 Participants

Almost 80 participants took the survey, leading to over 700 unique annotated segments. As legal professionals in the United States represent the American populace, it is also important that the annotators be relatively representative of the American populace. Table 3.1 shows the demographics of annotators in comparison to employed lawyers and the American population (American Bar Association, 2019; United States Census Bureau; Gallup, 2019). Percentages of participants are calculated as the number of responses annotated by a person with a certain identity divided by the total number of annotated responses. A more detailed description of how the descriptive statistics were calculated can be found in Appendix A.

3.4.2 Annotated Audio Files

The corpus includes almost all turn-changes taken in four oral arguments: Kahler v. Kansas (2019), Mitchell v. Wisconsin (2019), Virginia House of Delegates v. Bethune-Hill (2019), and Washington State Dept. of Licensing v. Cougar Den Inc. (2018). Each of these trials occurred in 2018 or 2019, covers a unique topic, and includes at least one female arguing before the court.[4] Information about the audio segments and annotations included in the corpus is listed in Table 3.2. The number of turns per attorney in the corpus ranged from 27 to 128. For justices, each of whom appeared in every oral argument, the number of turns per individual per trial ranged from 10 to 42; with one exception: Justice Clarence Thomas does not speak at all in any of the four oral arguments. Among justices, Justice Sonia Sotomayor and Justice Stephen Breyer are most highly represented, with over 130 turns each. As these

[4]The latter qualification narrows the selection considerably. The 2017-2018 term had the lowest number of women attorneys arguing before the court in at least seven years (Walsh, 2018). In 2018, only 15% of the people who argued before the court were women (Robinson and Rubin, 2019).

Indicator	Annotator Pool	American Bar Association Employed Lawyers	American Population
Gender/Sex			
Men	34.9%	64% ("Male")	49.2% ("Male")
Women	63.9%	36% ("Female")	50.8% ("Female")
Race			
Asian	8.6%	2%	5.6% ("Asian alone")
Black or African American	12.6%	5%	13.4%
Hispanic or Latino	1.3%	5%	18.3%
White	79.2%	85%	76.5% ("White alone")
Politics			
Left	70.5%	N/A	31%
Right	4.0%	N/A	30%
Independent or Other	22.5%%	N/A	38%

Table 3.1: Demographics of Annotators. Detailed explanations are listed in Appendix A.

measurements do not take into account length of turns, and the corpus does exclude turns below a certain length, it is worth noting that this measurement is not necessarily reflective of influence or representation in oral arguments.

3.5 Evaluation of Annotations

3.5.1 Annotation Distribution

Annotators grade each audio segment on a visual spectrum, as seen in Figure 3.3. The location on which an annotator places the slider is codified with a score between 0 and 100,

Number of oral arguments	4
Number of annotated segments	732
Number of annotators per segment	2
Number of unique annotators	77
Number of segments annotated by each annotator	1 - 26 (+ 2 “dummy” segments)
Number of Unique Male Participants	11 including Justice Thomas, who does not speak
Number of Unique Female Participants	9
Number of Justice to Non-Justice exchanges	338
Number of Non-Justice to Justice exchanges and	351
Number of Justice to Justice exchanges	22
Number of Non-Justice to Non-Justice exchanges	0
Number of Female to Female exchanges	127
Number of Male to Male exchanges	269
Number of Female to Male exchanges	165
Number of Male to Female exchanges	150

Table 3.2: Information about Annotated Segments

in which a 0 represents the most cooperative turn-change, and 100 is the most competitive turn-change. Annotators do not see the numerical score of their answer, though there are qualitative descriptions spread across the spectrum on the web interface.

The distribution of results in Figure 3.4 reflects the layout of the web interface. The highest peaks of score distribution are at either end of the spectrum and directly in the middle. This phenomenon indicates that annotators move the slider all the way to one end when an audio clip clearly sounds competitive or cooperative. The annotators leave the slider in place if the audio does not clearly fall into a category. There are also middling

Figure 3.3: Annotation Spectrum

peaks around where the survey interface has labels of "slightly competitive" and "slightly cooperative." These peaks indicate that annotators make use of the Likert-style guidelines, despite having the ability to drop the button anywhere on the slider.

When the labels for each segment are averaged, the extreme peaks subside, but are still present (Figure 3.5).

There is no gold standard label against which to check whether an annotation is "correct", so it is impossible to grade the annotations for accuracy. It is possible, however, to check whether an annotation is reliable; if audio files receive similar annotations by different people, then the annotation process can be considered reproducible (Artstein, 2017). We might also assume that reproducibly annotated segments are generally representative opinions of American English listeners, because of the linguistic, geographic, gender, age, and race diversity of the listeners. To determine whether the annotations collected are reliable, I analyze trends across individual annotators and across audio files, and propose five methods for interpreting the data.

Agreement

There are multiple ways to assess the level of agreement between annotators on the label of a segment. Below, I describe several coefficients which calculate observed inter-annotated agreement beyond the level of agreement by accident, or chance.

Cohen's κ reports the accuracy of annotations, or the amount of times two annotations

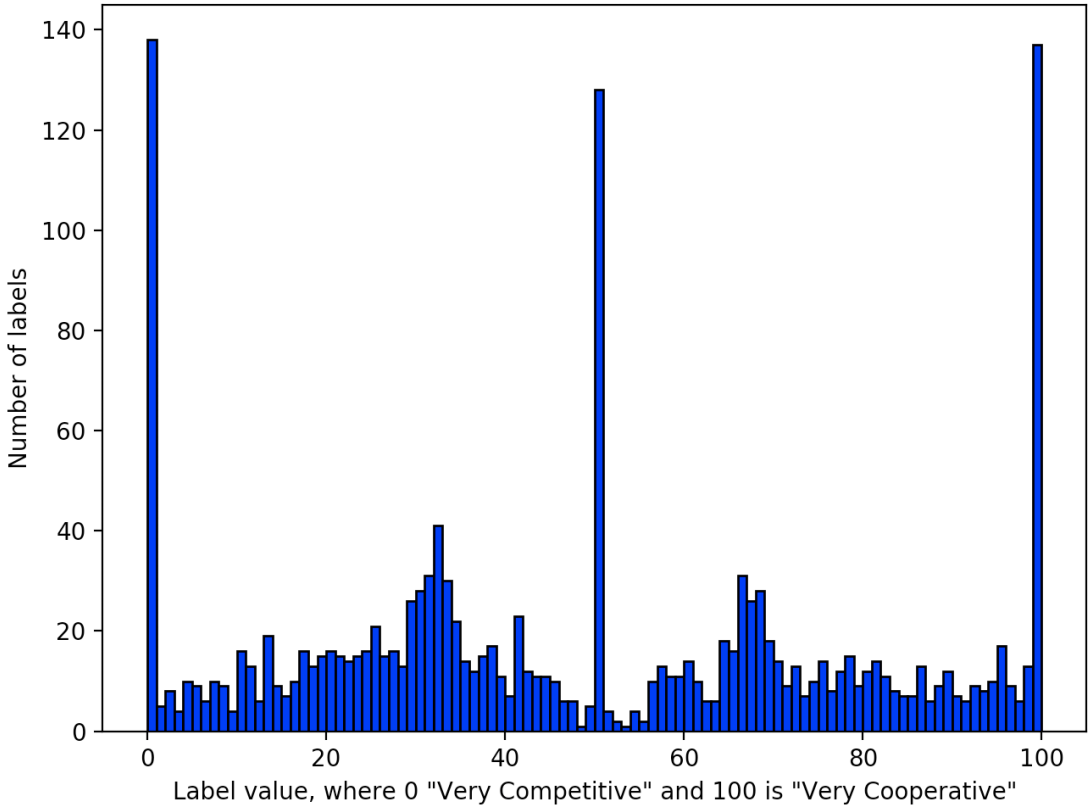


Figure 3.4: The Distribution of Labels in the Annotated Corpus

match, less the probability of chance agreement. Cohen's κ is defined by:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.1}$$

Where p_o is the frequency of observed agreement and p_e is the frequency agreement would occur by chance (Cohen, 1960). However, Cohen's κ considers every non-match to be a disagreement. In the raw numerical set produced by annotators, this is a poor measurement because a set of labels (99, 98) would influence the level of disagreement just as heavily as (99, 1) would.

