The Sound of Politics:

How Politicians Pitch Conflict, Representation, and Power

Mathias Rask

The Sound of Politics:
How Politicians Pitch Conflict, Representation, and Power

PhD Dissertation

Politica

©Forlaget Politica and the author 2025

ISBN: 978-87-7335-338-7

Cover: Svend Siune
Print: Fællestrykkeriet, Aarhus University
Layout: Annette Bruun Andersen

Submitted January 10, 2025
The public defense takes place March 28, 2025
Published March 2025

Forlaget Politica
c/o Department of Political Science
Aarhus BSS, Aarhus University
Bartholins Allé 7
DK-8000 Aarhus C
Denmark

Contents

Acknowledgements

I owe the greatest gratitude to several people who have supported me either directly or indirectly throughout the time writing this dissertation.

First, I would like to thank all my colleagues in the Department of Political Science at Aarhus University and in the Section of Political Behavior and Institutions. Thank you for engaging with my research ideas and research papers. I have learned a lot about specific articles, but even more valuable is the insight about what it takes and means to do thoughtful, innovative, and well-crafted research. The support, openness, and collegiality at the department have made writing this dissertation easier.

Second, I have had the pleasure of being supervised by an impressive set of people. I am truly and utterly grateful to Helene Helboe Pedersen for the time and dedication you have provided me over the last four years. You often say its your job, but your time and dedication go beyond what can be expected from the job description. You have always been available despite an incredibly busy calendar and provided me with comments on drafts within 24 hours, sometimes even a few hours. What goes surely beyond the job description is your human way of being. Your positivity about my project and my academic talent have meant more than I have probably expressed to you. I also thank Matt Loftis for his valuable insights and perspectives. The competencies of you and Helene were a match made in heaven before you moved to the public sector. I also owe a big thanks to Mathias Tromborg, who stepped in for Matt late in the process; the amount of time and effort you have invested to getting aligned with the state of the project has been admirable. The feedback you have provided has been indispensable.

Several other people who might not even know their influence should also be mentioned. Martin Vinæs Larsen and Martin Bisgaard, your willingness to share your knowledge with young researchers is admirable and has inspired me during the writing of the dissertation. You don't have to do so, but you *do*. Benjamin Egerod and Frederik Hjorth (also coauthor) should also be mentioned for the time during my bachelor and master degree at the University of Copenhagen. It was your courses and way of teaching that encouraged me to work in computational social science. Finally, Malte Dahl should be acknowledged for being the first to motivate me to pursue a Ph.D. during the fourth semester of my bachelor. Malte was my teacher in what was then called Methods II.

Gratitude also goes to my family. Mor, Henriette, and far, Henrik, you have been the most supportive parents one could ever imagine throughout our child- and adulthood. Your previous and continuing support has been invaluable to where Mikkel and I are in our lives. Mom, you encouraged and kept me focused on my school during my teenage years, where cycling took the majority of my awaken time. Thank you, Mikkel, for never complaining about all the time devoted to, first, the cycling career, and second, the praise I have received from pursuing a Ph.D. For that, I truly admire you. Mormor and Morfar, thank you for always being my biggest fan. There are no limits to your love.

The greatest gratitude of mine goes to Katrine, my wonderful and supportive wife. I am not aware of anyone willing to sacrifice the goals oneself to the same extent as you. Without your support, there would certainly not have been this dissertation – or our two lovely daughters, Ingrid and Edith, for that sake.

Preface

This report summarizes the PhD dissertation, *The Sound of Politics: How Politicians Pitch Conflict, Representation, and Power*. The dissertation was written as the conclusion of my PhD project at the Department of Political Science, Aarhus University (DK). The dissertation consists of this summary report and four research papers, three single-authored and one co-authored, presented below. The purpose of the report is to concisely present the findings and contributions of each individual paper and how they relate and interact. Furthermore, the report motivates the overall claim put forward in the dissertation, provides answer(s) to the research questions, and gives an overview of the data and methods used to produce the findings.

Paper A: Automated Annotation of Political Speech Recordings. *Working Paper*.

Paper B: Partisan Conflict in Nonverbal Communication. Coauthored with Frederik Hjorth. Revised and resubmitted to *Political Science Research and Methods*.

Paper C: Committed but Constrained: Explaining Why the Descriptive-to-Substantive Representation Link Weakens Over Time. *Working Paper*.

Paper D: When They Go High, We Go Low: Rhetorical Rewards of Governing. *Working Paper*.

Chapter 1
Introduction

THE SOUND OF THE HUMAN VOICE conveys meaning and information beyond the words themselves.[1] Nonverbal components of speech, including prosodic (e.g., pitch), spectral (e.g., formants), and general voice features (e.g., the fundamental voice frequency), transmit signals of a speaker's character traits, emotions, and sociodemographic characteristics, independent of verbal content.[2] From friendly talks to political debates, the way humans speak acts as a dynamic signaling tool that shapes social and political life by affecting how a speaker and the message are perceived by the audience.

Given that speech is essential to human communication, it is not surprising that political scientists have studied political speeches across nearly every subfield, from legislative studies to international relations. Politics occurs and unfolds in words (Grimmer & Stewart, 2013, p. 267), particularly the spoken, which serve as the primary medium through which politics is communicated to voters and by which politics operates and functions. This is evident by the fact that speech-making activities

[1] I use the term "sound" to refer to human perception of acoustic sound waves unless otherwise specified. In Chapter 4, I provide a physical definition of "sound" and discuss how it relates to its perceptual definition.

[2] I use the term "nonverbal speech" to refer to vocalized information conveyed in a speech beyond words. This term excludes body language and facial expressions, focusing exclusively on the sound produced in speech. "Vocal cues" or "vocal speech" are used interchangeably with "nonverbal speech" throughout the dissertation.

involving spoken words are central to most, if not all types of political institutions (Knox & Lucas, 2021, p. 649). On the campaign trail, candidates face each other in debates, attempting to persuade voters through rhetoric. In national parliaments, the elected candidates debate and contest legislation and policy on the floor. Furthermore, country leaders debate and negotiate international disputes and conflicts in the UN Security Council. In each of these cases, the raw data we study is spoken words.

The importance of the sound of human voices in transmitting signals about a speaker is not unique to politics but has long been recognized by evolutionists, rhetoricians, and psychologists. These studies all show that humans use and modulate their voice in social and human interactions. Starting with Darwin (1872), evolutionists view nonverbal speech as integral to social dynamics and ultimately natural selection in both humans (Puts et al., 2006) and non-human primates (Morton, 1977; Seyfarth et al., 1980). Drawing from evolutionary theories, psychologists see nonverbal speech as a channel for expressing emotions and social information (Bachorowski & Owren, 1995; Scherer, 2018; Scherer et al., 1984; Walton & Orlikoff, 1994). Rhetoricians, from Aristotle to Cicero and Quintilian, viewed nonverbal speech as essential in delivering efficient emotional appeals to persuade and connect with the audience (Guyer et al., 2021; Scherer, 2018).

With these insights in mind, it becomes evident why voters respond strongly to the sound of a politician's voice as a signaling tool. A series of studies show that a candidate's average vocal pitch, i.e., the perceived "highness" or "lowness" of a voice, shapes voting behavior (B. Banai et al., 2018; I. P. Banai et al., 2017; Gregory Jr & Gallagher, 2002; Klofstad, 2016; Klofstad et al., 2012, 2016; Tigue et al., 2012) and perceptions of candidates (Klofstad et al., 2015; Podesva et al., 2015; Surawski & Ossoff, 2006). Voters consistently favor candidates with lower-pitched voices (Anderson & Klofstad, 2012), likely interpreting this sound as a sign of competence, dominance, and composure, traits positively valued in political leaders (Laustsen et al., 2015). This extends beyond a candidate's average pitch. Voters' perceptions and preferences are also shaped by other specific prosodic features such as voice modulation (i.e., voice variance), the rate of speech, and the general sound of a candidate's voice (Damann et al., 2024).

Considering the close ties between spoken words and politics, it is perhaps surprising that political scientists study speeches mostly in a "speech unplugged" manner. Quantitative studies tend to study political speeches almost exclusively through transcripts, even when the corresponding audio recordings are publicly available (Knox & Lucas, 2021, p. 651-52). Focusing solely on text strips away information about the sound of politics (Cochrane et al., 2022). While the text of speech contains a wealth of relevant information in itself, treating spoken words as though they were written inevitably results in a loss of insights into a politician's traits, emotions, and sociodemographic characteristics (E. Ash et al., 2024; Damann et al., 2024; Rheault & Borwein, 2019; Zárate et al., 2024).

More recently, a new line of work has started using audio recordings to study political speeches, plugging speech back in. This literature has focused mostly on how nonverbal speech conveys the emotions of a political actor, such as a legislator or a judge, and how this relates to the link between descriptive and substantive representation (Dietrich, Hayes, & O'Brien, 2019; Rittmann, 2024) and the intensity of attitudes (Dietrich, Enos, & Sen, 2019; Knox & Lucas, 2021). Others have studied how politicians use their voice to shape perceptions of their ascriptive and valence attributes to monitor the representative-constituency linkage (Neumann, 2019). Finally, scholars have explored how music can be used to classify the mood of TV ads

(Tarr et al., 2023) and how text and audio can be aligned at the word level (Arnold & Küpfer, 2024). Although the use of audio recordings in empirical research remains scarce, this set of studies highlights the promise of using audio data to study politics (Rheault & Borwein, 2022). Specifically, this work provides the first indications that not only do voters react to the sound of politicians' voices as a signaling tool, politicians use their voice – intentionally or not – systematically and predictably.

1.1 Existing Challenges

This literature forms the basis of the overall claim advanced and recited throughout the report: *Politics is difficult to understand in silence but benefits from hearing the sound of politicians as they speak*. Evaluating this claim is not without challenges. Even though audio data receives growing attention, its integration into empirical research is still in its infancy. Where a search on "text as data" in politics yields nearly 9,000 hits on Google Scholar in October 2024, the equivalent search with "audio as data" yields 34 matches.[3] Compared to text data, which is firmly established in virtually any subfield of political science, audio data still lacks a systematic research agenda to fully unlock its promises in studying politics (Rheault & Borwein, 2022).

What explains this lack of attention? One possible reason is that working with audio recordings is computationally challenging. Not only does audio data require a different set of preprocessing techniques than text, it comes with a premium computational cost. Preprocessing speech audio requires domain-specific knowledge about preprocessing, handling, and manipulation of

[3] The Google Scholar (https://scholar.google.com/) search was conducted on October 29, 2024, using the strings *politic* AND "text as data"* and *politic* AND "audio as data"*. In comparison, image and video data return 267 matches using *politic* AND "image as data"* and 304 matches using *politic* AND "video as data"*, respectively.

digital signals, a field called digital signal processing (Rabiner & Schafer, 2011). Working with audio also presents a challenge in terms of computational costs and even basic storage issues. As noted by Rheault and Borwein (2022), a 30-second speech, on average, corresponds to a transcript of 75 words, which amounts to 150 bytes of data. In comparison, the corresponding audio recording of that same speech amounts to approximately 1.3 million bytes, a factor 9,000 times larger.[4] As audio archives consist of hundreds and thousands of hours of political speech, the size of recordings quickly becomes computationally costly, even when researchers preprocess the signal in a computational-friendly manner.[5]

Besides the computational costs and domain-specific knowledge needed to work with audio, integrating audio recordings into empirical research is challenged by two additional issues: one methodological and one theoretical. First, using audio recordings is *methodologically challenging*. Using audio of political speeches to measure nonverbal speech features requires annotations such as timestamps and speaker identities that allow segmenting the recording into distinct units of analysis (e.g., the speech or utterance level) before measurement. However, audio recordings are rarely as well annotated as transcripts of political speeches and are often inaccurate, incomplete, or entirely lacking. The absence of annotated audio archives poses a significant challenge to the use of audio in political science.

[4] A byte represents the smallest unit of binary digital information used to represent data on computers, typically encoded in eight bits (8-bits) consisting of 0s and 1s. One byte can take on $2^8 = 256$ different values. In this encoding, 1.3 million bytes amount to approximately 1.24 megabytes (MB).

[5] The size of audio recordings can be partially controlled by manipulating the sampling rate, which controls the number of samples used to represent the audio per second. The sampling rate can be lowered to ease computational and storage costs but only to a degree where frequencies are not aliased (Giannakopoulos & Pikrakis, 2014, p. 43-44). A typical choice is a sampling rate of 16,000 hertz (Hz), yielding 16K samples per second of audio. I return to this in Chapter 4.

Even disregarding the computational challenge, the lack of well-annotated audio impedes the integration of recordings into the methodological toolbox of political scientists.

Second, applying audio recordings is *theoretically challenging*. Measurement of nonverbal speech characteristics as an outcome (i.e., “a dependent variable”) and not as a predictor (i.e., “an independent variable”) requires a broader theoretical framework linking the work of evolutionists, rhetoricians, and psychologists to core political science concepts and theories to generate falsifiable observable implications of the sound of politics to avoid the “garden of forking paths” (Gelman & Loken, 2013). This pitfall arises partially from the challenge of linking single auditory features, such as pitch or formants, to single politically relevant behaviors, such as personal issue commitments (Dietrich, Hayes, & O’Brien, 2019) or judicial skepticism (Knox & Lucas, 2021) and partially by linking variation in auditory features to its causes.

1.2 Research Questions and Claims

The purpose of this dissertation is to provide answers, if not fully, then partially, to each of these challenges in order to evaluate the overall claim of the dissertation that politics has different sounds. However, the challenges are inherently interdependent. Answering the theoretical challenge makes little sense if the computational and methodological challenge persists. Based on this, I ask two related research questions that enable us to shed light on the sound of politics:

Research Questions (RQs)

RQ1: *To what extent can political speech recordings be automatically annotated?*
RQ2: *When do politicians change the sound of their voices?*

The first research question (RQ1) speaks to the methodological challenges of using audio recordings in empirical research. I answer this question by developing an automated pipeline based on pre-trained deep learning models that allow annotation of audio recordings of political speeches with near human-level accuracy without using any prior manually annotated data. This provides the foundations for integrating audio recordings at scale and for empirically testing theoretical claims.

The second research question (RQ2) addresses the theoretical challenge of using the sound of politicians' voices as an outcome and provides empirical answers. The "when" refers to the notion that politicians likely change how they speak depending on their political roles. This suggests a causal relationship. When the role changes, so does the sound of politicians' voices. The term "role" is broadly used to denote that politicians serve different functions and tasks, potentially at the same time.

In the dissertation, I focus on three different roles, *partisan*, *representative*, and *governing*, which are fundamental in democracies and tie into classic political science concepts about conflict, representation, and power. In their partisan role, politicians often highlight and emphasize partisan divisions to clarify the ideological stakes and rally their base. This is essentially about *conflict*. In their representative role, politicians advocate for specific constituencies or demographic groups, whether geographically defined (e.g., districts or states) or based on particular interests (e.g., minority communities, certain social groups), build trust with, and communicate responsiveness to the represented. This is essentially about *representation*. In their governing role, politicians introduce policy proposals, articulate policy decisions, and project their authority. This is essentially about *power*. The claim of the dissertation is that these roles causally change the sound of politicians' voices. The actual identification of this effect hinges on the design and statistical modeling, but the underlying relationship is theorized to be causal regardless of the identification. For example, when partisan conflict intensifies, this is hypothesized to be reflected in a more conflictual sound.

The key concept of the dissertation is "sound". By sound, I refer to the distinctive vocal quality associated with a speaker's intentions or roles as it is perceived.[6] Throughout the dissertation and its individual papers, I use a single nonverbal speech feature to characterize the general sound of a speech: speech-level average vocal pitch. While no feature can fully characterize the sounds of political conflict, representation, and power alone, vocal pitch has proven to be a surprisingly robust and accurate signal of a speaker's emotional arousal in both non-political (Bänziger & Scherer, 2005) and political settings (Dietrich, Hayes, & O'Brien, 2019; Rittmann, 2024) and trait perceptions (e.g., Tigue et al., 2012).[7] I use this to construct a measure of a politician's emotional arousal ranging from high composure to high activation based on speaker-standardized speech-level average vocal pitch. Higher values indicate greater arousal – low composure and high activation – and lower values indicate lower arousal – high composure and low activation – and operate within politicians by virtue of standardization. This means that values denote positive or negative changes from a speaker's own baseline, and the interpretation of this variation hinges

[6] In Chapter 4, I also provide a physical definition of sound.

[7] Vocal pitch is also surprisingly robust to changes in recording device, microphone quality, and so forth, making it ideal to study over longer periods (Vainio et al., 2023).

upon the role of the politician giving the speech. This scale dynamically captures the relationship between the role and the sound of a politician's voice in a given speech. In the role as partisan, the politician *raises their pitch*, transmitting sounds of agitation and disagreement. In the role as representative, the politician *raises their pitch*, transmitting sounds of engagement. In the role of governing, the politician *lowers their pitch*, transmitting sounds of composure.

1.3 Case Selection and Empirical Setting

I focus on when politicians change their vocal pitch when giving speeches in *legislative debates*. Speeches in legislative debates serve an important function in legislatures around the world as members of the legislature "present different policies to the public by setting them out in legislative debates" (Bäck & Debus, 2024, p. 249), mostly to highlight already known policy positions and highlight partisan differences (Laver, 2021). Legislative speeches are central to the tasks of partisans, representatives, and those who govern, offering a direct and indirect connection to voters. I study the sound of legislative speeches in a single empirical setting; the Danish Parliament. While legislative debates are widely studied in political science (Bäck et al., 2021), research on speech in the Folketing is scarce (Willumsen, 2021). Still, as I will argue in Chapter 5, the Folketing offers both theoretical and methodological advantages that make it an ideal case to study the sound of politics. I compile a multimodal corpus containing text-audio data of all plenary speeches in the Folketing given from 2000-2022. This includes six national elections, 28 parliamentary terms, and 2,186 debates with 850,357 speeches. The corpus is compiled using the method developed in Paper A and used in the empirical articles (B, C, and D).

1.4 Outline of the Dissertation

The dissertation contains this summary report and four self-contained articles, three individually authored and one co-authored. The articles are listed in Preface and illustrated visually in Figure 1.1. Paper A (single author) addresses RQ1 and the methodological challenges that come with using audio recordings that lack annotations in empirical research. In the paper, I develop an annotation pipeline that allows me to automatically annotate audio recordings of political speeches without using any prior human-annotated data. To validate the pipeline, automated annotations are compared with human annotations.

Papers B, C, and D address RQ2 and conduct empirical tests of whether political roles shape the sound of politicians' voices. The papers synthesize work from psychology on how deeply rooted vocal signals shape perceptions and classic political science concepts to generate empirical predictions about how a political role affects a speaker's vocal pitch. In Paper B (coauthored with Frederik Hjorth), we investigate whether vocal signals in legislative speeches reflect the prevailing patterns of partisan conflict, as indicated by a heightened pitch. In Paper C (single author), I ask whether the descriptive-to-substantive representation link is sustained and communicated in the vocal signals of politicians, also indicated by a heightened pitch. Finally, in Paper D, I analyze whether politicians who assume governing roles speak their power by signaling composure, here indicated by a lower pitch relative to when they are not assuming governing roles. The papers are summarized in Table 1.1.

	Paper		
	B	**C**	**D**
Title	Partisan Conflict in Nonverbal Communication	Committed but Constrained: Explaining Why the Descriptive-to-Substantive Representation Link Weakens Over Time	When They Go High, We Go Low: Rhetorical Rewards of Governing
Political Role	Partisan	Representative	Governing
Δ Pitch	Higher	Higher	Lower
Perceptible Sound	Agitation	Engagement	Composure

Table 1.1: Articles addressing RQ2: Title, role, expected change in pitch, and resulting sound.

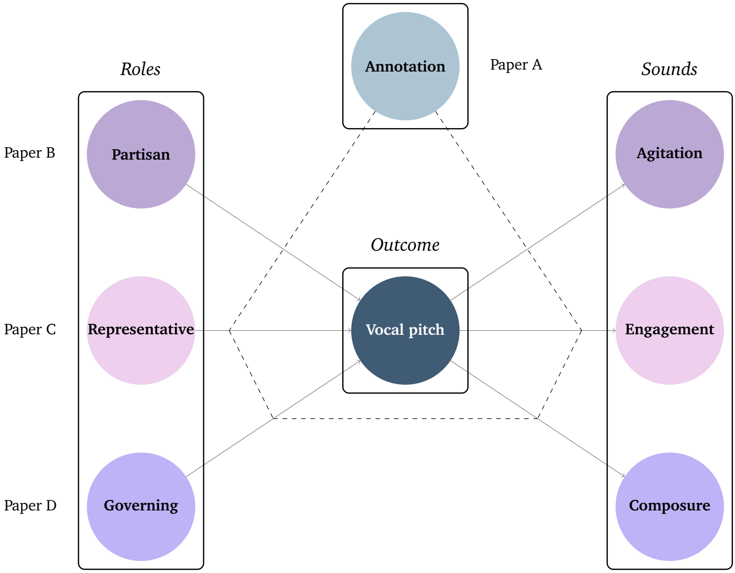


Figure 1.1: Dissertation overview.

Chapter 2
Pitching Sounds

THIS CHAPTER PRESENTS THE LITERATURE used to develop the theoretical framework in Chapter 3. The chapter proceeds as follows. First, literature from psychology is reviewed for how humans make inferences about emotions and traits from the sound of a speaker's voice. As the dissertation uses vocal pitch as the outcome, the review considers the sounds expressed in the pitch accordingly. This part establishes that politicians can be expected to vary the sounds of their voices, as conveyed in their vocal pitch, for two reasons: 1) when their speech has an emotional component and (2) when they are motivated or expected to signal certain traits. Second, based on the review in the first part, an emotional arousal scale of vocal pitch is outlined ranging from fully composed (lower than average pitch) to fully activated (higher than average pitch). This part establishes that variation in pitch can arise from multiple sources and that each source, for example, a political role, gives a different interpretation of the meaning of the variation.

2.1 Vocal Signaling and Perceptions

The human voice is incredibly flexible and can convey a range of signals in different contexts. At the same time, humans possess well-designed psychological systems that enable them to draw inferences about others based solely on the sound of their voices. This makes vocal speech a rich signaling device and source of perceptions. In the following, I focus on how vocal pitch signals traits and emotions and how humans draw vocal inferences based on their perceptions of those signals.

2.1.1 Emotional Arousal

The role of the human voice in signaling emotions has been known since the ancient Greeks (see, e.g., Scherer, 1993) and the pioneering work of Darwin (1872). This occurs predominantly through the fundamental frequency (F0) – the acoustic analog of the vocal pitch – by which a speaker gives a speech (Banse & Scherer, 1996; Titze, 1994). Each individual speaks with a baseline vocal pitch, but it can vary between and even within speeches. This variation conveys information about the emotional state of the speaker at the time of speaking.

Whether emotionally driven variation in pitch is strategic or not is a highly disputed question. One account prescribes that pitch changes occur due to subconscious biological origins (e.g., Dietrich, Hayes, & O'Brien, 2019); another posits that the effect is conscious and intentional (Scherer et al., 2003, p. 232). In the former, variation in vocal pitch is uncontrollable, and in the latter, it is controllable. The core of this theoretical dispute revolves around whether humans can emulate physiological effects. As noted in Knox and Lucas (2021), it is highly likely that trained speakers, such as politicians, can imitate virtually any behavior, including behavior with biological origins. Anecdotal evidence supports this notion as multiple politicians have attended voice training throughout their careers to intentionally shape how they are perceived by voters.[1] That trained speakers can emulate physiological effects is already evident by the fact that most research on the vocal expressions of emotions relies on actor portrayals (Scherer et al., 2003, p. 232-233). This rules out that changes in pitch are completely unintentional (Knox & Lucas, 2021, p. 651). A reconciling, arguably more realistic, perspective views pitch variation as simultaneously intentional and unintentional, which allows both physiological and emulative effects to be present. This does not imply that trained speakers always modulate their voice intentionally, only that they have the ability to do so.

The link between vocal pitch and inference and emotions has been investigated using both dimensional and discrete perspectives (Mauss & Robinson, 2009). Where the dimensional model posits that any emotions experienced by an individual are a combination of valence (also called sentiment) and arousal (also called activation), the discrete perspective prescribes that each emotion can be uniquely classified and categorized based on the individual's experience, physiology, and behavior (Mauss & Robinson, 2009, p. 211). Although both models have been investigated in relation to vocal pitch, "the most consistent association reported in the literature is between arousal and vocal pitch, such that higher levels of arousal have been linked to higher pitched vocal samples" (Mauss & Robinson, 2009, p. 222) with vocal pitch consistently "influenced by affect-related

[1] To mention a few, Margaret Thatcher, former prime minister (PM) in the U.K., attended voice training to lower the frequency of her voice to appear more authoritative and "presidential" (Moore, 2013; Pisanski, Cartei, et al., 2016), Barack Obama, former president of the United States, volitionally changed his use of African American English across contexts and audiences (Alim & Smitherman, 2012), and Ed Miliband, former PM in the U.K., used different phonological variants depending on the composition of the audience (Kirkham & Moore, 2016).

arousal" (Owren & Bachorowski, 2007, p. 240). The link to affect-related valence (Johnstone & Scherer, 2000; Leinonen et al., 1997) is unclear, most likely because similarly aroused emotions have different valence (Mauss & Robinson, 2009). The same goes for discrete emotions, which can only be partially distinguished by combining ten acoustic features, including vocal pitch (Banse & Scherer, 1996).[2] In summary, this literature shows that the emotional information contained in pitch variation is strongest when considering the level of emotional arousal experienced or signaled by a speaker (Bachorowski & Owren, 1995).

The dimensional model underlying this finding is the so-called circumflex – or core affect – model of emotion (Russell, 1980; Russell & Barrett, 1999), which consists of a horizontal and a vertical dimension. The horizontal corresponds to the valence (i.e., sentiment) of a speech and captures whether the speech reflects a positive or a negative tone. The vertical corresponds to arousal and captures whether the speech is activated or subdued (Cochrane et al., 2022, p. 100). The variation in vocal pitch conveys information about this vertical dimension. This means that a higher pitch is a signal of higher emotional activation and a lower pitch is a signal of lower activation, but the change is agonistic to the valence of this shift. In other words, a higher pitch may be the result of both anger and joy. Furthermore, the scale is symmetric. This implies that a higher pitch is indicative of higher activation and higher composure at the same time and similarly for a lowering of the pitch.