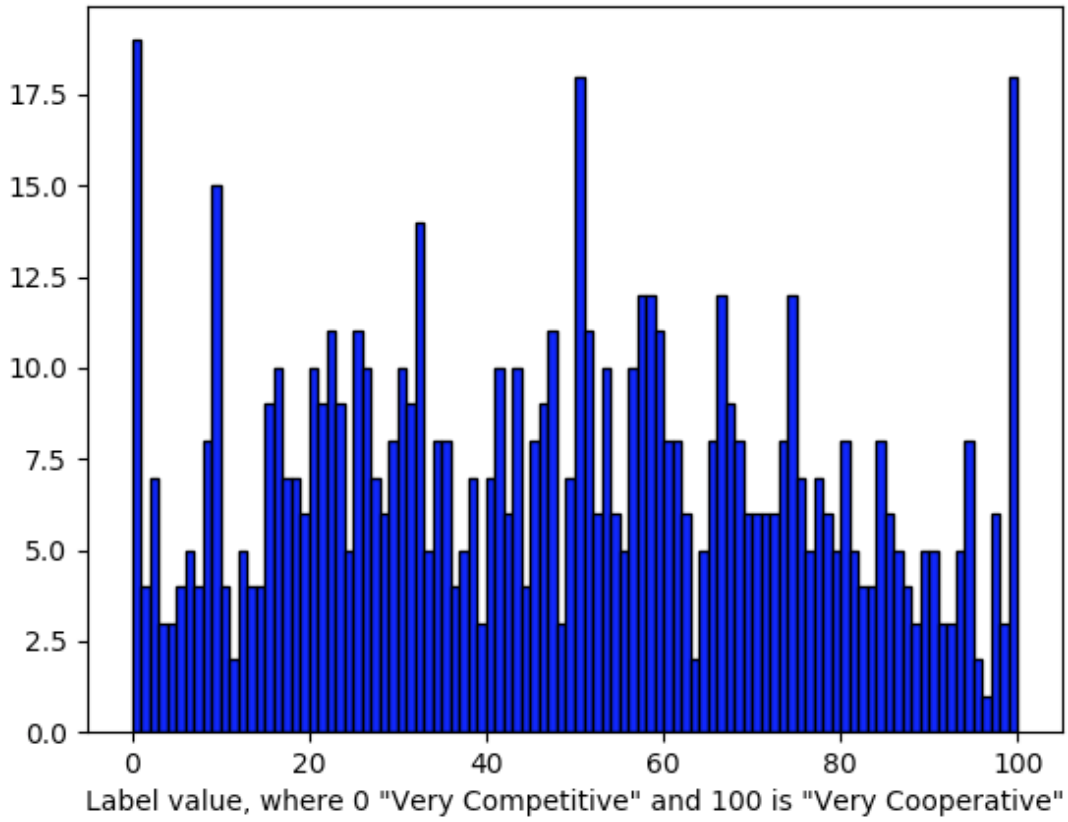


Figure 3.5: The Distribution of Averaged Segment Labels

A weighted Cohen's κ mimics the above calculation, except that it also takes into account a matrix of weights. In the weight matrix, the values down the diagonal are zero, indicating perfect agreement, while other values have increasing weights depending on their distance from the diagonal (Cohen, 1968). For example, in the matrix below, if the first annotation were a 75, and the second annotation 100, then the pair would be weighted with a 1, for mild agreement.

Second Label \ First Label	50	75	100
50	**0 (perfect agreement)**	1 (mild agreement)	2 (low agreement)
75	1 (mild agreement)	**0 (perfect agreement)**	1 (mild agreement)
100	2 (low agreement)	1 (mild agreement)	**0 (perfect agreement)**

Table 3.3: Example Weighted Cohen's κ Weight Matrix

The equation multiplies each agreement or disagreement by the accordant weight:

$$\kappa_w = 1 - \frac{\sum \nu_{ij}\rho_{oij}}{\sum \nu_{ij}\rho_{eij}} \tag{3.2}$$

where ρ is the frequency of a label for a segment and ν is the weight assigned to the two labels (Cohen, 1968). In the results below, I calculate ν by squaring the distance between the two values:

$$\nu_{ij} = (l_i - l_j)^2 \tag{3.3}$$

Krippendorff's α also incorporates the distance between each set of labels, thereby considering pairs which are closer together more agreeable than pairs that are further apart (Krippendorff, 2019). This coefficient calculates agreement by counting and scaling the pairs of labels which disagree, and dividing this by the probability that an arbitrary disagreement would occur:

$$\alpha = 1 - \frac{d_o}{d_e} \tag{3.4}$$

where d_o is observed disagreement and d_e is expected disagreement, and in which disagreement is defined by the sum of all the label pairs scaled by distance:

$$d = \frac{\sum(l_i * l_j * dist(l_i, l_j))}{\sum l(\sum l - 1)} \tag{3.5}$$

where l_i and l_j are the labels on a segment, and dist(l_i, l_j) is the distance between the two labels. Like in Cohen's κ, I calculate this value as the square of the distance between the two labels. The α measure demonstrates the reliability of the data in its rawest form, with results closest to 1 being the most reliable.

Correlation

Recent studies have argued that correlation coefficients should also be used to evaluate the reliability of annotations. While agreement coefficients calculate the extent to which two labels are the same, correlation coefficients measure the extent to which one label can predict another label. Exploring how the labels provided by individual annotators correlate with the set as a whole can provide insight into the consistency of a specific annotator(Amidei et al., 2019).

As such, I use Spearman's ρ to calculate annotator correlations below. Spearman's ρ evaluates the monotonic relationship between two sets of values; in other words, the degree to which higher labels match with other higher labels and lower labels match with other lower labels, without regard to linearity. Instead of using the values of the labels, as other correlation coefficients do, Spearman's ρ makes use of the rank of labels within the set. Spearman's ρ is calculated by dividing the covariance of two ranks by the standard deviation of the ranks (Spearman, 1904).

3.5.2 Reliability of Individual Annotators

Two people annotate each audio file in the corpus, so measuring the agreement of annotators is one way to explore the reliability of the annotation process. As the annotators in this study are relatively representative of the U.S. population and the population of registered legal professionals in terms of demographics, it can be predicted that the agreements come from shared assumptions and opinions across the full population.[5]

Difficulties

There are several factors that make standardization of labeling across annotators difficult. First, not all annotators completed the full survey, and the range of annotations completed

[5]The demographics of the annotators in comparison to the American Bar Association's list of employed lawyers, as well as to the U.S. population, can be found in Table 3.1.

by each annotator ranges from one (in addition to the two dummy annotations used for normalization) and 26. Second, as all annotators reviewed a random sample of clips, it can not be expected that the values of annotations would be centered around a norm.

Correlation by Annotator

Because the annotators are anonymous and the level of formal training for the task is limited, I check the set of annotators for the level of disagreement of each annotator on the labels they produce. For each annotation produced by a particular annotator, I compare the other annotation of the same audio file. I calculate Spearman's ρ of the values produced by that particular annotator against the other annotations produced by other annotators for the same files. Of the 76 annotators, 50 annotators have correlations of above 0.50, and 60 have correlations above 0.40. The mean correlation is 0.58, indicating that in general scores produced by individual annotators correlate with the scores produced by others.

3.5.3 Reliability of Audio File Annotations

In the same way that consistency among individual annotators can be a hallmark of reliability, consistent labels on individual segments can also provide insight into annotation reliability. It can be assumed that high variability in labels on a single annotation could indicate a mistake by an annotator or that the segment itself could not be annotated within the parameters of the annotation instructions.

Qualitative analysis of the audio files with the lowest inter-annotator agreement demonstrates patterns that bolster this assumption. There are several instances of extremely high disagreement in which depending on interpretation, an audio segment could be conceived as falling into both or either category. For example, in several clips the speakers talk over each other but the content of their speech is polite: phrases like "If I may, your honor"(Mitchell v. Wisconsin, 2019) and "Sorry, sorry your honor"(Washington State Dept. of Licensing v. Cougar Den Inc., 2018) might sway the annotator to mark an otherwise competitive interaction as cooperative.

3.5.4 Reliability of Labels

I also explore the symmetry of the agreement of annotators. Reflecting on the actions a user can take, I divide the spectrum into three categories: cooperative (move the slider to the right), neutral or unsure (do not move the slider), and competitive (move the slider to the left). For every label that falls into one of these categories, I calculate Spearman's ρ for the labels attached to that segment in Table 3.4.

	Cooperative	Neutral	Competitive
Spearman's ρ [correct]	0.172	-.016	.311

Table 3.4: Agreement Asymmetry by Category

From these results, we can assume that the most reliable labels are associated with competitive-sounding turns, while cooperative turns may appear less clear. The turns that are neither cooperative nor competitive have the ρ value of a null hypothesis, which is predictable as annotators were instructed to leave the slider in place if they are unsure of a result.

3.5.5 Annotation Standardization

To minimize differences between individual annotators, I experiment with several methods of interpreting scores. There are several segments that receive more than two labels; for these segments, the two most similar labels are selected. I compare these configurations by the level of agreement achieved between annotators on each audio file.

- Numerical: I use the raw score provided by the annotator.

- Calibrated: As the samples provided to each annotator were random, and as not all annotators completed the full survey, typical normalisation methods such as a z-score calculation or mean-centering the responses are not feasible. To account for the

differences in how individual annotators utilize the sliding spectrum, I calibrate every annotation performed by that annotator to a range between 0 and 100. For example, if the lowest score provided by an annotator is 32, that score is calibrated to 0, and all other scores were adjusted linearly. Because the first two audio clips in each survey fell clearly into either category, we can be sure that there is at least one audio clip that can represent each extreme case.

- Quintiles: I divide the score into five categories, organized according to quintiles for the entire set.

- Five Bins: I divide the score into 5 categories, organized according to the labels on the chart and the peaks in the raw score distribution (e.g. the first category is scores ranging from 0 to 20).

- Tertiles: When the annotator sees the spectrum on the survey, the button is set to 50, right in the middle of the slider. Therefore, the annotator is faced with a binary choice: to leave the button in place, or slide it in either direction. I divide the score into three categories, organized according to tertiles for the entire set.