[2] While it is intriguing to measure discrete emotions with vocal pitch, it also comes with a premium. As noted by Mauss and Robinson (2009): "However, these links were complex and multivariate in nature, involving post hoc comparisons that were novel to the literature and in some cases perhaps not theoretically motivated. Thus, replications are crucial to having greater confidence in the findings reported in this study" (p. 222).

The link between emotional arousal and vocal pitch has been exploited by political scientists to study political representation. Based on small variations in vocal pitch and a large-scale text-audio corpus of floor speeches in the U.S. House of Representatives from 2009-2014, Dietrich, Hayes, and O'Brien (2019) argue and show that legislators' vocal pitch varies systematically across speeches with existing theories of the descriptive-to-substantive representation link and issue ownership. Female legislators speak with a consistently higher pitch than male legislators, and this generalizes to legislators' broader issue commitments as predicted by theories of issue ownership (e.g., Petrocik, 1996). The former is replicated in the German Bundestag, highlighting the usefulness of using vocal pitch in measuring emotional arousal in politicians' speeches (Rittmann, 2024).[3]

2.1.2 Leadership Traits

Humans draw inferences not only about emotions but also about traits from the sounds of a speaker's voice. Those traits can be classified as descriptive and subjective. The former (e.g., sex, size, and strength) are observable, perceptible, and potentially factually true; subjective traits are fundamentally unobservable and only exist perceptually (e.g., charisma, trustworthiness, or attractiveness). The importance of perceptions emphasizes that descriptive and subjective traits might be highly related despite being theoretically distinct. For instance, a speaker perceived as strong – a descriptive trait – might also be perceived as dominant – a subjective trait – for that exact reason. This suggests that descriptive traits might causally mediate the link between vocal

[3] Vocal pitch as an indicator of a politician's emotional arousal can also be used to study the attitudes of Supreme Court Justices (Dietrich, Enos, & Sen, 2019) and party competition (Arnold & Küpfer, 2024). Furthermore, Knox and Lucas (2021) study how the sounds of human voices can be used to study skepticism using a comprehensive set of features trained on speaker-specific supervised Hidden Markov Models.

and subjective characteristics in human psychological systems (e.g., Puts et al., 2006).

In the following, I focus on two subjective traits – *dominance* and *competence* – where the vocal pitch has clear political ramifications in terms of power, status, and leadership. Inferences of dominance and competence are entirely perceptual and are based on psychological impressions shaped by deeply rooted evolutionary foundations that are found to predict the electability of political candidates.[4] In a seminal study of eight presidential elections in the United States, lower-pitched presidential candidates receive more votes than higher-pitched candidates (Gregory Jr & Gallagher, 2002). This finding has later been generalized to presidential (I. P. Banai et al., 2017) and parliamentary elections around the world (B. Banai et al., 2018), national legislative elections (Klofstad, 2016), and experimental settings (Klofstad, 2016; Klofstad et al., 2012; Tigue et al., 2012).[5]

Dominance

The first trait found to mediate the effect of vocal pitch on the selection of political leaders is dominance. This is highly related to the physical characteristics of a speaker, a descriptive trait where speakers who have a lower pitch are evaluated as both larger and stronger and are perceived as more physically and socially

[4] I treat dominance and competence as distinct traits, but others view the former as preceding the latter in the causal chain. For example, Laustsen and Petersen (2015) argue that if "dominance-related traits exclusively influenced a leader's competence with respect to securing collective action, dominant leaders should be universally preferred" (p. 287). This suggests that dominance causally precedes competence.

[5] Subjective traits that also carry political implications but are not covered in this report include work on vocal attractiveness (Borkowska & Pawlowski, 2011; Collins, 2000; Feinberg et al., 2005; Oguchi & Kikuchi, 1997; Surawski & Ossoff, 2006; Zuckerman & Driver, 1989; Zuckerman & Miyake, 1993), charisma (Niebuhr et al., 2017; Novák-Tót et al., 2017; Signorello, 2019, 2021), and trustworthiness (Fish et al., 2017; O'Connor & Barclay, 2017; Oleszkiewicz et al., 2017; Schild et al., 2020; Schirmer et al., 2020).

dominant (Ohala, 1984; Puts et al., 2006, 2007). Dominance is particularly relevant when making impressions and persuading others based on the sound of the voice because “it is an innately used and recognized signal” (Burgoon et al., 1996, p. 316). The notion of dominance is closely related to the perceptions of leadership (Tigue et al., 2012) and to the more general concept of the social power of a speaker (Aung & Puts, 2020), that is, control over the results valued by others (Fiske & Berdahl, 2007).

The link between vocal pitch and preference for leaders who are more dominant and better leaders is often theorized to be adaptive. Evolutionary psychology points to two mechanisms. The first is male dominance competition, also called intrasexual selection, where more dominance ensures greater access to valued resources and mates (Puts et al., 2006). Dominance might be physical as in non-human animals but also social through leadership and persuasion (Henrich & Gil-White, 2001). The second explanation is leadership selection and a preference for dominance because it solves a range of collective action problems when living in groups (B. Banai et al., 2018).

Because dominance is adaptive, lowering the pitch is an efficient way for public speakers, such as politicians, to showcase their status and power (Carney et al., 2005; Gregory Jr & Gallagher, 2002; Kalkhoff et al., 2017) because it is associated with interpersonal deference and power relations (Gregory, 1994) and signals followership (Cheng et al., 2016). This perception is not gendered but generalizes to stereotypical masculine and feminine political positions (Anderson & Klofstad, 2012). Although the preference for dominance applies to different positions, it might not be universally preferred due to its externalities (Laustsen & Petersen, 2015) or the ideology of the follower (Laustsen et al., 2015), but its value increases in times of war and conflict, as it signals dispositional abilities to prevail (Tigue et al., 2012).

Competence
The second trait found to mediate the effect is competence. Like dominance, the notion of competence is closely tied to perceptions of leadership and the ability to "properly perform his/her job, identifying and employing the appropriate policies that enable her to get the job done" (Galeotti & Zizzo, 2018, p. 27). Dominance is more related to a speaker's "physical prowess", and competence is more associated with "integrity" (Tigue et al., 2012, Table 2). Competence may not be an important trait in social contexts, but it is universally valued in politics to the extent that it signals competent leadership. A rich set of studies shows that voters draw inferences about a speaker's competence from the vocal pitch, and lower-pitched politicians are consistently evaluated as more competent than higher-pitched politicians (Klofstad, 2016; Klofstad et al., 2012, 2015; Tigue et al., 2012).

The theoretical underpinnings explaining the link between vocal pitch and perceptions of competence are less understood. As noted by Klofstad et al. (2015), the "result may be surprising, as it is unlikely that people with lower-pitched voices are inherently more competent leaders" (p. 11). This notion is empirically supported as vocal pitch is not statistically associated with actual leadership abilities (Klofstad & Anderson, 2018). However, the link probably persists because it "reflects correlates between voice quality and leadership capability that were relevant at some earlier time in human evolutionary or cultural history" (Klofstad et al., 2015, p. 11) and is "associated with competent leadership" (B. Banai et al., 2018, p. 2).

2.2 Arousal Scale and Drivers

So far, the literature review has established two main points about the information encoded in the vocal pitch of a politician:

1) the pitch changes as the politician becomes more emotionally activated (higher pitch) or more composed (lower pitch), and 2) the pitch changes when the politician wishes to leave off impressions of certain traits. This shows that variation in pitch can be explained by both variation in emotional arousal and variation in the motivation to signal character traits. As the source of this variation is different, the pitch in a given speech is likely a weighted combination of the context of the speech (e.g., campaign vs. legislative debate) and the role(s) held by a politician (e.g., backbencher vs. minister) at the time of speaking. For example, Touati (1993) shows that Jacques Chirac, former president of France, changed his vocal pitch substantially from a pre-electoral speech to a post-electoral press conference from an average pitch of 163 and 117, respectively.

These observations suggest that the signaling of emotions and traits can be viewed as a symmetrical and unified scale of arousal that varies from low activation and high composure to high activation and low composure. This scale is illustrated visually in Figure 2.1. Higher arousal manifests itself in politicians speaking with a higher than average vocal pitch, and vice versa for lower arousal. In other words, because dominance is associated with a lower pitch, this can be thought of as speaking with less arousal, as indicated by lower activation and greater composure. Likewise, because engagement is associated with a higher pitch, this can be thought of as speaking with greater arousal, as indicated by greater activation and lower composure. This illustrates two important points.

First, the scale operates within politicians, but not between them, as different levels of vocal pitch, due to physiological and speaker-specific differences, might confound changes in trait- or emotion-driven arousal. This is accounted for empirically by standardizing speakers' vocal pitch such that values are interpreted in relation to a speaker's own baseline (see Chapter

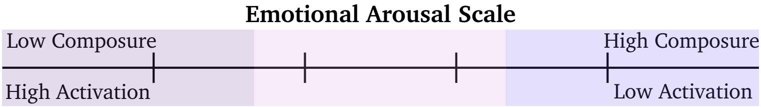


Figure 2.1: Scale of emotional arousal from low activation/high composure to high activation/low composure.

6). Second, the scale is invariant to the source of the variation. Speakers can be more or less aroused for multiple reasons, and while variation in vocal pitch encodes a speaker's arousal, the meaning and interpretation of the sound depend on the underlying driver. This is illustrated with the measurement model shown in Figure 2.2. The model makes clear that arousal has multiple drivers, with three of those investigated in the dissertation. As a result, the interpretation of the pitch variation is based on the paper-specific predictor and data partitioning.

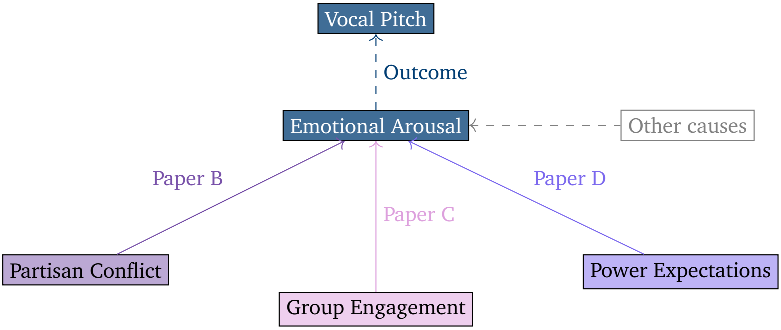


Figure 2.2: Directed acyclical graph illustrating links between pitch, arousal, and drivers of arousal.

Chapter 3
Pitching Politics

THE THEORETICAL CLAIM OF THE DISSERTATION is that politicians adapt the sound of their voices – as conveyed in vocal pitch – when they assume different political roles. Since vocal pitch is a dynamic signaling tool, politicians are expected to vary their pitch systematically in response to their roles. The dissertation considers three key roles — *partisan*, *representative*, and *governing* – each of which plays vital democratic, representative, and institutional functions in liberal democracies. The partisan role (Paper B) is central to political life in all, particularly parliamentary, governmental systems (e.g., Thomassen, 1994), and is a defining feature of conflict and competition in democracies (Druckman et al., 2013). The representative role is synonymous with the question of representation (Paper C) and centers on the task of making the positions and priorities of the represented visible and heard (Pitkin, 1967). The governing role (Paper D) is tied to leadership positions and concerns the responsibilities of governance and those who hold the power to make executive decisions that affect citizens, for better or worse.

3.1 Political Roles as a Mechanism

The concept of "roles" is a useful analytical tool to understand *when* politicians vary the sounds of their voices. Although the three substantive papers (B, C, and D) each considers how a single role affects the sound of a politician's voice, the concept of roles provides an overarching framework to connect, structure, and organize the articles theoretically. Where political roles have been treated as both dependent and independent variables in legislative studies (Blomgren & Rozenberg, 2015, p. 3), I deploy the concept of roles as more of an intervening and mediating variable that helps to explain how a legislator's party membership, personal background, and institutional positions shape the dynamics of the vocal pitch (Andeweg, 2014b, p. 282). This is reminiscent of Bailer et al. (2022), who utilize roles "as a mechanism that influences strategic choices during parliamentarian career cycles" (p. 538). Treating political roles as a mechanism illustrates the subtle but important point that the concept is related to but distinct from that of positions. Positions are functional in the sense of having well-defined tasks, duties, and responsibilities, and roles are subjective in the sense of being shaped more by norms, goals, and motivations (Andeweg, 2014b, p. 270). This distinction proves useful in connecting the articles and in developing theoretical arguments for why a political role is expected to be systematically reflected in vocal pitch.

3.2 Partisan Role

The first role concerns partisanship and is the subject of Paper B. Parties are virtually ubiquitous in democratic legislatures (Martin et al., 2014, p. 17). In contemporary democracies, the large majority of legislators are elected under party labels making political parties the main “agents of representation” (Dalton, 2017, p. 609). Parties are also an essential component of organizing legislative activities (Saalfeld & Strøm, 2014, p. 372) and rely on a division of labor where its members are assigned the position of spokesperson with the task of representing the party on a specific policy portfolio (Andeweg & Thomassen, 2011). However, the most important function of parties is arguably providing competition and healthy conflict to the voters (Schattschneider, 1960), which strengthen democratic accountability and the consistency of opinion formation among voters (Levendusky, 2010). The competitive nature of partisanship is even institutionalized in core functions of legislatures. Parties are embedded in the procedures of legislative debates with legislators granted access to the floor based on their party membership in an “institutionalized arguing game of ‘us’ versus ‘them’ (between politicians’ parties)” (Mollin, 2018, p. 209).

The importance of partisanship in organizing legislative debates suggests that legislative speeches are a rich source of information about how legislators communicate and express partisan conflict. Legislators take the floor due to their position as party members, but the nature of partisanship is only truly competitive to the extent that they play the role of a partisan. The “partisan role” prescribes that partisanship matters not only by virtue of party membership or the organizing factor but by an “object of emotional involvement or attachment” for the individual legislator (Wahlke et al., 1962, p. 352). In their seminal study of legislative roles in U.S. state legislatures, *The Legislative System: Explorations in Legislative Behavior*, Wahlke, Eulau, Buchanan, and Ferguson define the partisan role as “the roles they feel called upon to take by virtue of their membership [in a party]” (p. 343, Chapter 13). Affect towards one’s party is an important driver in this definition and is even deemed a necessary condition for partisanship to meaningfully matter in shaping legislative behavior (p. 352, Chapter 13).

How do legislators express this emotional attachment? Existing studies measuring partisan conflict in speeches rely almost exclusively on the verbal dimension, analyzing differences in word choices as indicators of party polarization in speech transcripts. For example, Gentzkow et al. (2019) argue that if "parties speak more differently today than in the past, these divisions could be contributing to deeper polarization" (p. 1308). Using a similar approach, Peterson and Spirling (2018) argue that party polarization can be measured by the extent to which different parties are "more or less distinguishable over time, in terms of what they choose to say" (p. 121). A similar approach is suggested by Rheault and Cochrane (2020) where party embeddings, computed from word embeddings fitted with party-specific indicators, can be used to measure ideological positions and therefore the polarization between parties and individual legislators. Finally, Proksch, Lowe, et al. (2019) argue that sentiment, i.e., the relative use of positive and negative words, identifies political conflict.

All of these approaches generate valuable insights into partisan competition, but they do not necessarily capture the emotional underpinnings. As noted by Cochrane et al. (2022), the emotion of a speech is less about syntax and word choice but more about nonverbal components (Cochrane et al., 2022, p. 99). This includes, for example, intonation patterns as expressed in vocal pitch contours (Bänziger & Scherer, 2005). To the extent that pitch is a reliable indicator of a speaker's emotional arousal as outlined, vocal pitch is an important channel through which legislators signal conflict. Notably, while the pitching of partisan conflict might be partially correlated with textual measures (Gennaro & Ash, 2023, Appendix Table A.1), the signal is likely to provide information about conflict independently of other partisan differences in language usage.

To study how legislators signal partisan conflict in their vocal pitch, I focus on speeches in which legislators find themselves in interpersonal conflict (Deutsch, 1973). While they often take the floor on behalf of their group, i.e., their party, they do not debate with parties but with other persons. In other words, while the nature of the conflict is partisan and group-based in the aggregate, it is expressed at the interpersonal level, such as when legislators from opposing parties debate each other. I refer to such an exchange as *dyadic*. In dyadic speeches, the *sender* is the speaking legislator, and the *target* is the addressee of the speech whether it be an individual legislator or a party. The partisan structure of legislative debates means that dyadic exchanges occur frequently, but legislators can also emphasize the conflict themselves by verbally targeting other parties or legislators in their speeches net of the institutional structure.[1]

There are at least two sources of partisan conflict that are worth exploring in the context of dyadic speeches. The first is *partisan polarization*. I follow existing work and refer to party polarization as "the degree of ideological differentiation among political parties in a system" (Dalton, 2008, p. 900). To the extent that partisanship has an emotional attachment, and variation in vocal pitch signals partisan conflict, it follows that *legislators should speak with a higher vocal pitch than average in dyadic speeches where the target is more ideologically different*. In a partisan role, this suggests that variation in pitch can be described as having the sounds of agitation and disagreement where the pitch

[1] This leaves open the possibility that what appears to be partisan conflict is confounded by social polarization, which relates to legislators' social ties and friendships (Dietrich, 2021), or the respect between the speakers (Caldeira et al., 1993). On average, however, this seems unlikely given the partisan nature of friendship where party identification is among the strongest predictors of friendship within legislatures (Caldeira & Patterson, 1987).

is used to emphasize tension and highlight conflict between parties.

In European multiparty systems, party competition generally occurs at the level of so-called blocs rather than individual parties (Bale, 2003). This is the case for the Danish parliament (Green-Pedersen & Thomsen, 2005),[2] and therefore the expectation regarding partisan polarization is tested at the bloc level with dyadic exchanges characterized as in-partisan and out-partisan based on bloc affiliation.

The second source of partisan conflict worth exploring is *policy conflict*. Although political conflict predominantly operates at the partisan level in party-centered systems, substantial policy conflict also occurs within as well as between blocs. I refer to a policy conflict as a situation in which the position on a bill differs between two parties. Compared to polarization, which refers to ideological differences, policy conflict is about differences in positions on a single legislative bill. Policy conflict can be viewed as a nested version of partisan polarization where ideological differences coincide with policy differences in the case where blocs are completely homogeneous. In reality, however, within-bloc variation in policy conflict is to be expected. Parties within the same bloc can disagree as long as the policy conflict does not fracture the coalition or empower opposition parties trying to turnover government (Green-Pedersen & Thomsen, 2005, p. 159). From this it follows that *legislators should speak with a higher vocal pitch than average in dyadic speeches when the speaker and target disagree on a bill.*

[2] In the recent election in 2022, however, the historical major left-wing party, the Social Democrats, joined forces with the historical major right-wing party, Denmark's Liberal Party, to create a majority coalition government disrupting the otherwise stable bloc system (Kosiara-Pedersen & Kurrild-Klitgaard, 2018). Speeches given after this election are not contained in the multimodal text-audio corpus used in the dissertation but provide an intriguing opportunity to disentangle the mechanism even further.

3.3 Representative Role

The second role concerns the legislator as a representative and is the subject of Paper C. Representational roles are arguably the most studied in the context of role theory, encompassing both the dimensions of "style' and "foci" (Blomgren & Rozenberg, 2015, p. 12-14). Style refers a legislator's approach to fulfilling their representational duties, whether as a delegate, adhering closely to constituents' preferences; a trustee, exercising independent judgment; or a politico, balancing the two roles depending on the issue at hand (Wahlke et al., 1962, Chapter 12). Foci pertain to the scope of interests a legislator prioritizes, whether national, territorial, or non-territorial (Andeweg, 2014b, p. 274).

A prominent non-territorial focus concerns the representation of social groups. Although any legislator can focus on representing social groups, descriptive representatives – legislators who share one or more characteristics with social groups – are generally assumed to prioritize their social groups more than non-descriptive representatives (Mansbridge, 1999, p. 642). This observation is often formulated using the concepts of descriptive representation – "standing for" – and substantive representation – "acting for" – coined by Pitkin (1967) in her classic conceptual work, *The Concept of Representation*, with a large body of work documenting an empirical relationship for ethnicity, gender, and class (for ethnicity; Grose 2005; Preuhs 2006; Rocca and Sanchez 2008; Saalfeld 2014; Saalfeld and Bischof 2013; Wilson 2010, for gender; Bratton and Ray 2002; Carroll et al. 1994; Celis et al. 2008; Childs and Krook 2009; Pearson and Dancey 2011; Wängnerud 2009, for class; Carnes 2012; Carnes and Lupu 2015; O'Grady 2019).

The empirical link between descriptive and substantive representation is often theorized on the basis of two competing models of legislative behavior. The first is the so-called presence model (Preuhs, 2006), which posits that descriptive representatives are more likely to act on behalf of their group members due to their personal background (Phillips, 1998). Descriptive representatives share background and experiences with members of their social groups, leading to greater commitment, engagement, and willingness to prioritize the issues of the groups. The second is the so-called accountability model (Broockman, 2013) according to which descriptive representatives are more likely to prioritize their groups' policy issues, not because they are committed, engaged, or willing per se, but because they face electoral incentives that compel them to act on behalf of their groups.

Despite the divergent theoretical underpinnings, both models can be understood through the lens of representative roles (Celis & Wauters, 2013) and the theories proposed by Searing (1994) and Strøm (1997), respectively. Although neither discusses the descriptive-to-substantive representation link, each theory offers a way to understand the representative role as a mechanism linking the social backgrounds of legislators to their behavior. The presence model can be interpreted in relation to the motivational approach advanced by Searing (1994) in his study of Westminster's political roles, while the accountability model fits the strategic approach proposed by Strøm (1997). Both theories share the notion that institutions constrain and shape role adoption, but they fundamentally disagree on whether a legislator's goals are predominantly personal or strategic.

The motivational approach posits that role adoption is mainly a function of a legislator's personal preferences and motivations. In this account, roles are either preference- or position-based, with the former (e.g., backbenchers) being less institutionally constrained than the latter (e.g., ministers) (Blomgren & Rozenberg, 2015, p. 21-22). The preferences of a legislator are both rational and emotional, but the latter are "the principal energizing forces in all parliamentary roles" (Searing, 1994, p. 19). Rational goals refer to typical career goals, such as reelection, emphasized by rational choice scholars (Mayhew, 1974), while the latter are psychological and deeply personal to the legislator (Andeweg, 2014b, p. 271). Where the incentives provided by career goals might shift over a legislator's career, those provided by the emotional ones are generally fixed.

The strategic approach argues that role adoption is mainly a response to a legislator's career goals: reselection, reelection, party office, and legislative office (Strøm, 1997). In this rational choice account, political roles are defined as "strategies for the employment of scarce resources toward specific goals" (p. 155). This contrasts sharply with the motivational approach. Where Searing (1994) sees emotional incentives as the factor distinguishing positions from roles, Strøm (1997) views variation in the pursuit of career goals between legislators as what separates roles from positions. In this view, roles are distinct from positions because legislators pursue different goals; some pursue reselection, others reelection, and still others pursue office (Andeweg, 2014b, p. 270).

These generate vastly different interpretations of how the role of the representative mediates the descriptive-to-substantive representation link. The motivational approach posits that the social background of legislators motivates them to take on a role of "group representative" due to emotional underpinnings (Celis & Wauters, 2013). The strategic approach postulates that legislators take on the role as a strategy to achieve their career goals. In case of the former, this generates the expectation that legislators who are descriptive representatives should assume the role of group representative throughout their careers to the extent that the role rests on stable emotional incentives. In case of the latter, legislators who are descriptive representatives should assume the role of group representative only to the extent that it is a viable strategy in achieving their career goals. Put differently, the motivational approach prescribes a stable and time-invariant link between descriptive and substantive representation, whereas the strategic approach prescribes a dynamic and time-variant link.

What should we observe empirically if one theory or the other is correct? Using Strøm (1997) and his rational choice approach, Bailer et al. (2022) argue that descriptive legislators adopt their representative role as a response to their strategic goals and choices faced throughout their legislative careers (p. 538). Because of this, descriptive representatives should deprioritize their groups over time as they "acquire expertise and credibility within parliament" (p. 539) enabling them to shift focus from their early career efforts in portraying them as group representatives to new policy areas unrelated to their social identity. Using interpellations to measure attention to group-specific issues, Bailer et al. (2022) find empirical evidence consistent with this theoretical expectation. On average, descriptive representatives devote more attention to group issues than non-descriptive representatives, but this declines substantially as they gain seniority, diminishing the substantive value of legislators' social backgrounds over time.

On the surface, this seems to be consistent with the accountability model, but the pattern does not rule out the motivational approach, as the diminishing value could be confounded by institutional constraints. As legislators gain experience and credibility, their career goals may shift from reelection to office (Davidson, 1969), but they are also assigned institutional positions for the exact same reasons (Wüst, 2014). The latter functionally limits their selective prioritization of issues independently of their career goals. For example, a ministerial position requires legislators to devote attention to the portfolio associated with the specific ministry. However, this does not necessarily mean that the emotional goals change, but might also be a result of changes in the institutional constraints surrounding a legislator.

This suggests that representation might occur through both attention and engagement. Issue attention, the frequency with which a policy issue is raised, relates to the prioritization of a group, and issue engagement, the emotional arousal with which a policy issue is raised, relates to the commitment to a group, may but do not necessarily covary. Both the accountability and the motivational model predict a decline in issue attention devoted to group-specific policy areas throughout a legislator's career due to institutional constraints. That is, *the frequency with which descriptive representatives raise the issues of their groups decreases with a legislator's seniority*.

However, the models disagree on the trajectory of issue engagement. The accountability model suggests that descriptive representatives' engagement with the issues of their groups declines with time, so *the vocal pitch of descriptive representatives when they talk about the policy issues of their groups declines with their seniority*. The motivational model suggests that engagement remains the same net of career constraints, such that *the vocal pitch remains unchanged when legislators who are descriptive representatives discuss their groups' policy issues over time*. If so, the vocal pitch should not be related to the seniority of the legislator. Variation in vocal pitch signals partisan conflict when legislators participate in dyadic exchanges with outpartisans and group representation when legislators raise and engage with their group's policy issue on the floor.