	Numerical	Calibrated	Quintiles	Five Bins	Tertiles
Spearman's ρ	0.556	0.557	0.539	0.547	0.525
Krippendorff's α	0.553	0.555	0.538	0.542	0.524
Weighted Cohen's κ	0.553	.555	0.538	0.542	0.524

Table 3.5: Agreement of Different Data Interpretations

Interestingly, the metrics differ minimally across the data interpretations, with the Calibrated data set indicating the highest reliability.

3.5.6 Distributions of Turn Types in the Annotated Corpus

Using these interpretations of the annotations, certain distributional trends emerge. For each segment, I take the sum of the average label for all cooperative turns in a certain turn category, and divide that by the number of turns of that category in the full dataset. I define a turn as cooperative when it has an average label that is greater than the top tertile of the corpus. I do the same for competitive turns, and define competitive as any value less than the bottom tertile. The percentages in Table 3.6 are calculated from the Numerical interpretation.

Turn Participants	Percentage of Competitive Turns	Percentage of Cooperative Turns
Male-to-Female	0.247	0.387
Female-to-Male	0.400	0.248
Male-to-Male	0.338	0.331
Female-to-Female	0.323	0.362
Justice-to-Justice	0.636	0.045
Attorney-to-Justice	0.481	0.160
Justice-to-Attorney	0.154	0.524

Table 3.6: Percentage of Competitive and Cooperative Turns by Turn Participants

In general, there is more significant variation between categories related to role in court than there is to gender. Interestingly, justices are highly likely to speak competitively to one another. This could be due to the fact that they rarely speak to each other (only 22 times in a corpus of over 700 exchanges), and that in many of the moments in which a turn change involves two justices, both are speaking at the same time about some controversial topic to an attorney. Attorney-to-justice turns have a distribution which peaks at slightly before a full competitive rating, with a median of about 30, where 0 is competitive and 100 is cooperative. Over 80% of justice-to-attorney turns are cooperative or neutral; this could

be due to the power differential between the roles, which could incline attorneys to avoid speaking competitively with justices. There are no attorney-to-attorney turns.

The gender variance, though less pronounced, is present. More female-to-male turns sound competitive to annotators; in only 25% of turns is the exchange cooperative. When speaking after a male speaker, female speakers more commonly speak cooperatively. Turns between speakers of the same gender have tri-modal label distributions similar to that of the whole corpus (see Figure 3.4), while the median male-to-female turn is neutral, and the distribution of female-to-male turns peaks both at neutral and competitive.

The distribution of labels per person gives further insight into the impact of role on turn-type. Figure 3.6 shows the prevalence of certain labels when the speaker on the horizontal axis is the first speaker in a turn. Significantly, every justice has a median label that is more cooperative that every attorney, demonstrating that it is much less common for a justice to be the first speaker in a competitive turn. Figure 3.7 shows the prevalence of labels for the second speaker in a turn, and follows the same pattern; every attorney has a median label more cooperative than every justice, suggesting that it is much less common for an attorney to be the second speaker in a competitive turn. There are no clear gender trends, which may indicate that due to the small number of speakers in the corpus, certain speech features being captured are unique to individuals, and not to their genders as a demographic.

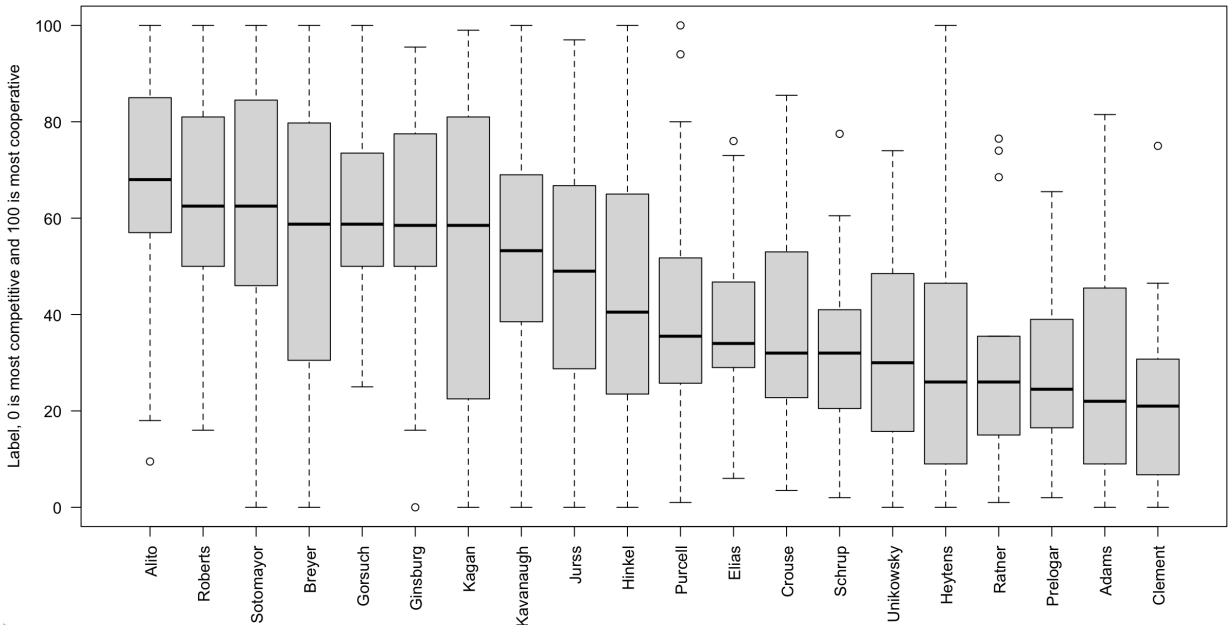


Figure 3.6: Label Frequency for the First Person in a Turn

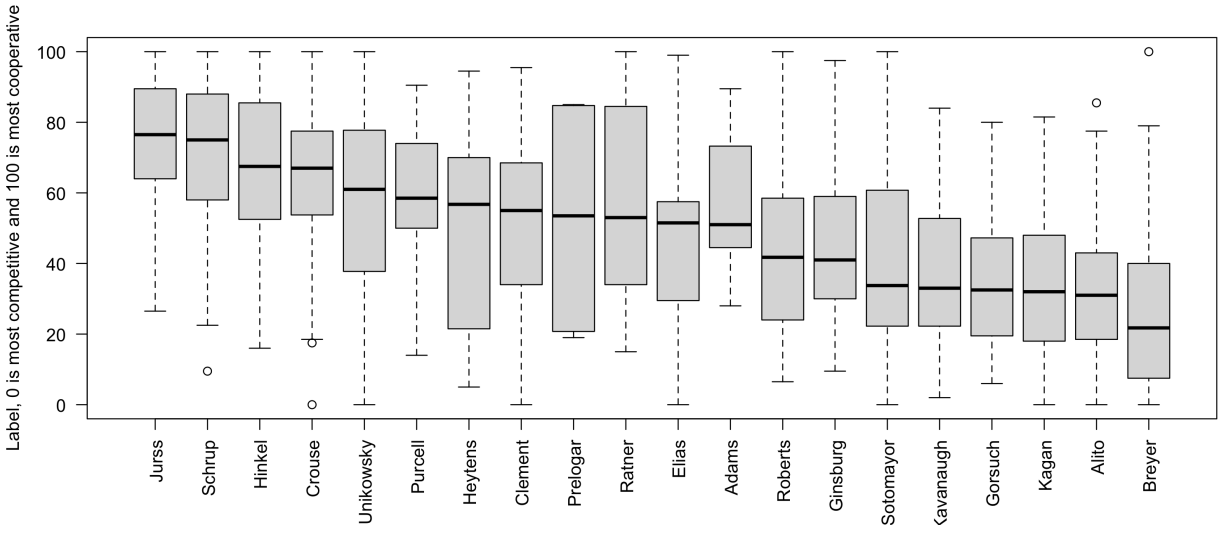


Figure 3.7: Label Frequency for the Second Person in a Turn

Chapter 4

FEATURE EXTRACTION

The next step in exploring whether a computer can automatically predict whether a human would hear a turn as "competitive" or "cooperative" is to extract speech features from each audio segment.

Following the hypotheses described in Section 2.2, I calculate the attributes, or features, of each audio segment in the corpus, related to the interval of time between turns, the loudness of the speakers' voices as perceived by the annotator, and the patterns in pitch used by speakers. In addition to the features outlined in my hypothesis, I extract sets of features determined by previous studies to correlate with levels of certain emotions in speech. In the rest of this section, I describe the feature extraction process in detail.

4.1 Silence

Linguists have demonstrated in previous studies that the length of overlap in speech between speakers in a turn-change is often associated with more competitive speech (Kurtić et al., 2010, 2013). In these studies, if one speaker is talking and another speaker starts speaking, and both speakers continue speaking at the same time, a turn is more likely to be considered competitive. Alternatively, a shorter overlap, or even a short duration of silence, between two speakers' turns, can indicate a more cooperative exchange.