Legislative speeches provide a promising way to disentangle these competing explanations. When legislators speak on the floor, they convey information about what they say and how. The first is related to the verbal content of a speech, such as the topic, and can be used to measure issue attention, and the latter is related to the delivery of the speech, including intonation patterns expressed in the variations and contours of the vocal pitch (Bänziger & Scherer, 2005) and can be used to measure issue engagement. I study this using women and lower social class as indicated by lower-educational groups. I use the empirical approach proposed by Dietrich, Hayes, and O'Brien (2019) and couple the policy issue of each speech with a measure of the vocal pitch, allowing "examination of legislators' emotional intensity around different issue areas" (Dietrich, Hayes, & O'Brien, 2019, p. 941).

3.4 Governing Role

The third role concerns legislators in governing roles and is the subject of Paper D. I use the example of a minister, that is, the head of a ministry that holds the executive power and authority to implement policy. The Folketing has a strong tradition of drawing ministers from the members of the legislature, resembling the Westminster tradition (Strøm, 1997, p. 168) where ministers must be MPs by constitutional provision (Andeweg, 2014a, p. 535).

Among the three roles considered in the dissertation, the governing role is arguably closest related to the position itself. In his distinction between "preference roles" and "position roles", Searing (1994) classifies frontbench positions such as ministers and party leaders as positional and backbencher positions such as the constituency member or the policy advocate as preferential (Andeweg, 2014b, p. 272). The difference between the two types of roles is determined by the degree of functional constraints when holding an institutional position. As Searing writes: "Position roles are associated with positions that require the performance of many specific duties and responsibilities. Preference roles, in contrast, are associated with positions that require the performance of few specific duties and responsibilities" (Searing, 1994, p. 12). When the job description is fixed, as for the position roles, there is less room for interpretation of how to play the position (Andeweg, 2014a, p. 534). As a consequence, personal preferences and goals, emotional and rational, are less influential in shaping position roles than in shaping preference roles (Andeweg, 2014b, p. 270).

Holding a governing position such as minister classifies as a position role due to the high number of fixed tasks and duties the politician is required to complete. This functionally constrains their legislative speeches (Hjorth, 2024). They are formally constrained in their speeches by the nature of their position, such as introducing legislation. They are also substantively constrained in terms of topics they can raise in their speeches. For example, ministers are asked to deal with more technical and complex policy issues and to handle potential crises, such as a pandemic (Louwerse et al., 2021). The same set of formal and subjective constraints do not apply to legislators in non-governing positions.

Despite the high number of functional constraints faced by governing legislators by nature of the position, what makes being a minister a role rather than just a position are perceived norms. Wahlke (1962) defines a role as a "coherent set of norms of behavior which are thought by those involved in the interactions being viewed to apply to all persons who occupy the position of legislator" (p. 8). Norms can be thought of as "expectations of behavior", suggesting that norms have substantial predictive value to the extent that politicians comply with the expectations associated with their position (Andeweg, 2014b, p. 270). The emphasis on the importance of norms in playing a role makes it highly related to the "logic of appropriateness" (March & Olsen, 1989, p. 154) where ministers adapt their behavior as a function of what is deemed appropriate when holding a position of political power.

This understanding of roles suggests that the governing positions are not only functionally constrained but also constrained by the role itself. These role constraints operate as a disciplining factor through the "logic of appropriateness" where ministers strive to comply and align with their role expectations independently of the functional constraints they face. Although parts of the motivation to comply with the expectations of holding a governing position might be explained by "the logic of consequentiality" (March & Olsen, 1989, p. 154-155) where ministers are assumed to behave in ways that maximize their utility (Andeweg, 2014b, p. 270), the main driver is theorized to be intrinsically rooted motivation to comply with the expectations of others (Fishbein & Ajzen, 1975).[3] This suggests that the behavior of ministers can be predicted by identifying the expectations associated with holding a governing position and how politicians signal them.

Legislative speeches provide an ideal setting for studying this because they contain information on the nonverbal expressions a minister "gives off" when giving a speech (Goffman, 1959, p. 2). As described in Chapter 2, the perceptions of a candidate's voice derived from the vocal pitch significantly influence the selection of political leaders, presumably because they shape perceptions of the characteristics of a candidate (e.g., Klofstad, 2016). The literature consistently finds that candidates with a lower pitch are more associated with traits of competence and dominance than candidates with a higher pitch (e.g., Klofstad et al., 2015). This suggests that these traits are both highly valued and expected in political leaders. Although the link between vocal pitch and perceptions of leadership traits has only been established between candidates, the speaker-specific arousal scale suggests that this can be thought of as speaking with greater composure.

[3] This mechanism is similar to social pressure in explaining voter turnout (Gerber et al., 2008).

In other words, candidates who speak that way, as indicated by a lower pitch, are perceived as more dominant and competent. To the extent that competence and dominance are traits expected in political leaders, this suggests that *legislators lower their vocal pitch when assuming governing roles*. Where a heightening of the pitch has the sounds of agitation and engagement in the context of partisan and representative roles, a lowering has the sounds of composure in the context of assuming governing roles.

To study how legislators change their vocal pitch when holding a governing position, I compare the vocal pitch before and after entering government using an within-legislator design. This serves as a direct test of the hypothesis, but is not able to disentangle the theoretical mechanism. A lowering of the pitch is consistent with both an effect of role and functional constraints. To parse this further, the data are partitioned to disentangle the two explanations, for example, by exploiting the variation in functional constraints in the types of debates that feature in the yearly parliamentary calendar and by taking policy topics into account.[4]

3.5 Theoretical Expectations

The theoretical framework generates a total of six testable hypotheses, two for B, three for C, and one for D. The individual articles are referred to for further tests and auxiliary results. Note that pitch is used as the outcome in all but one of the expectations set out in Table 3.1. The hypothesis on issue attention (the third listed in the table, the first listed for Article C) is included to clearly contrast and compare the difference in the trajectories between issue attention and issue engagement.

[4] Empirical tests of the mechanism are not reported in the summary report of the dissertation but are found in Paper D. The results are more consistent with the role mechanism and the functional mechanism.

Article	Label	Hypothesis
B	H_1^B	Pitch is higher in speeches targeting outbloc parties (polarization)
	H_2^B	Pitch is higher in speeches targeting parties with whom they disagree on bill (policy)
C	H_1^C	The frequency of addressing group issues declines over time
	H_2^C	Pitch declines in speeches addressing group issues over time (accountability)
	H_3^C	Pitch remains the same in speeches addressing group issues over time (motivation)
D	H_1^D	Pitch declines after legislators assume governing roles relative to before

Table 3.1: Theoretical expectations for each of the substantive papers.

Chapter 4
Audio Data

STUDYING THE SOUND OF POLITICS requires using recordings, not just speech transcripts. This moves the analysis from text to audio data. Although large-scale analysis of digitized speech text has become a methodological household tool among political scientists (Denny & Spirling, 2018; Grimmer & Stewart, 2013; Wilkerson & Casas, 2017), virtually no attention has been devoted to large-scale analysis of digitized speech audio (for exceptions, see e.g., Arnold & Küpfer, 2024; Dietrich, Enos, & Sen, 2019; Dietrich, Hayes, & O'Brien, 2019; Knox & Lucas, 2021; Neumann, 2019; Rittmann, 2024; Tarr et al., 2023). As noted by Knox and Lucas (2021), "a wide range of research has already studied audio recordings of political speech – but in virtually every domain, researchers have done so by extracting transcripts, then discarding the remainder of their data" (p. 651-652). Hence, to utilize the information conveyed in audio recordings, we first need to obtain basic knowledge about how to work with audio data. Accordingly, this chapter serves three main purposes. The first is to explain the basics of working with audio data and define "sound" from a physical and physiological perspective. The second is to explain how sound relates to vocal pitch – the key outcome of the dissertation. The third and final purpose is to explain the role and importance of audio annotations in preprocessing recordings.

4.1 Sound Theory

To understand the basics of working with audio data, we must first understand what constitutes "sound". In its simplest form, any audio recording contains a digitized version of an audio signal representing a sound. The physical definition of what we call "sound" refers to a vibration that travels as an acoustic wave when air molecules oscillate as a result of air pressure applied by a source (Camastra & Vinciarelli, 2015, p. 15).[1] Possible sources include traffic noise, animal calls, explosions, and human speech.

From a physical perspective, any sound wave, regardless of the source, can be fully characterized by amplitude (A) and frequency (f). The amplitude refers to the amount of pressure used to produce the wave (Rabiner & Schafer, 2011, p. 27), and the frequency, measured in hertz (Hz), denotes the number of times a wave repeats itself during a second (Camastra & Vinciarelli, 2015, p. 15-16). The frequency f is defined as $\frac{1}{T}$ where T is the time it takes for the wave to return to its equilibrium position. These basic components are illustrated in Figure 4.1a, which shows the waveform representation of a sine wave with a single frequency of 5 Hz and a fixed pressure. As shown in the figure, the amplitude/pressure A corresponds to the wave's

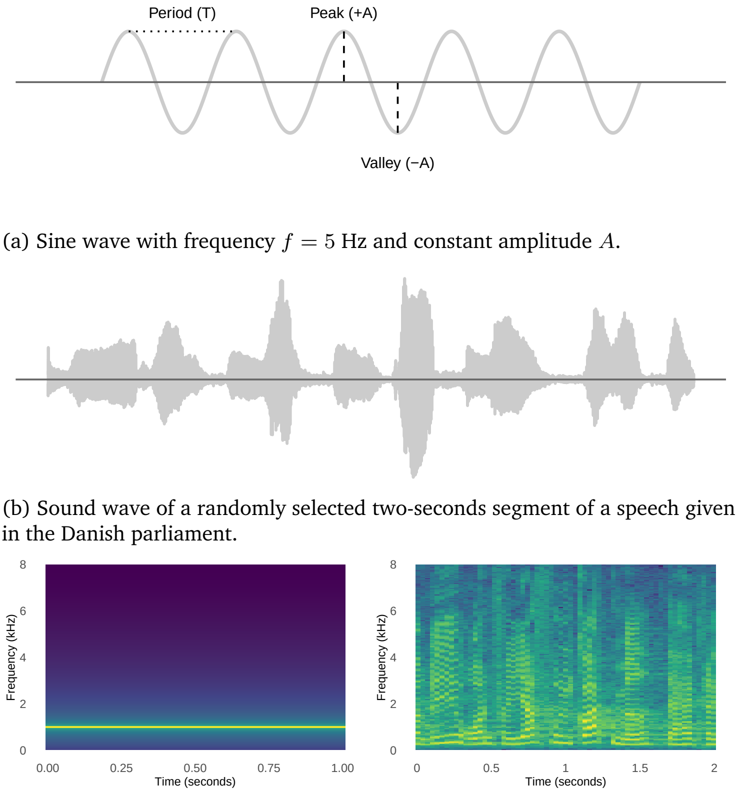
[1] Aristotle alluded to this understanding of "sound" as early as 350 years BCE in this book *On the Soul*. In Book II), part 8, he writes:

> What has the power of producing sound is what has the power of setting in movement a single mass of air which is continuous from the impinging body up to the organ of hearing. The organ of hearing is physically united with air, and because it is in air, the air inside is moved concurrently with the air outside. Hence animals do not hear with all parts of their bodies, nor do all parts admit of the entrance of air; for even the part which can be moved and can sound has not air everywhere in it (Aristotle, 350 B.C.E.)

.

highs (peaks) and lows (valleys). In this case, the period T corresponds to 0.2 seconds because $f = \frac{1}{T} = \frac{1}{0.2} = 5$ Hz.

The sine wave in Figure 4.1a only has a single frequency, but real-word sounds are always more complex and contain multiple frequencies simultaneously. This means that sounds are an energy distribution over different frequencies, which can be viewed as a "sum of single frequency sounds" (Camastra & Vinciarelli, 2015, p. 18). This is illustrated in Figure 4.1b, which shows a randomly selected two-second segment of a speech given by a legislator on the floor of the Danish parliament. This speech waveform is more complex than the sine wave, and the frequencies are hardly visible to the naked eye. Frequencies constituting a speech wave can be recovered using a Fourier transform (Sneddon, 1995), an application of Fourier's Theorem stating that complex periodic waves, such as those generated by human speech, can be decomposed into individual sine waves, each having a single frequency.



(a) Sine wave with frequency $f = 5$ Hz and constant amplitude A.

(b) Sound wave of a randomly selected two-seconds segment of a speech given in the Danish parliament.

(c) Spectrogram of sine wave with $f = 1,000$ Hz.

(d) Spectrogram of speech wave in Figure 4.1b.

Figure 4.1: Waveform and spectrogram representations of sound waves.

The lowest of those frequencies contained in a wave is called the fundamental frequency (F0) and is perceived as the pitch of a sound. The higher frequencies are called harmonics or formants and are integer multiples of the F0 by Fourier's theorem (Oppenheim, 1999). This means that the frequencies of complex sound waves, such as those produced by human speech, can be written as the sum of ω, 2ω, 3ω, and so on. Most energy is concentrated in the first 10-12 formants for human speech (Camastra & Vinciarelli, 2015, p. 19), indicating that a man or a woman vocalizing with an average F0 of around 100 and 300 Hz distributes most of the energy below 2,000-4,000 Hz.

This holds implications for how audio recordings are digitized and the rate by which the analog signal is sampled. For example, the speech wave plotted in Figure 4.1b is a digitized representation of an analog signal. The digital signal is a sampled and discretized version of the analog and continuous audio signal with the sampling rate denoting the number of samples per second. This can be thought of as a population and sample relationship where the analog signal is the population we want the sampled digital signal to represent. The population-sample analogy is useful in relation to the sampling rate. For a discretized signal to properly represent the continuous signal, the former must sample the latter at a rate that enables reconstruction of the signal's frequencies. This can also be expressed mathematically using the Nyquist-Shannon sampling theorem. The theorem says that an analog signal can be fully reconstructed by a digital signal by sampling at twice the rate of the highest frequency in the analog signal (Rabiner & Schafer, 2011, p. 98). This yields a lower bound for the sampling rate. If the energy is concentrated mostly below 4 kHz, a sampling rate of 8 kHz is sufficient to characterize the frequencies in human speech. In practice, a sampling rate of 16 kHz is used, meaning that the digital recording can reconstruct frequencies up to 8 kHz. This generates 16,000 samples per second, meaning that the interval between each measure is a fixed $\frac{1}{16,000} = 6.25\text{e} - 5$ seconds. This generates a massive 960,000 samples per minute and 57.6 million per hour of audio.

This illustrates a valuable insight about digitized audio recordings: A recording is simply a time series with a very high number of values per second. In other words, a recording is simply a vector of signed integers, often with a bit depth of 16.[2] This is valuable because it means we can approach speech audio the same way we approach time series. On the flip side, audio recordings face the same challenges as other time series data such as non-stationarity.

This non-stationarity of human speech poses a problem for the application of the Fourier transform, which assumes that the signal is stationary (Hammond & White, 1996). This issue is resolved by applying the Fourier transform on shorter windows, e.g., 20 of 25 ms, assuming that the signal is stationary within the windows. This is typically referred to as the short-time Fourier transform and yields as many distributions as there are windows. This is, for instance, the methodology used to obtain the spectrogram representations of a sound as illustrated in Figure 4.1c and Figure 4.1d, and how pitch is computed using the autocorrelation function as proposed by (Boersma et al., 1993), which is the common method used in the Praat software. A spectrogram visualizes the frequency distribution of a signal as it varies over time colored by the amplitude. This representation only yields additional information for complex signals such as human speech composed of multiple frequencies. The spectrogram representation in 4.1d is the equivalent of the speech

[2] A bit depth of 16 means that each value can take on $2^{16} = 65,535$ different values. When signed, this ranges from -32,768 (-1×2^{15}) through 32,767 ($2^{15} - 1$). The bit depth of a signal is typically referred to as quantization, which describes the discrete representation of the non-countable values in the analog signal (Camastra & Vinciarelli, 2015, Section 2.3).

waveform in Figure 4.1b. The spectrogram shows that frequencies below 2,000 Hz are present in the speech with most frequencies concentrating below 500 Hz. The F0 of this region corresponds to the vocal pitch of the speech (Rheault & Borwein, 2022). The higher frequencies are the harmonics.

4.2 Pitch Theory

The F0 of a sound is a physical property of a signal, denoting the number of times a sound wave repeats itself in a second and generally corresponds to what humans perceive as pitch (Camastra & Vinciarelli, 2015, p. 19). Pitch is popularly defined as the perceived "highness" or "lowness" of a sound (Klofstad, 2016, p. 2) and allows humans to order complex sounds from low to high. This can be viewed as a sound's dominant or most detectable frequency. This moves the definition of "sound" from the realm of physics to the realm of perceptions. When understood perceptually, the Merriam-Webster dictionary defines "sound" as the "sensation perceived by the sense of hearing".[3] This aligns with the term when describing "the sound of politics".

Pitch perceptions are not linear with F0 but instead follow the mel scale capturing that humans do not detect frequencies linearly (Stevens et al., 1937). For example, while we can tell the difference between 100 Hz and 200 Hz without further ado, we have more difficulties hearing the difference between 1,000 Hz and 1,100 Hz. The purpose of the mel scale is to correct this such as the equal distances corresponding to equal changes in the perceived pitch. The relationship can be expressed with the equation put forward by O'Shaughnessy (1987):

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right) \tag{4.1}$$

[3] https://www.merriam-webster.com/dictionary/sound.

I plot the relationship between mels and hertz in Figure 4.2, which shows the log-like functional shape between the mel and the hertz scale. Throughout the dissertation, I rely on hertz when reporting pitch measures.

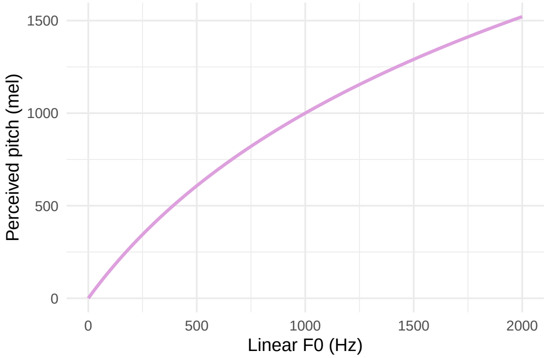


Figure 4.2: Mel vs. hertz for frequencies in the range 0-2000 Hz with a step of 10.

The pitch is a key feature of speech production. When humans speak, air is expelled from the lungs and propagated through the vocal folds (also called vocal cords), which cause the folds to vibrate. This vibration determines the F0 and the perceived pitch of a speech wave. The determinants of the F0 of a speech can be mathematically expressed with the equation proposed by Titze (1994):

$$\mathrm{F0} = \frac{1}{2\mathrm{L}}\sqrt{\frac{\sigma}{\rho}} \tag{4.2}$$

where L is the length of the vocal folds, σ is the tension applied to folds, and ρ is the mass or density of the tissue of the folds. The equation offers two main takeaways.

First, the pitch is heavily influenced by the physiology of a speaker's throat. The length (L) and mass (ρ) are largely determined by sex, but there is considerable intrasexual variation (Titze, 1989). Because males generally have longer vocal folds than females, the inverse relationship captures that the folds vibrate more slowly resulting in a lower vocal pitch for men (Puts et al., 2006, p. 284). The same goes for the density of the vocal fold tissues (ρ), which scales proportionally with pitch. Those physiological factors are more or less fixed within the same speaker and cause a bimodal shape with males speaking with an average $F0$ of around 100-120 Hz and females around 200-220 Hz, i.e. twice as high (Simpson, 2009). While the length (L) of the folds is largely physiologically determined, the effective length used in phonation can be altered by using the laryngeal muscles to either increase or decrease the rate of the vibration. For example, Pisanski, Mora, et al. (2016) show that humans exaggerate their body size by lowering their vocal pitch in sexual interactions. This is achieved partly by the tension applied to the vocal folds (i.e., σ in 4.2) and partly by manipulating the length of the surface area used to vocalize (i.e., L in4.2).

Second, the pitch is influenced by speech-specific variation. While the length can be partially manipulated despite being largely physiologically determined (Pisanski et al., 2018), tension applied to the folds (ρ) is dynamic and proportionally affects the pitch largely independent of the physiology.

The takeaways indicate that vocal pitch simultaneously encodes the speaker's physiology, specifically of the throat, and their motivations, intentions, and emotions. This provides concrete guidelines for how to approach the empirical analysis. Because the F0 is a "characteristic specific of each individual" (Camastra & Vinciarelli, 2015, p. 19), this creates substantial speaker heterogeneity as each speaker has a unique voice with a different baseline rate of vibration. This suggests that vocal pitch should be measured to remove physiological differences, such as standardizing by each speaker. I elaborate on this in Chapter 6 where the details of the pitch measurement are presented.

4.3 Why Annotations Matter

Utilizing audio recordings in empirical work is not without its challenges. The biggest challenge is arguably the lack of accurate annotations in the large and publicly available audio archives of political speeches. The question of how to obtain audio annotations in cases where they are lacking is the subject of Chapter 5. For now, we only consider why annotations matter. Annotating is not to be confused with labeling, which is to assign categories to data, for example coding the policy issue in a TV advertisement (Tarr et al., 2023) or classifying the tone in an utterance (Knox & Lucas, 2021). Annotating means imposing structure on unstructured data. These tasks are often used interchangeably, but address and solve different questions and issues.

The role and importance of annotations are best illustrated by considering the requirements to test one of the hypotheses in Table 3.1. Consider the first hypothesis as an example, $\mathbf{H}_1^{\mathbf{B}}$. Analyzing whether legislators speak with a higher pitch in dyadic speeches where the target is an outpartisan (i.e., when playing the partisan role) requires information along two dimensions. We need to know (1) *when* the speech occurs to compute the vocal pitch, i.e., the timestamps of when a speech starts and ends, and (2) *who* the speaker is to categorize the partisanship.[4]

The example above demonstrates the role and importance of annotations in three ways. First, annotations are only necessary

[4] We also need information about *what* is said to define whether the speech is dyadic, its topic, sentiment, and so on. The first two are about annotations, and the third is about alignment. I elaborate on how I achieve this in the corpus used in the dissertation in Chapter 6.

when dealing with multi-speech and multi-speaker recordings, such as entire legislative or campaign debates. If the recordings can be downloaded at the level of each speech, such as plenary proceedings in the German Bundestag, and the unit of analysis is speeches, no further annotation is required.

Second, annotations might have profound consequences for downstream measurements. Consider the case of pitch measurement, which we know is largely speaker-specific from Equation 4.2, with speeches as the unit of analysis and the original recording at the debate level. To analyze this recording at the speech level, the debate-level recording is segmented into its distinct speeches by timestamp annotations that denote the start and end times of each speech. The accuracy of these timestamps affects the validity of the empirical analysis through its implications for downstream measurement. If the timestamps are inaccurate, what appears to be a speech given by a single speaker might be a combination of speeches by two or more different speakers. This means that the pitch estimate of the speech reflects a weighted combination of the vocal fold vibrations of two or more different speakers. This is particularly problematic for features that are largely physiologically determined.

Third, the level of the annotations determines the level of the analysis. Following extant work (e.g., Dietrich, Hayes, & O'Brien, 2019), each substantive paper in the dissertation analyzes vocal pitch at the speech level suggesting that speech-level annotations are needed to segment the recording. If researchers want to study pitch at the level of utterances, sentences, or words, corresponding annotations are required independently of how pitch is computed. Note, however, that the level of analysis is unrelated to the computation of pitch, which is done on short windows of the underlying signal and then aggregated to the level regardless of the level imposed by the annotations. For example, a one-minute speech returns $\frac{60.0}{0.025} \times \frac{0.025}{0.0125} = 4,800$ pitch estimates using a window length of 25 ms with a shift of 12.5 ms. This returns a vector of estimates, which are then reduced to a scalar by, for example, the mean or the standard deviation, yielding speech-level measures of the average pitch and pitch modulation, respectively.

Annotations are clearly important for utilizing audio recordings in empirical research, and it not surprising that existing work using audio recordings in political science tends to use archives where annotations are already available and accurate. This includes the U.S. Supreme Court where annotations are available at the utterance level (Dietrich, Enos, & Sen, 2019; Knox & Lucas, 2021), the U.S. House of Representatives where annotations are available at the speech level (Dietrich, Hayes, & O'Brien, 2019), and the German Bundestag where the hierarchical structure of the archive permits downloading recordings at the speech level.

These archives are the exception rather than the rule, however. I exemplify this with three archives of legislative debates in which annotations are inaccurate, incomplete, or absent. In the Danish Folketing, annotations are inaccurate. The recordings include speech-level timestamps and speaker information, but closer inspection reveals substantial errors in the accuracy of the timestamps, potentially up to more than ten minutes. In the U.K. House of Commons, annotations are often incomplete. Recordings of question hours are fully and accurately annotated at the level of each speech, but recordings of regular legislative debates are incomplete. In this case, only recordings of question hours can be readily utilized in empirical research. Lastly, annotations are absent in the Irish Dáil Éireann, which essentially renders large-scale analysis of the archive infeasible. Notably, this is not only an issue for recordings of legislative debates but also pertains to recordings of committee proceedings (Kappos, 2024), campaign debates (Proksch, Wratil, & Wäckerle, 2019), and local government meetings (Barari & Simko, 2023). It even extends to single-speaker recordings where speech must be distinguished from non-speech if the purpose is to study speech production only (Neumann, 2019).

Chapter 5
Annotation Pipeline

THE PREVIOUS CHAPTER established the role and importance of audio annotations in analyzing sound empirically. In this chapter, I demonstrate how speech recordings can be annotated at the speech level using an automated annotation pipeline developed in Paper A without prior human-annotated data. The pipeline integrates three tools from computer science – speaker diarization, automatic speech recognition, and speaker recognition – to construct an end-to-end workflow that automatically annotates speech recordings with timestamps and speaker identity at the speech level. It relies entirely on pre-trained and open-source deep learning models and leverages a weakly supervised learning approach for speaker identification developed in the paper. The weakly supervised setup enables automation of the otherwise manual compilation of human-annotated data. The setup does not require any retraining when targeting new speakers, making it highly flexible and scalable to large archives. I show that audio recordings can be automatically annotated to an extent fully comparable to a human benchmark. The pipeline is applied to compile the dataset used in the empirical analysis described in Chapter 6.