It is difficult to extract the length of an overlap in speech automatically, because in the Supreme Court, audio is recorded in a single channel, instead of having speakers each have their own channel that is easily separable [1]. As a proxy, I calculate the duration of silence

[1]State-of-the-art speech diarization systems cannot accurately differentiate between different voices speaking at the same time. This is known as the Cocktail Party Problem, and is a high priority challenge in signal processing (Wang D., 2008).

between the speech of the first and second speaker, with the assumption that a segment with no silence detected could indicate that there is speech overlap, and any duration of silence greater indicates that the speakers did not speak at the same time.

From each file, I use `FFmpeg` (FFmpeg Developers, 2016) to extract any amount of silence. I extract silence at a -30dB noise tolerance at a length one tenth of a second or longer. This length boundary, determined based on studies of the duration of closure of the vocal tract in spontaneous speech, excepts the silence in an audio segment that might arise from stops or voice onset transition within a turn (Yao, 2007). For each silent period found in a segment, if that silent period overlaps with the point at which a turn is indicated in the time stamps provided by The Oyez Project, I use the duration of that silence period to represent the duration of time between turns, as shown in Figure 4.1: where S_d is the duration of silence

$$S_d = \begin{cases} S_e - S_s, & \text{if } S_s < T < S_e \\ 0, & \text{otherwise} \end{cases}$$

Figure 4.1: Silence Duration

used for the feature, S_s is the time-stamp of the start of a silence detected that is within the noise tolerance and duration floor, S_e is the end time-stamp of that detected silence, and T is the time-stamp of a turn as detected by The Oyez Project.

Approximately 80% of the segments do not have a gap in speech by this definition, so the value of the feature for these segments is 0. It can be assumed then that 80% of files contain some level of speech overlap. Among all segments, the maximum silence length is 2.67 seconds.

4.2 Amplitude

Linguistic studies of interruptions, which are sometimes associated with competition in turn-taking, have found higher intensity and amplitude measurements around the turn-change in

turns that are perceived to be interruptions (Yang, 2003; Kurtić et al., 2013).

While the amplitude at which each speaker speaks may be distorted by the recording devices in Supreme Court oral arguments, the amplitude of each speaker on the recording is what is perceived as loudness by any listeners of the recordings. As such, I incorporate amplitude-related features into models as perceptual measurements.

I split the turn segment into two sub-segments; the first speaker's speech up until the point of the turn, and the second speaker's speech from the point of the turn until the end of the segment given to the annotators. For each of these sub-segments, I measure the maximum amplitude of the segment, the mean amplitude of every sample in the segment, and the root mean square amplitude of the segment, using `SoX` (Sound eXchange, 2015). The files are sampled at 48 kHz.

Each speaker has a range of amplitude unique to their voice, so I take the z-score of amplitude measurements of each speaker in a turn across all instances in the corpus in which that speaker speaks in the same role in the turn (e.g., I normalize each time Justice Ruth Bader Ginsburg is the first speaker in a turn separately from every time she is the second speaker in a turn).

4.3 Pitch

Gorisch et al. (2012); Truong (2013); Yang (2003); Wichmann and Caspers (2001) and others have shown the contour of pitch near a turn-change to influence the way listeners perceive a turn exchange. The pitch contour can be quantified by taking measurements of fundamental frequency (F0) across samples of an audio segment.

Figure 4.3 shows the spectrogram of a turn-change that annotators consider to be competitive. The F0 measurement, represented by the blue line, jumps in the middle of the figure when Justice Stephen Breyer interrupts attorney Andrew Hinkel. They overlap briefly, then, starting at the red vertical line, Justice Breyer steadily lowers his pitch as he takes the floor. In the 3.5 second segment, the difference between the maximum F0 measurement at 245.1 Hz, and the rest of the conversation, which hovers around 75 Hz, demonstrates how

significantly the pitch can change.

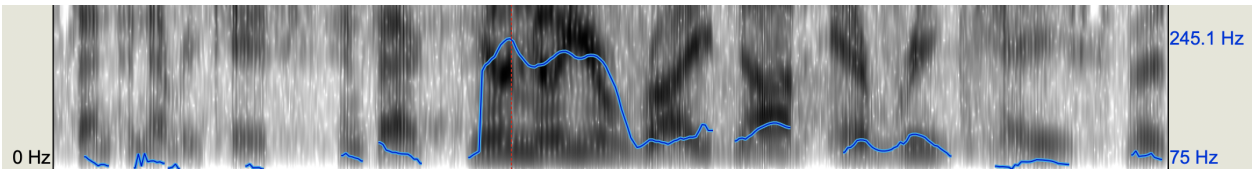


Figure 4.2: Example Pitch Contour in a Competitive Turn

I split each audio segment and take pitch measurements for each speaker. Instead of extracting every F0 measurement of every sample in a segment, I calculate the approximate slope and offset of the pitch contour for each speaker. I use `openSMILE` (Eyben and Schuller, 2015) to calculate four pitch values for each segment:

1. the slope of a linear approximation of an F0 contour, smoothed by a moving average filter;

2. the offset of a linear approximation of an F0 contour, smoothed by a moving average filter;

3. the envelope of the slope of a linear approximation of an F0 contour, smoothed by a moving average filter;

4. the envelope of the offset of a linear approximation of an F0 contour, smoothed by a moving average filter.

4.4 Feature Analysis

Using a linear mixed effects model implemented in R, I explore the effect pitch and amplitude have on tertile labels (Bates et al., 2015). For each speaker in a turn, I use maximum amplitude z-scored by speaker (Figure 4.3) and the slope of the linear approximation of the pitch contour as fixed effect (Figure 4.4). The interaction effect between the measurements

for both speakers in a turn is also listed. The label is the dependent variable, the gender and role of each speaker in a turn are additional fixed effects, while individual speakers are random effects.

From the model, we can see that the pitch and amplitude of the first speaker are negatively related to the score; in other words, lower amplitude and pitch in the first speaker are associated with more cooperative interactions.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.51	1.15 – 1.88	**<0.001**
X1_Maximum_amplitude_zscore	-0.17	-0.22 – -0.11	**<0.001**
X2_Maximum_amplitude_zscore	-0.02	-0.08 – 0.03	0.424
gender_1	-0.09	-0.26 – 0.09	0.321
gender_2	0.07	-0.11 – 0.26	0.440
role_1	-0.17	-0.51 – 0.16	0.313
role_2	-0.86	-1.20 – -0.52	**<0.001**
X1_Maximum_amplitude_zscore * X2_Maximum_amplitude_zscore	-0.00	-0.05 – 0.05	0.911
Random Effects			
σ^2	0.47		
$\tau_{00\ person\_1}$	0.02		
$\tau_{00\ person\_2}$	0.03		
ICC	0.09		
$N_{\ person\_1}$	20		
$N_{\ person\_2}$	20		
Observations	711		
Marginal R^2 / Conditional R^2	0.237 / 0.306		

Figure 4.3: Linear Mixed Effects Model; Amplitude

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.52	1.15 – 1.90	**<0.001**
X1_F0_sma_linregc1	-0.02	-0.03 – -0.00	**0.024**
X2_F0_sma_linregc1	-0.00	-0.02 – 0.02	0.931
gender_1	-0.08	-0.26 – 0.10	0.366
gender_2	0.06	-0.14 – 0.25	0.564
role_1	-0.18	-0.53 – 0.16	0.297
role_2	-0.85	-1.20 – -0.50	**<0.001**
X1_F0_sma_linregc1 * X2_F0_sma_linregc1	-0.00	-0.01 – 0.00	0.458
Random Effects			
σ^2	0.49		
$\tau_{00\ person\_1}$	0.02		
$\tau_{00\ person\_2}$	0.03		
ICC	0.10		
$N_{\ person\_1}$	20		
$N_{\ person\_2}$	20		
Observations	711		
Marginal R^2 / Conditional R^2	0.198 / 0.275		

Figure 4.4: Linear Mixed Effects Model; F0

4.5 Feature Sets

There are several sets of features that are commonly used in speech classification. Using `openSMILE`, I extract features summarized across a speaker segment (as opposed to by-sample features) from three feature sets for each speaker in each audio segment in the corpus. I describe these feature sets in Figure 4.5.

As I did with amplitude-related measurements, I take the z-score of features in each feature-set for each speaker in the corpus as a first speaker, and each speaker in the corpus

Figure 4.5: `openSMILE` Feature Sets

Feature Set	Number of Features	Primary Focus
prosodyShsViterbiLoudness (Eyben and Schuller, 2015)	36	pitch-related features
eGeMAPSv01a (Eyben et al., 2016)	88	a limited number of linguistically and psychologically-informed features
emobase (Eyben and Schuller, 2015)	988	live emotion recognition

as a second speaker. I normalize these features to ensure that features of an individual speaker's unique voice sway a model's ability to make predictions about speakers in general. In addition, many features in these feature-sets measure slopes or deltas, which do not need to be normalized.

Chapter 5
EXPERIMENTS

Using the labeled data from the corpus, I explore whether machine learning can accurately predict the distribution of competitive and cooperative turns across a larger, unlabeled corpus of segments.