5.1 Pipeline Architecture

The pipeline is constructed to annotate debate-level recordings with speech-level annotations in a fully automatic fashion. To achieve this, I integrate three methods that originate in computer science – speaker diarization (SD), automatic speech recognition (ASR), and speaker identification (SI) – into a two-stage workflow. In the first stage, speech-level timestamps are obtained to segment the recording into individual speeches. In the second, a speaker identity is assigned to each speech segment, as detected by the first stage. It is beyond the scope of the dissertation to explain these tools in detail, but a brief explanation of their respective purposes is required to understand the architecture and objective of the pipeline.

Computational Tools: Problems and Solutions

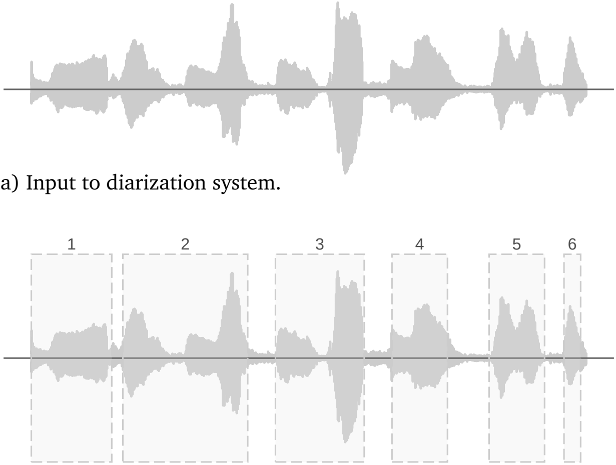
Each of the three methods aims to solve a distinct problem when analyzing audio recordings. The first, SD, serves the purpose of segmenting an audio recording into its separate speech segments by speaker (Park et al., 2022). State-of-the-art systems operate unsupervised, and speakers are only generically identified by their voice characteristics but not their identity. The output is a set of speech segments, each with a pair of timestamps denoting the segment's start and end times and a generic speaker label. The second, ASR, serves the purpose of automatically transcribing speeches and has been used by political scientists to transcribe campaign speeches when only the recording is available (Proksch, Wratil, & Wäckerle, 2019). I utilize ASR to set up a fuzzy string matching scheme where pre-existing speech-level transcripts (e.g., ParlSpeech V2) are linked to ASR-generated transcripts to retrieve reference audio for target speakers automatically. The third, SI, is a variant of speaker recognition (SR) and aims to identify the identity of a single-speaker recording. Virtually any SR system operates with supervised learning where a classifier is trained on manually compiled reference audio to learn the voice profile of each target speaker. I cast this as a weakly supervised learning problem by utilizing fuzzy string matching to compile references for target speakers automatically.

Stage 1: Timestamp Annotation

The first stage of the pipeline retrieves the timestamps associated with each speech in a recording. I obtain this by using the open-source and pretrained SD system from `pyannote.audio` (Bredin & Laurent, 2021; Bredin et al., 2020), which performs end-to-end diarization using neural network building blocks and achieves state-of-the-art performance on multiple datasets (Bredin, 2023). The system enables fine-tuning of individual components of the workflow and can be used off-the-shelf to achieve a high baseline accuracy. The input and output are illustrated in Figure 5.1. Each segment corresponds to a single speech uttered by a single speaker and has a corresponding tuple of timestamps with start and end times in the format hh:mm:ss.

Stage 2: Speaker Annotation

The second stage assigns speaker identities to the speech segments identified in the first stage. I conduct this using a cosine similarity approach allowing speakers to be flexibly identified and adapted such that no re-training is required when new target speakers are recognized (Tumminia et al., 2021). Because there are multiple targets, this amounts to a multi-class classification problem. The identification is done with the cosine function, pairwise comparing references to speech segments. A speaker is assigned when the most similar pair exceeds the classification threshold.



(a) Input to diarization system.

(b) Output from diarization system.

Figure 5.1: Illustration of input to and output from diarization system. An unsegmented recording is forwarded to the system and returns a segmented recording where each segment corresponds to a speech uttered by a single speaker. In the illustration, the system outputs a total of six speech segments.

Weakly Supervised Speaker Identification

Where conventional speaker classification is done in a supervised setup with manually compiled references for each target speaker to be identified, I leverage a weakly supervised learning approach developed in Paper A. The identification works like any other SI task, but the difference is in how the references are compiled. The method exploits fuzzy string matching to link and compare ASR-generated transcripts computed on the speech segments to pre-existing transcripts. When a speech segment is matched to a speech in a pre-existing transcript, it automatically retrieves information about the speaker's identity.

This is what makes the annotation pipeline fully automatic. Whenever a speaker can be located in a corresponding pre-existing transcript, the compilation of reference segments can be done automatically. The transcript need not correspond to the recording we want to annotate but only assumes that a speaker must be located in *any* pre-existing transcript. If so, references can be automatically compiled, and the speaker can be identified in any other speech recording whether it has a corresponding transcript or not. Luckily, this assumption is mostly trivial in the context of politics. Consider, for example, an application where we want to annotate a campaign debate between party leaders leading up to a national election. In this case, we can compile reference segments with the weakly supervised setup by utilizing the recordings and the corresponding transcripts of legislative debates where the party leaders are located with great likelihood.

5.2 Pipeline Analysis

I now turn to the validation of the pipeline to shed light on the extent to which audio recordings of political speeches can be automatically annotated. The automated annotations generated with the pipeline are validated against a human benchmark to score how computational annotation of speech audio compares to manual annotation. To the extent that recordings can be automatically annotated, the difference between automated and manual annotations should be minor. The validation data consists of manually annotated recordings of legislative debates in the Danish Parliament, which is an ideal case to test the generalizability and usability of the pipeline for at least two reasons.

First, the Danish language is not as widespread in machine learning as, e.g., English, Spanish, or German. This provides a test of whether the pipeline works on smaller languages. Second, legislative debates are a boundary case for testing the diarization system. Compared to most diarization applications, legislative debates are characterized by a large number of unique speakers taking unequal turns, both in frequency and duration, and their length, often more than five hours. In the following, I focus on a selective set of results. The full set of validation exercises is reported in Paper A.

Diarization Analysis

First, I validate the pre-trained diarization system's ability to detect speech segments. To do so, I compare the automated timestamps generated from SD to human-generated timestamps for the first 50 speeches in a randomly sampled recording of a debate in the Danish Parliament. The performance is evaluated using the Diarization Error Rate (DER) metric capturing the joint ability to detect speech segments and distinguish speakers. A lower DER indicates that the system detects speech segments and discriminates between speakers more accurately. The evaluation is reported using summary statistics of the DER and its components (false alarm, missed detection, and confusion) across five batches.

	False alarm	Missed detection	Confusion	**DER**
Avg.	4.0	5.3	2.9	**0.02**
Std.	3.3	3.8	6.1	**0.01**
Min.	0.5	0.6	0.0	**0.01**
Max.	7.2	11.1	13.8	**0.04**

Table 5.1: Summary statistics of the diarization performance. Units are in seconds except the DER, which refers to the metric itself. The evaluation is done across five batches.

The results are shown in Table 5.1. The pre-trained system performs incredibly well without any fine-tuning. The average DER is a low 0.02 with a standard deviation of only 0.01, corresponding to around one second of error per minute of speech. Importantly, this error mostly occurs because the system fails to detect speech (i.e., “missed detection”), secondarily because non-speech is falsely assigned as speech (i.e., “false alarm”), and less so because speakers are confused with each other (i.e., “confusion”). This shows that timestamps for speech segments can be automatically retrieved with the same accuracy as human-annotated timestamps.

The accuracy of automated timestamp annotations proves promising for the accuracy of downstream measurement. This is evaluated by comparing the pairwise average pitch for the automated and manually annotated speech segments. The measures are almost perfectly correlated with a correlation coefficient of $\rho = 0.99$. I illustrate this visually in Figure 5.2 where the estimates align perfectly along the 45-degree line. This shows that the automatically annotated timestamps are accurate in themselves but also generate accurate downstream measurement.

Speaker Analysis

Having established that speech segments can be automatically obtained with human-level accuracy, I now turn to the analysis of the speaker annotation. For this task, I evaluate the extent to which reference segments compiled with the weakly supervised setup can identify target speakers across 20 new recordings. To conduct the evaluation, I diarize each recording using the procedure outlined in the first stage to obtain speech segments similar to those illustrated in Figure 5.1b.

All reference segments and speech segments are then projected into x-vector embeddings (Snyder et al., 2018) using the pre-trained representations from pyannote.audio computed with

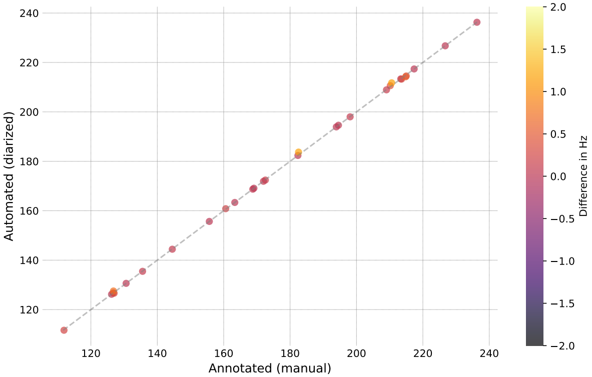


Figure 5.2: Relationship between pitch estimates computed with automatically and manually annotated timestamps. Results are only reported for speech segments ten seconds or longer for visualization purposes.

the time-delay neural network (TDNN) to construct fixed-length ($D = 512$) vectors from varying-length segments (Bredin et al., 2020; Coria et al., 2020). While the embeddings encode a variety of factors such as whether the speaker is emotionally aroused, the type of speech, and so on, the main variation is supposed to be the speaker's vocal characteristics. I visualize this in Figure 5.3, which shows reference embeddings colored by speakers after being reduced to two dimensions using the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm. The embeddings cluster by speaker, suggesting that the pre-trained x-vector can be used as input to a speaker identification model.

As a first test, I show the raw distribution of pairwise cosine similarity scores between speakers with and without reference audio. The similarities are computed based on the references and the speech segments in the 20 unseen recordings. Intu-

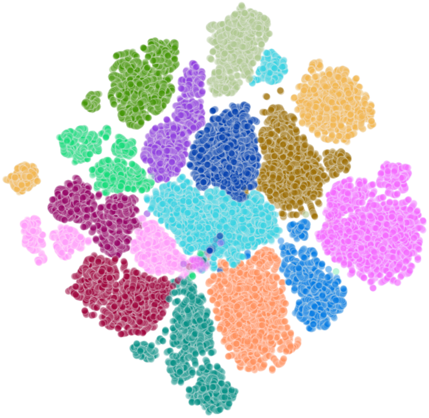
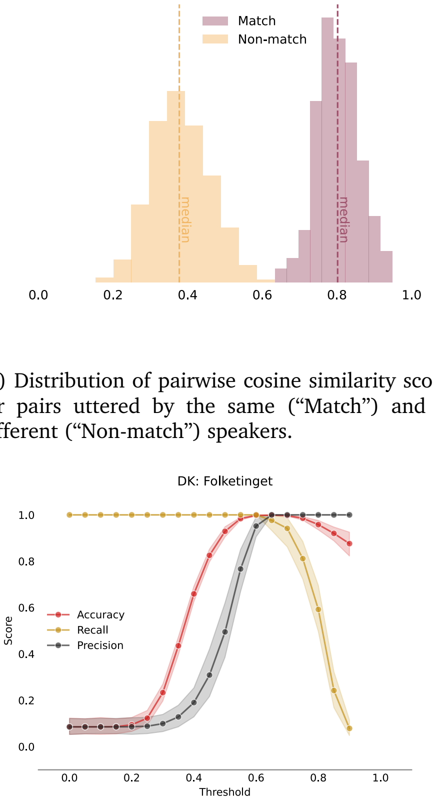


Figure 5.3: Visualization of reference embeddings in two dimensions. Embeddings are reduced using t-SNE.

itively, if the pipeline is able to annotate speakers automatically, the pairwise similarity should be higher when the pair belongs to the same speaker. I show this in Figure 5.4a. The similarity distribution is bimodal with peaks for matched and non-matched pairs. The average similarity is 0.80 for pairs uttered by the same speaker compared to 0.38 for different speakers. Notably, the two barely overlap, suggesting that speakers can be identified with almost full accuracy if the threshold is selected carefully. This is further supported by the classification results shown in Figure 5.4b. For a threshold of around 0.70, speakers can be classified with high precision and recall simultaneously. In total, Figure 5.4 shows that speaker annotations can be conducted automatically and with human-level accuracy using the automated pipeline.



(a) Distribution of pairwise cosine similarity scores for pairs uttered by the same ("Match") and by different ("Non-match") speakers.

(b) Classification metrics as a function of cosine similarity thresholds $\lambda \in \{0.0, 0.1, \ldots, 0.9\}$.

Figure 5.4: Speaker identification results. Panel (a) shows the pairwise similarity scores for pairs uttered by the same and different speakers. Panel (b) shows the classification metrics as a function of different thresholds.

Chapter 6
Empirical Strategy

Testing the theoretical expectations spelled out in Chapter 3 requires an empirical strategy with (1) repeated observations of the same legislator over time, (2) accurately annotated audio recordings, (3) aligned text and audio data, (4) biographical information about legislators, and (5) a speaker-dependent pitch scale. This chapter presents the setting, data, measurement, and modeling strategy used to study the “sound of politics”. Each substantive paper uses slightly different modeling strategies due to the distinct nature of the independent variables. I focus on the general strategy used throughout the dissertation that cuts through the paper-specific choices. This leaves out certain details for which I refer to individual papers.