5.1 Data Division

I divide the labeled corpus into two sub-corpora: 80% of the corpus into a training set and 20% into an evaluation set. Each set has relatively comparable gender and role distribution across turn-type; for example, 21% of the turns in the full corpus, training set, and evaluation set are male-to-female, and 49% of all turns in each set are attorney-to-justice. I use the 80% subset for training the model, and the 20% set for testing the model, so that the model can be evaluated on entirely unseen labeled data. The training data is divided randomly into subsets for grid search for hyperparameter optimization of the models.

5.2 Models

I use two experiment groupings for label prediction of each segment. Each grouping makes use of a different class of supervised machine learning algorithms.

5.2.1 Classification

In the first set of experiments, I use the mean of a segment's two raw labels given by annotators in a 0 to 100 scale, then, as described in Section 3.5.5, I categorize the averages into quantile-based bins. For these categories of labels, I use classification-based models to map the set of features of a segment to a predicted class. Because classification techniques do

not evaluate the ordering of classes, these models remove the assumption of a spectrum and evaluate competitive, cooperative, and neutral labels as unrelated bins. For classification experiments, I use Random Forest and Support Vector Machine classifiers.

Random Forest Classifier

A Random Forest classifier (RF) is, eponymously, a forest of decision tree classifiers. Each decision tree makes decisions based on the features of audio segments, to branch off toward the leaves of the tree, or class labels. For example, a subsection of a tree might have two branches, one where a turn-change has a period of silence and therefore leads to a "cooperative" leaf, and one where a turn-change has no silence, and leads to a "competitive" label leaf. A RF classifier, then, produces the most common class prediction of a segment based on a forest of decision trees (Liaw and Wiener, 2002).

Support Vector Classifier

A Support Vector Machine classifier (SVC) groups each instance, or audio segment, into classes based on label, and determines a hyperplane that divides those classes. If each instance is a p-dimensional vector of features, the class groupings will be divided by a p-1-dimensional hyperplane, which maximizes the distance between classes on one side and the other. Once trained on the training set, the model will then use the features in the evaluation set to predict which side of the hyperplane each instance will fall on (Weston, 1998).

In Figure 5.4, I use the SVM and RF classifiers provided by SciKit Learn using the SciKit Learn Laboratory Toolkit, with features scaled by standard deviation and centered around a mean, and a micro-averaged F_1 score as a grid search objective (Pedregosa et al., 2011).

5.2.2 Regression

In the second set of experiments, I use the average of a segment's two raw labels given by annotators in a 0 to 100 scale. For this continuous range of values, I use regression-based models to map the set of features of a segment to a predicted value.

Multilayer Perceptron Regressor

The first type of model, a Multilayer Perceptron regressor (MLP), is a feedforward neural network which uses regression to map an input instance, or the features of an audio segment, to a label. The MLP transforms the features using a non-linear activation function to a space in which different instances can be linearly separated depending on their label. After multiple hidden layers, or transformations, in which the network self-corrects errors using feedback from the labels in the training set, the network outputs the predicted label of the instance (S. K. Pal, 1992). Using the model trained on the training instances, I run the model on the evaluation instances, and do not allow the model to learn from the labels. I report the ability of the network to predict the labels of the unseen instances in Table 5.3.

Support Vector Regression

The second model, a Support Vector Machine regressor (SVR) is a support vector machine which uses regression to predict a label for an audio segment. As there is not necessarily a straight line that maps a vector of features to a label, the SVR uses a kernel function to map the linear regression function into higher dimensional feature space and determine a regression that can map the features of an instance to a label (Awad M., 2015).

In the results in Table 5.3, I use the MLP and SVR implementations provided by SciKit Learn using the SciKit Learn Laboratory Toolkit, with features scaled by standard deviation and centered around a mean, and the R^2 score as a grid search objective (Pedregosa et al., 2011).

5.2.3 Metrics

For the classification models, I use the micro-average of the harmonic mean, or F_1 score as the objective for models and report the score as a determination of the quality of the model. I calculate F_1 as

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{5.1}$$

where recall is

$$recall = \frac{tp}{tp + fn} \tag{5.2}$$

and precision is

$$precision = \frac{tp}{tp + fp} \tag{5.3}$$

where tp is the number of instances in which the label is predicted correctly, fp is the number of instances in which a label is predicted for an instance with a different label, and fn is the number of times the label should have been predicted but was not (Tharwat, 2018). I calculate the metrics for each label, then find the unweighted mean of these metrics. I report the F_1 for each label, and the micro-average (with metrics calculated globally, and not by label), for the entire model.

For the regression models, I train the models with the objective of the highest R^2 score, and record the R^2 score of each model. To calculate the R^2 score, I calculate the residuals of the prediction, or the difference between the actual score and the prediction for each instance, to find the residual sum of squares, where

$$SS_{residual} = \sum e_i^2 \tag{5.4}$$

where e_i is the residual score of an instance. I also calculate the total sum of squares as

$$SS_{total} = \sum (y_i - \bar{y})^2 \tag{5.5}$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{5.6}$$

where y_i is the actual label of an instance. The R^2 score combines these metrics as

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}. \tag{5.7}$$

I use the implementation from Pedregosa et al. (2011).

5.2.4 Baseline

As a baseline for each model, I calculate the quality of classifiers in predicting the labels of the instances in the evaluation set when every label is entered (artificially) as the median label in the set. The results listed in Table 5.4 can be compared to this baseline to see the added value of having a trained model.

5.3 Results

The learners are most effective at predicting the labels of segments which are at either end of the spectrum; in other words, the segments with the most cooperative and the most competitive labels. For the two classifiers, I report the results of the category which captures the most extreme labels. So, for tertile-category labels, I report the results of the model in predicting the tertile of segments labeled most cooperative and the tertile of segments labeled most competitive. Likewise, for quintile-category labels, I report the results of predictions of the most competitive and cooperative quintiles. I do not report the middle, or neutral, quantile(s), which have the noisiest data and the least consistent labels (as annotators were instructed to leave the slider in the middle if they were unsure).

For each group of features, I find that the features that have been normalized by speaker produce inferior F_1 results compared to the equivalent non-normalized sets (Table 5.1). In the tertile-category set with segments that have the most cooperative category label, this leads to a 0.02-0.1 loss. In segments with the most competitive category label, the loss is larger: .01-.2. This is surprising, considering features such as pitch and amplitude generally vary by speaker sex due to speaker physiological differences. One theory as to why this could occur is that the sex of speakers correlates so strongly with the label of a turn that the

raw features, influenced by sex differences, actually help the models predict the correct label. Alternatively, female speakers may be limited in their vocal range compared to typical ranges of speech (Figure 5.1).[1] Finally, many features in these sets are deltas of measurements, so the features that would normally require normalization have limited effects on the models.

	eGeMAPS		Prosody	
Model	Competitive	Cooperative	Competitive	Cooperative
SVC with Normalized Features	.596	.547	.686	.544
SVC with Raw Features	.636	.551	.687	.561
RF with Normalized Features	.579	.558	.614	.636
RF with Raw Features	.617	.593	.640	.611

Table 5.1: Range in F_1 score using two feature sets for normalized and raw features

Another indicator that gender may be supporting the quality of the classifier is that when I add the gender and role of the speakers to the feature sets, models trained on certain feature sets see a modest rise in F_1 scores. For example, the Prosody feature set, which hypothetically would better differentiate gender because of average pitch differences between genders, produces higher F_1 scores for segments with competitive labels when gender is added. The gender and role of each speaker in a competitive exchange is so influential as to be indicative of the label. Other feature sets, such as eGeMAPS, which include features not typically correlated with gender, have a modest decrease in F_1 scores in the Random Forest model when gender and role are not included (Figure 5.2). While the extra features help predict competitive labels, cooperative label predictions suffer slightly, indicating a less significant relationship.

In general, the differences between models are relatively small. The RF models are better at predicting the labels at the extremes of the spectrum, while the SVCs capture

[1]It has been documented that women in professional settings tend to lower the pitch of their voices (Pemberton et al., 1998).

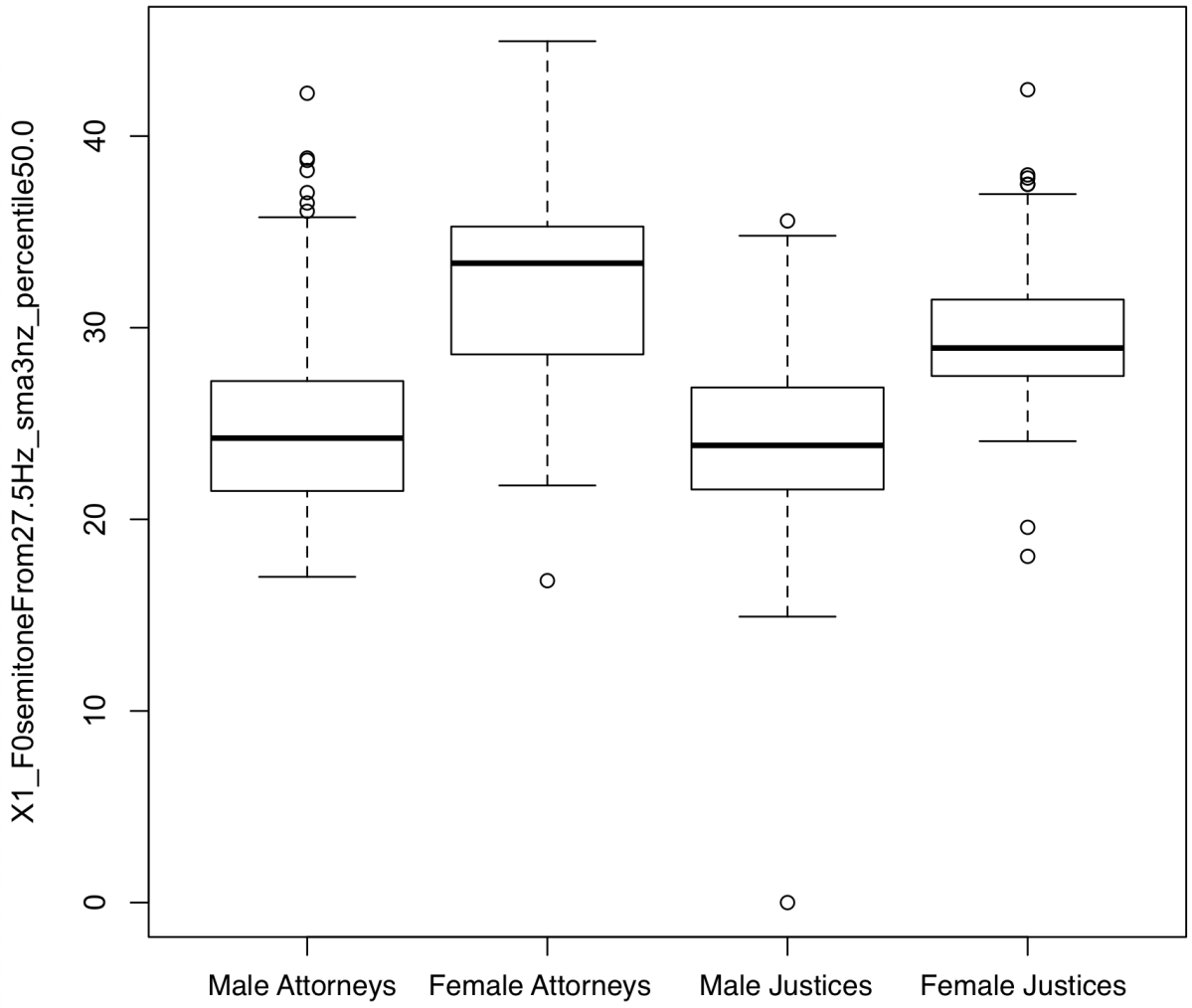


Figure 5.1: 50th Percentile F0 by Gender and Role

higher overall accuracy by improving accuracy in middling labels. Among regression models, the MLP performs poorly, and could benefit from a larger training set. Figure 5.3 shows the results of using the raw (not normalized), continuous labels to train regression-based models. Gender and role are included in each feature set. The SVR trained on Prosody features gives the highest results.

	eGeMAPS		Prosody	
Model	Competitive	Cooperative	Competitive	Cooperative
SVC with gender and role features	0.636	0.551	0.687	0.561
SVC	0.623	0.604	0.653	0.579
RF with gender and role features	0.617	0.593	0.640	0.611
RF	0.654	0.630	0.628	0.587

Table 5.2: Range in F_1 score using two raw feature sets with and without gender and role labels

	eGeMAPS		Prosody		emobase		All Sets Combined	
Model	R^2	Pearson	R^2	Pearson	R^2	Pearson	R^2	Pearson
SVR	0.344	0.595	0.376	0.618	0.232	0.509	0.281	0.544
MLP	0.241	0.530	0.241	0.530	0.185	0.511	0.202	0.539

Table 5.3: R^2 and Pearson Correlation results by Feature Set for Regression Models

The largest differences in performance come from the different feature sets. The small set of silence, amplitude, and pitch features which were hypothesized by previous work to be defining features of competitive and cooperative exchanges resulted in low scores. The eGeMAPs feature set, which is often praised for its limited number of features and relatively high performance, produced the highest classification scores for predicting tertile labels. The Prosody feature set produced the best regression on the continuous labels and the quintile labels, indicating that it may produce more fine-grained predictions. The emobase feature set, despite its size, produced inferior results to the two much smaller `OpenSmile` feature sets. Surprisingly, a combination of all the feature sets produced modestly inferior predictions, with the exception of predictions of cooperative exchanges. Figures 5.4 - 5.6 show the results of the four types of models with the four highest-performing feature sets in predicting the

three types of labels. Each of these models includes features with gender and role, and none of the features are normalized.

Model	eGeMAPS	Prosody	emobase	All Sets Combined
Tertile Labels				
SVC	0.542	0.549	0.521	0.507
Baseline	0.458	0.225	0.423	0.423
RF	0.549	0.521	0.542	0.565
Baseline	0.437	0.563	0.239	0.408
Quintile Labels				
SVC	0.308	0.429	0.435	0.366
Baseline	0.211	0.141	0.197	0.155
RF	0.408	0.401	0.423	0.345
Baseline	0.261	0.063	0.169	.0162

Table 5.4: Micro-Averaged F_1 Labels by Feature Set for Classifiers

5.3.1 Discussion

Considering that the granular level of labels produced by annotators may be relatively noisy, the most successful classifier of unseen data uses tertile-category labels for Prosody features to train an RF model. This model captures the peaks in label distribution and predicts both the competitive and cooperative categories with relative success, and has an average accuracy across all labels of 0.528 which is within the range of state-of-the-art emotion recognition models. Though other models have higher accuracy across the entire set of labels, this model is better at predicting the non-neutral labels, a feat not captured by the average accuracy.

However, depending on the task at hand, various models are better at predicting different labels, suggesting that the spectrum between competitive and cooperative turns may be imbalanced, or not a spectrum at all.

Model	eGeMAPS	Prosody	emobase	All Sets Combined
Tertile Labels				
SVC	0.636	0.687	0.610	0.483
RF	0.617	0.640	0.606	0.435
Quintile Labels				
SVC	0.515	0.536	0.563	0.476
RF	0.603	0.563	0.567	0.391

Table 5.5: F_1 of Most Competitive Quantile by Feature Set for Classifiers

Model	eGeMAPS	Prosody	emobase	All Sets Combined
Tertile Labels				
SVC	0.551	0.561	0.574	0.667
RF	0.593	0.611	0.590	0.650
Quintile Labels				
SVC	0.412	0.429	0.435	0.508
RF	0.478	0.507	0.528	0.586

Table 5.6: F_1 of Most Cooperative Quantile by Feature Set for Classifiers

For example, models trained with quintile-category labels can predict extreme categories with high precision and recall, but struggle significantly with middling categories (Table 5.8). The tertile-category labels capture some of the middle categories, but with a trade-off for precision (Table 5.7).

This section aims to show the breadth of ways this corpus can be used for automatically labeling turns. As such, any researcher using the corpus to predict turns should adjust the models, label categories, and feature sets according to the task.

	0	1	2	Precision	Recall	F_1
0	[35]	6	4	0.565	0.778	0.654
1	18	[9]	19	0.391	0.196	0.261
2	9	8	[34]	0.596	0.667	0.630

Table 5.7: Confusion Matrix for eGeMAPS Random Forest Classifier with Tertile Labels

	0	1	2	3	4	Precision	Recall	F_1
0	[17]	5	2	0	0	0.405	0.708	0.515
1	8	[13]	2	3	3	0.433	0.448	0.441
2	6	4	[2]	6	9	0.222	0.074	0.111
3	8	5	2	[8]	12	0.421	0.229	0.296
4	3	3	1	2	[18]	0.429	0.667	0.522

Table 5.8: Confusion Matrix for eGeMAPS Random Forest Classifier with Quintile Labels

Chapter 6

CONCLUSION

This study introduces a corpus of segments of speech from U.S. Supreme Court oral arguments that include a turn-change between speakers. The segments, annotated by legal practitioners for competitiveness and cooperativeness, provide insight in the ways that justices and attorneys speak with one another in the unique speech setting of an oral argument. I demonstrate that classifiers and regressors trained only on phonetic and acoustic features extracted from the audio segments can achieve a level of predictive accuracy comparable to state-of-the-art emotion recognition research.

6.1 Continued Work

In depth studies of gender bias and inequality are critical to the oversight of an institution as influential as the Supreme Court.[1] There is demand in the social sciences for even broader analysis; within one month of the this writing, several cross-cutting studies have criticized increasing bias in the Supreme Court and federal appeals courts, especially in regards to poverty and race (Ruiz et al., 2020; Cohen, 2020). While the models presented in this study analyze linguistic trends in relation to gender, the labeled corpus could be integrated with other demographic or content-related information to provide a fine-grained analysis of intersectional fields.

An expansion of this corpus to include more instances of speech from justices and a larger number of unique attorneys would help prevent over-fitting of classification and regression models to the relatively small training set. Any future annotating of segments should require

[1]Similar studies of other institutions, such as using police dash-camera footage to analyzed bias, show evidence of a growing field (Voigt et al., 2017).

an increased training period for annotators and avoid easy-to-error phone surveys, to raise the accuracy of labels as measured by inter-annotator agreement.

There are other strategies for feature extraction which may lead to improvement. Text-based n-grams and syntax trees could raise the prediction abilities of classifiers, as an unfinished sentence is one of the defining qualities given to annotators for a competitive exchange. Taking measurements at some frame-per-sample rate, instead of summarizing the measurements for the entire audio segment, could provide a more fine-grained picture of trends related to labels. Features could also be extracted from the entire oral argument, and not just the short window around turns. For example, the frequency of pauses per speaker could be extracted for each speaker in an oral argument. Normalizing features across all of one speaker's speech in all oral arguments, instead of just across all speech segments in the corpus, could lead to more effective speaker-based normalization.

In addition to an expanded training set or other features, classification could be improved by altering the training and evaluation sets or training other classifiers. For example, cross-validation leaving one speaker out for each experiment could help improve speaker-independent predictions. Semi-supervised or unsupervised methods, such as bootstrapping the labeled training data with unlabeled turns, or using clusters of segments by label and features to label unseen segments, could provide higher accuracy predictions.

With improved predictive models, a larger set of turn-changes across all Supreme Court oral argument recordings and possibly other court recordings could provide fodder for future statistical social science studies of speech trends in the U.S. judicial system.

BIBLIOGRAPHY

Albornoz, E., Milone, D., and Rufiner, H. (2011). Spoken Emotion Recognition Using Hierarchical Classifiers. *Computer Speech & Language*, 25:556–570.

Ambady, N., Krabbenhoft, M. A., and Hogan, D. (2006). The 30-Sec Sale: Using Thin-Slice Judgments to Evaluate Sales Effectiveness. *Journal of Consumer Psychology*, 16:4–13.

Ambady, N. and Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, 64:431–441.

American Bar Association (2019). ABA National Lawyer Population Survey: 10-Year Trend in Lawyer Demographics.

Amidei, J., Piwek, P., and Willis, A. (2019). Agreement is Overrated: A Plea for Correlation to Assess Human Evaluation Reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, Tokyo, Japan.

Artstein, R. (2017). Inter-annotator Agreement. In Ide N., P. J., editor, *Handbook of Linguistic Annotation*, chapter 11, pages 297–313. Springer, Dordrecht.

Awad M., K. R. (2015). Support Vector Regression. In: Efficient Learning Machines. *Efficient Learning Machines*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bias (2003). *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 6 edition. The International Statistical Institute.

Bramsen, P., Escobar-Molano, M., Patel, A., and Alonso, R. (2011). Extracting Social Power Relationships from Natural Language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA. Association for Computational Linguistics.

Chira, S. (2017). The Universal Phenomenon of Men Interrupting Women. *The New York Times*.

Cohen, A. (2020). *Supreme Inequality: The Supreme Court's Fifty-Year Battle For a More Unjust America*. Penguin Random House, New York.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, J. (1968). Weighted Kappa: Nominal Scale Agreement Provision For Scaled Disagreement or Partial Credit. *Psychological Bulletin*, 70(4):213–220.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47:1–12.

Delaney, R. (2000). Dialect Map of American English. http://robertspage.com/dialects.html.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Eyben, F. and Schuller, B. (2015). OpenSMILE:): The Munich Open-Source Large-Scale Multimedia Feature Extractor. *SIGMultimedia Rec.*, 6(4):4–13.

FFmpeg Developers (2016). ffmpeg tool. http://ffmpeg.org/. Software.

Gallup (2019). Party Affiliation. https://news.gallup.com/poll/15370/party-affiliation.aspx.

Goldberg, J. A. (1990). Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power and rapport-oriented acts. *Journal of Pragmatics*, 14:883–903.

Gorisch, J., Wells, B., and Brown, G. (2012). Pitch Contour Matching and Interactional Alignment Across Turns: An Acoustic Investigation. *Language and Speech*, 55:57–76.

Jacobi, T. and Schweers, D. (2017). Justice, Interrupted: The Effect of Gender, Ideology and Seniority at Supreme Court Oral Arguments. *Virginia Law Review*, 103(7):1379–1496.

Jiang, W., Wang, Z., Jin, J., Han, X., and Li, C. (2019). Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network. *Sensors*.

Kahler v. Kansas (Oct. 7, 2019). *No. 18-6135*. United States Supreme Court.

Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology*, pages 277–360. SAGE Publications, 4 edition. University of Pennsylvania.

Kurtić, E., Brown, G., and Wells, B. (2010). Resources for Turn Competition in Overlap in Multi-Party Conversations: Speech Rate, Pausing and Duration. pages 2550–2553. 11th Annual Conference of the International Speech Communication Association (INTERSPEECH).

Kurtić, E., Brown, G., and Wells, B. (2013). Resources for Turn Competition in Overlapping Talk. *Speech Communication*, 55:721–743.

Laskowski, K. (2010). Modeling norms of turn-taking in multi-party conversation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 999–1008. Association for Computational Linguistics.

Liaw, A. and Wiener, M. (2002). Classification and Regression by RandomForest. *R News*, 2:18–22.

Ling, G., Mollaun, P., and Xi, X. (2014). A Study on the Impact of Fatigue on Human Raters When Scoring Speaking Responses. *Language Testing*, 31(4):479–499.

Mitchell v. Wisconsin (Jan. 21, 2019). *No. 18-6210*. United States Supreme Court.

Noroozi, F., Sapiński, T., Kamińska, D., and Anbarjafari, G. (2017). Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pemberton, C., McCormack, P., and Russell, A. (1998). Have women's voices lowered across time? a cross sectional study of australian women's voices. *Journal of voice : official journal of the Voice Foundation*, 12:208–13.

Pettarin, A. (2017). Aeneas. https://github.com/readbeyond/aeneas. Software.

Prabhakaran, V. and Rambow, O. (2014). Predicting Power Relations Between Participants in Written Dialog from a Single Thread. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344, Baltimore, Maryland. Association for Computational Linguistics.

Richardson, L. (2007). Beautiful Soup Documentation.

Robinson, K. S. and Rubin, J. S. (Jan. 30, 2019). Women Argue Only a Fraction of Supreme Court Cases.

Ruiz, R. R., Gebeloff, R., Eder, S., and Protess, B. (2020). A Conservative Agenda Unleashed on the Federal Courts. *The New York Times*.

S. K. Pal, S. M. (1992). Multilayer Perceptron, Fuzzy Sets, and Classification. *IEEE Transactions on Neural Networks*, 3:683–697.

Sound eXchange, S. (2015). http//sox.sourceforge.net/.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101.

Tannen, D. (1994). *Gender and Discourse*. New York: Oxford University Press.

Tharwat, A. (2018). Classification Assessment Methods. *Applied Computing and Informatics*.

The Oyez Project (2019). About Oyez. https://www.oyez.org/about.

The Supreme Court of the United States (2019). Transcripts and Recordings of Oral Arguments.

Truong, K. (2013). Classification of Cooperative and Competitive Overlaps in Speech Using Cues from the Context, Overlapper, and Overlappee. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1404–1408.

United States Census Bureau. Quick facts, United States, 2013-2017. https://www.census.gov/quickfacts/fact/table/US/LFE046217.

Virginia House of Delegates v. Bethune-Hill (Mar. 18, 2019). *No. 18-281*. United States Supreme Court.

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., and Eberhardt, J. L. (2017). Language from police body camera

footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Walsh, M. (Aug. 1, 2018). Number of Women Arguing Before the Supreme Court has Fallen off Steeply. *American Bar Association Journal*.

Wang, Z., Zechner, K., and Sun, Y. (2018). Monitoring the Performance of Human and Automated Scores for Spoken Responses. *Language Testing*, 35(1):101–120.

Wang D., H. G. (2008). Cocktail Party Processing. *Computational Intelligence: Research Frontiers. Lecture Notes in Computer Science.*, 5050.

Washington State Dept. of Licensing v. Cougar Den Inc. (Oct. 30, 2018). *No. 16-1498*. United States Supreme Court.

Weston, J., W. C. (1998). Multi-class Support Vector Machines. *Royal Holloway University of London Department of Computer Science*.

Wichmann, A. and Caspers, J. (2001). Melodic cues to turn-taking in English: Evidence from perception. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Yang, L. (1996). Interruptions and Intonation. In *Proceeding of Fourth International Conference on Spoken Language Processing*.

Yang, L. (2003). *Current and New Directions in Discourse and Dialogue. Text, Speech and Language Technology*, volume 22, chapter Visualizing Spoken Discourse. Springer.

Yao, Y. (2007). Closure Duration and VOT of Word-Initial Voiceless Plosives in English in Spontaneous Connected Speech. *UC Berkeley Phonology Lab Annual Report*, pages 183–225.

Young, S., Gunnar, E., Mark, G., H. T., and Kershaw, D. (2015). The HTK Book Version 3.5 Alpha.

Zhang, M. (2013). Contrasting Automated and Human Scoring of Essays. *ETS R and D Connections*.

Appendix A

ANNOTATOR DEMOGRAPHIC QUESTIONNAIRE

This appendix details the responses of participants to questions in the demographic questionnaire. I grouped annotator responses into overarching categories to create the labels listed in Table 3.1.

- Profession
  - Student: “2L”, “Student, 3L”, “3L”, “Law student, 2L”, “Third-year law student”, “2nd year law student”, “Law Student 2L”, “3L law student”, “2L Law Student”, “Law student- second year”, “3L Law Student”, “2L in law school”, “Second year law student”, ”U of Mich. Law School”, “Law Student (2L)”, “’Student, 2L Law/First Year Master of Public Policy”, “Law student (3L)”
  - Attorney: “Attorney”, “Lawyer”
  - Other: “Paralegal”, “Directory, Licensing and Compliance”, “Manager”, “Magistrate judge”
- Gender
  - Female: “Female”, “F”, “Cis woman”, “FEMALE”, “f”
  - Male: “Male”, “M”, “Man”, “m”
  - Not listed
- Race

  - White: “White”, “White, Asian, Hispanic”, “White/ Caucasian”, “Black/White”, “Wh (jew)”, “Caucasian”, “W”, “WHITE”, “Mixed: South Asian/White”, “White/jewish”
  - African-American/Black: “Black/White”, “African American/Black”, “African Amer”, “Black”, “African American”, “African-Ameican”
  - Asian: “Chinese, “White, Asian, Hispanic”, “South Asian”, “Mixed: South Asian/White”, “Asian”
  - Hispanic: “White, Asian, Hispanic”
  - Not listed

- Politics
  - Left: “Dem”, “Democrat”, “Democratic”, “Democratic Party”, “Democrat (reluctantly)”, “Democratz”, “DSA”, “D”, “DEMOCRATIC SOCIALIST”, “Independent with democratic leanings”, “Democratic party”, “Democrat/progressive”, “democrats”, “No, but generally lean left”
  - Right: “Republican”
  - Other: “ Libertarian”, “Independent”, “green” Not listed, “N/A”, “None”, or “No”

- Dialect/Sociolect: Annotators were given the option to pick one or more varieties of English from the following map Delaney (2000), as well as the options of “African American Vernacular English”, “Chicano English”, “New York Latino English”, “I do not speak American English”, and “I speak American English but do not identify with any of these dialects”.

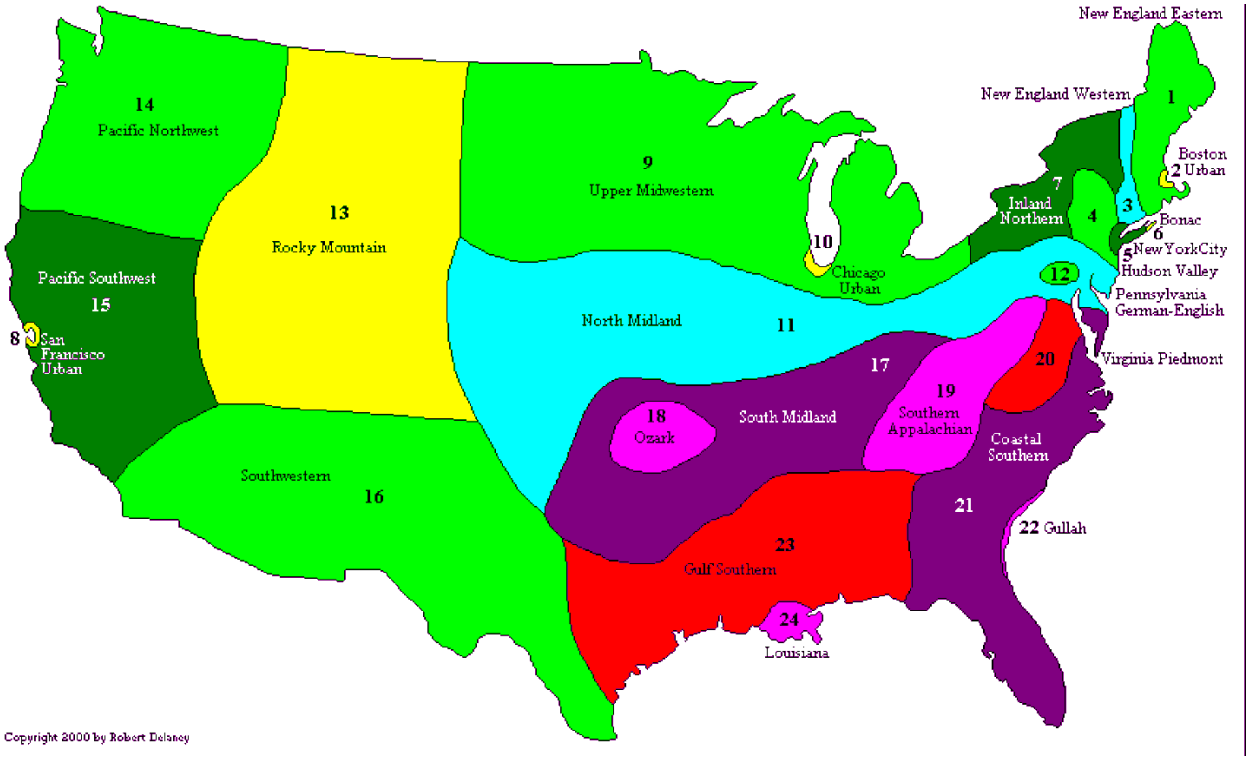
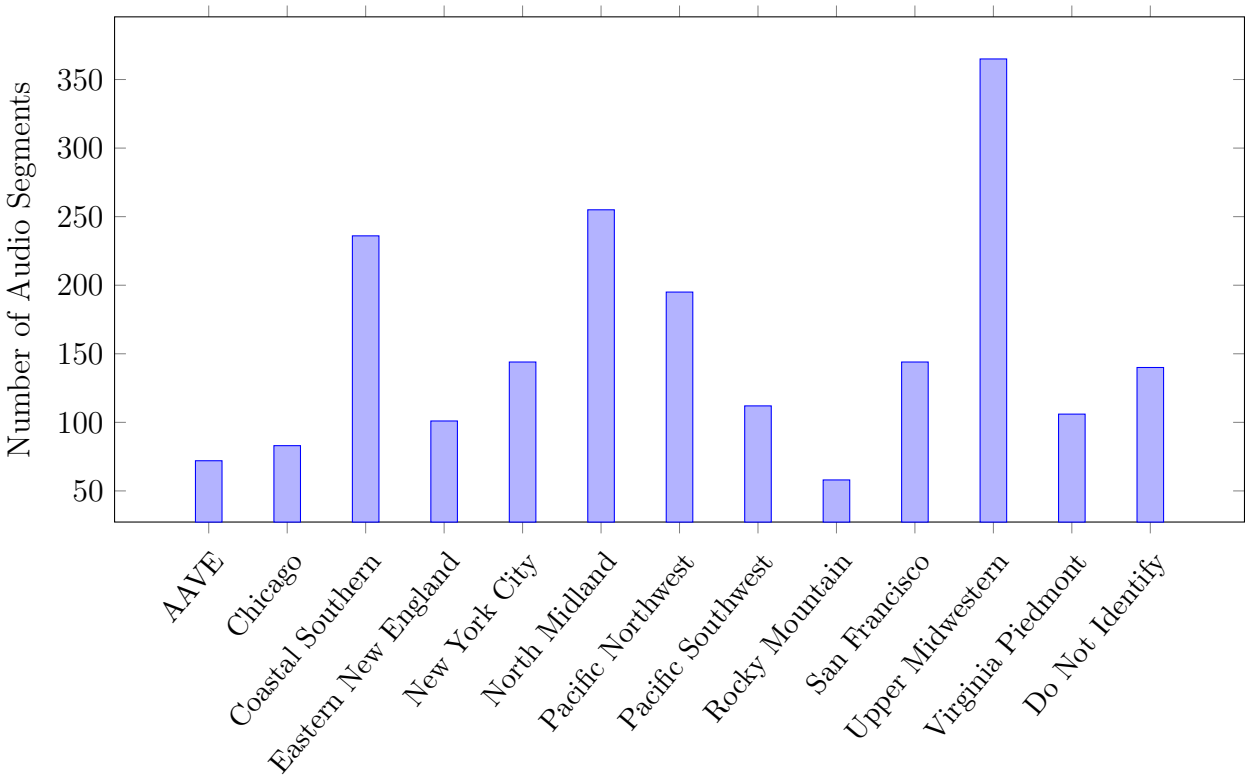


Figure A.1: Map of American English Dialects

Figure A.2: The results of the dialect/sociolect question are listed below, with number of times a speaker of that dialect annotated an audio segment. Only categories which label more than 50 segments are listed. Some annotators listed more than one answer.



Appendix B

SPEAKERS

A mapping of speaker to gender and role.

Speaker	Gender	Role
Adams	F	Attorney
Alito	M	Justice
Breyer	M	Justice
Clement	F	Attorney
Crouse	M	Attorney
Elias	F	Attorney
Ginsburg	F	Justice
Gorsuch	M	Justice
Heytens	F	Attorney
Hinkel	M	Attorney
Jurss	F	Attorney
Kagan	F	Justice
Kavanaugh	M	Justice
Prelogar	F	Attorney
Purcell	M	Attorney
Ratner	M	Attorney
Roberts	M	Justice
Schrup	F	Attorney
Sotomayor	F	Justice
Thomas	M	Justice
Unikowsky	M	Attorney