6.1 Setting: Parliamentary Speeches

The empirical setting of the dissertation is speeches given in legislative debates in the Danish Parliament, Folketinget. Debates in legislatures occupy an integral component of democracies and have gained widespread attention from political scientists in recent decades with the increased availability of digitized archives and the advance of computational methods to measure quantities of interests (Bäck et al., 2021). While the substantive importance of legislative debates is questioned by scholars (e.g., Austen-Smith, 1990; Laver, 2021), it is undisputed that speeches provide parties and legislators with a unique opportunity to publicly communicate their positions and priorities to the (s)electorate (e.g., Hill & Hurley, 2002; Proksch & Slapin, 2012). Compared to roll call votes, which are highly constrained by party (Schwarz et al., 2017; Snyder Jr & Groseclose, 2000), particularly in party-centered systems, speeches are a way for legislators to individualize their behavior net of their party. This makes parliamentary speeches an ideal measurement device to study individually driven behavior, such as role effects on vocal pitch, where the theoretical mechanisms operate at the level of each legislator. In addition to being a valuable measurement device, parliamentary speeches add theoretical, empirical, and methodological value to the empirical strategy of the dissertation.

Theoretical Value

Theoretically, parliamentary speeches are important in the tasks and functions for each of the three roles considered in the dissertation. The structure of legislative debates is a testament to their partisan nature. Legislators are generally granted access to the parliamentary floor based on party membership and then deliver a speech in the context of a stylized debate between parties (Mollin, 2018, p. 209). Speeches also provide a mechanism for making representation work as the "public visibility of debates makes them a privileged channel for legislators to signal to their constituents that they are acting on their behalf" (Fernandes et al., 2021, p. 1036-37). Lastly, parliamentary speeches are a core task for legislators in governing positions. Ministers are tasked with introducing legislation on the floor and required to participate in the recurrent question hours where legislators, typically from the opposition, scrutinize and criticize the government (Green-Pedersen, 2010).

Empirical Value

Empirically, parliamentary speeches are an ideal case to study the text and audio of political speech jointly, providing an opportunity to analyze verbal and nonverbal communication simultaneously. Digitized transcripts and recordings are widely and publicly available across legislatures. This holds empirical value for at least three reasons. First, the audio enables the study of legislators' nonverbal speech, which can give insights into a legislator's vocal style independently of the text. This feature is exploited in Paper D. Second, the joint availability offers a multimodal dataset that allows analysis of the interaction between verbal and nonverbal speech. This feature is exploited in Paper B and Paper C. Third, the recordings make it possible to study the emotional arousal of speeches, which cannot be retrieved from transcripts (Cochrane et al., 2022). This feature is essential for the validity of vocal pitch as an indicator of emotional arousal.

Methodological Value

Methodologically, legislative debates are arguably the best type of data source to obtain repeated observations for the same speaker over time due to its granularity and long coverage period. Within-variation is a precondition for studying the theoretical expectations for Paper C and D where the hypotheses concern within-legislator variation. Legislative debates are a recurrent event in the parliamentary calendar, which means frequent participation by legislators. This allows the construction of a panel dataset with repeated measures of the vocal pitch for each legislator with only the availability of digitized audio recordings of the debates limiting the period covered in the panel. This makes it possible to use fixed effects to study the impact of political roles on vocal pitch.

6.2 Case: The Danish Folketing

The empirical case of the dissertation is the Danish Parliament, the Folketing, which is an appealing empirical case for several reasons, but one is certainly data methodological. While audio archives are widely and publicly available across political institutions and legislatures (Dietrich, Hayes, & O'Brien, 2019, p. 941), the Folketing has digitized audio recordings spanning two decades starting in 2000. At the time of writing and to the best of my knowledge, this is longer than any other institution except C-SPAN. Table 6.1, summarizes the duration of existing work on audio data in political science. The span covered by a corpus has implications for statistical power but also influences what can be learned from the recordings substantially. Most notably, the span of the Folketing's archive provides more within-legislator variation and offers an ideal case to study career effects (Paper C) and transitions in and out of position (Paper D).

The Danish Parliament also constitutes an interesting case for the emotional attachments of partisanship (Paper B). The Danish multiparty system is characterized by high party cohesion with legislators toeing the party line in legislative voting. While the party system is highly fragmented, it operates almost as a two-party system revolving around two blocs "blocs" of left-

Article	Country	Institution	Level	# Recordings	Period	Hours
Dietrich, Hayes, and O'Brien (2019)	United States	House of Representatives	Debate	863	2009-2014	6,432
Knox and Lucas (2021)	United States	Supreme Court	Case	NA	2010-2016	95
Dietrich, Enos, and Sen (2019)	United States	Supreme Court	Case	1,773	1982-2014	2,639
Rittmann (2024)	Germany	Bundestag	Speech	50,198	2011-2020	4,542
Arnold and Küpfer (2024)	Germany	Bundestag	Speech	15,553	2017-2021	NA
Neumann (2019)	United States	Senate	Speech	69,241	2017-2019	54
This dissertation	**Denmark**	**Folketing**	**Debate**	**2,282**	**2000-2022**	**10,139**

Table 6.1: Overview of data sources used in audio applications in political science and the corpus used in this dissertation.

and right-leaning parties (Green-Pedersen & Thomsen, 2005) but without the negative externality of "bipolar disagreement" (Christiansen, 2021, p. 756). The political tone remains moderate despite the *de facto* two-party system with lower levels of partisan polarization compared to, for instance, the United States (Lind et al., 2023). This arguably makes the Folketing a least likely test of whether partisanship is also reflected in the sound of politicians' voices.

6.3 Data: Multimodal Speech Corpus

The dataset used in the dissertation is a multimodal text-audio corpus of parliamentary speeches given in the Danish Parliament from 2000 to 2022. This period spans six national elections, 28 parliamentary terms, and 2,282 debates. Of those, 2,260 debates have digitized audio recordings. This contains 850,357 speeches with an average of 389 speeches and a standard deviation 249 for each debate.

Audio and Text Acquisition

The recordings are downloaded using web-scraping as `.m3u8` playlists. This format is a plain text file that specifies where multimedia files are hosted either locally or in the cloud. Each playlist is then converted into an audio recording stored in `.wav` format with a sampling rate of 16 kHz using the command-line tool `ffmpeg`. I retrieve the audio from two sources. The recordings are downloaded from the Danish Royal Library from 2000-2009 and the Parliament's self-hosted archive from 2010-2022. The transcripts are retrieved using two different sources. Transcripts obtained from the ParlSpeech V2 corpus (Rauh & Schwalbach, 2020) from 2000-2018 and `.xml` files scraped from the Parliamentary Hansard from 2019-2022. The two transcripts are homogenized and combined to form a full-text corpus of parliamentary speeches spanning 2000-2022.

Annotation and Alignment

The next step is annotation and alignment. First, the recordings are annotated using the automated annotation pipeline developed and validated in Chapter 5 (Paper A). The result is a set of timestamps denoting when speech segments start and stop as well as speaker identities. Next, the audio and text data are aligned. For this task, I match each recording to a unique transcript to create pairs wherein the alignment occurs. A total of 2,186 pairs are identified corresponding to about 97% of the available audio. Within each pair, each speech segment in the recording is matched to a single speech in the transcript by comparing ASR-generated transcripts on the audio segments to the written speeches in the transcript using fuzzy string matching. The transcripts match only approximately because the latter is alatim and the former is verbatim.[1] Text and audio are successfully aligned for 95% of the speeches, excluding procedural chair speeches.

[1] Applying ASR on the speech segments *de facto* aligns the text and audio at the speech level. However, I opt for the non-verbatim solution because it contains no misspellings or other lexical mistakes, which is crucial for constructing the independent variables in Paper B.

Corpus Preprocessing

When the text and audio data have been aligned, I preprocess the multimodal corpus. Each substantive paper uses slightly different preprocessing and partitions of the overall dataset, but they share five common steps. First, only speeches with aligned text and audio are kept. Second, procedural speeches are removed. This is done by filtering away all speeches uttered by chairs as these do not provide valuable information about when politicians change the sound of their voices. Third, speeches shorter than 40 words are removed for two reasons: Short speeches are typically procedural or interjections that carry little substantive information, and they are prone to measurement error when using audio data as the segment boundaries constitute a larger share of the overall segment. Fourth, only speeches given by legislators from Danish parties are included. The Danish Parliament has four North Atlantic mandates saved for two representatives from Greenland and the Faroe Islands. Their speeches are removed as their behavior and role orientations differ substantially from those of Danish politicians (Jensen, 2002; Stæhr Harder, 2022).[2] Fifth and finally, legislators who have given less than 50 speeches are removed for standardization purposes (see 6.4).

Additional Data Sources

The multimodal corpus is augmented with two additional datasets. The first is biographical information from the Danish Legislator Database (DLD) (Klint et al., 2023), which provides information about gender, education, birth year, and election status for all legislators from 1849 to 2022. From this, I com-

[2] The North Atlantic mandates are an intriguing case to study the link between descriptive and substantive representation in Paper C, but the sample size is not large enough for statistical analysis. A total of 19 politicians have given 2,827 speeches as a North Atlantic mandate in the Folketing during the period covered by the corpus. However, the average seniority is only five years.

# Step	Preprocessing	Sample size	# Debates	Avg. # of speeches
1	Full corpus	850,357	2,282	389
2	Annotated and aligned speech	511,444		
3	Procedural speech	449,972		
4	Short speech	435,525		
5	Danish parties	380,266		
6	Fifty speeches	377,937	2,145	176

Table 6.2: Sample size by each processing step.

pute the seniority and social class of each legislator used in Paper C). The second dataset contains information on government compositions and their duration. This is scraped from the Folketing's website and is used to calculate the start and end date of government spells for each legislator exploited in Paper D.[3]

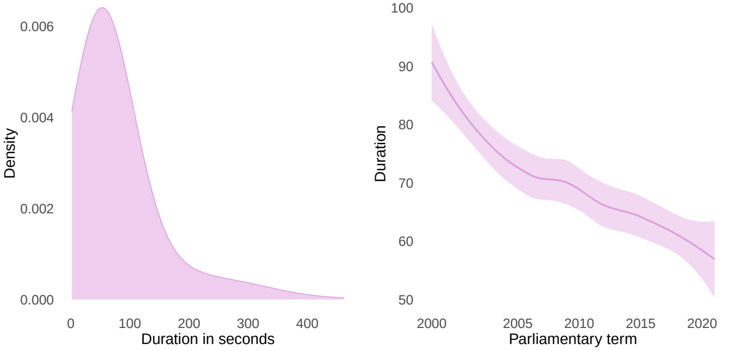
Corpus Description

The compiled and preprocessed dataset consists of 377,937 parliamentary speeches given across 2,145 legislative debates. The corpus contains 509 legislators giving 743 speeches on average with a standard deviation of 758. The average speech duration is 76 seconds with a standard deviation of 68 with a heavily right-skewed distribution (Figure 6.1a) and an overall downward-trending trajectory during the parliamentary terms covered in the corpus (Figure 6.1b).

6.4 Measurement: Vocal Pitch

The outcome of the dissertation is the speech-level average vocal pitch. The speech-level vocal pitch encodes the "sound" of a politician's voice and is used to measure the emotional arousal of a speaker in a given speech using the speaker-specific arousal scale outlined in Chapter 2. In the following, I describe how the

[3] https://www.ft.dk/da/folkestyret/regeringen/regeringer-siden-1953.



(a) Distribution of speech duration as a density with a bandwidth of 50.

(b) Development of speech duration as a function of parliamentary terms.

Figure 6.1: Duration of parliamentary speeches in the corpus.

vocal pitch is computed, the use of speaker standardization, and its robustness. Lastly, I validate the measure as an indicator of a legislator's emotional activation–composure using a series of validation exercises.

Computation

The vocal pitch is computed at the level of each speech using the `communication` package in `R` (Knox & Lucas, 2021). The package computes the F0 on the speech signal using two pitch detection algorithms with 25 ms hamming windows with 12.5 ms overlap. The hamming windowing minimizes the spectral leakage caused by the discontinuities arising from windowing. I exploit that the package estimates pitch by two algorithms to reduce measurement error by only considering a window estimate valid if both algorithms return a non-zero estimate. A non-zero estimate means the algorithm has classified the window as voiced speech, and a zero estimate that the window is classified as either silent or unvoiced speech. Keeping only voiced windows follows best practices of extant work (e.g., Dietrich, Hayes, & O'Brien, 2019). A speech-level estimate of the pitch is then computed by averaging over all the non-zero windows in a speech. This ensures that the resulting speech-level measure only relies on estimates from windows where it can be credibly estimated. This is illustrated in Table 6.3 with a hypothetical example with a speech with seven windows.

	Algorithm			
Window	FSV	MHS	Valid	Estimate
1	0.00	170.28		0.0
2	0.00	317.54		0.00
3	310.86	313.27		312.10
4	306.88	306.52		306.70
5	302.06	303.01		302.54
6	181.61	0.00		0.00
7	214.12	218.50		216.31
				284.41

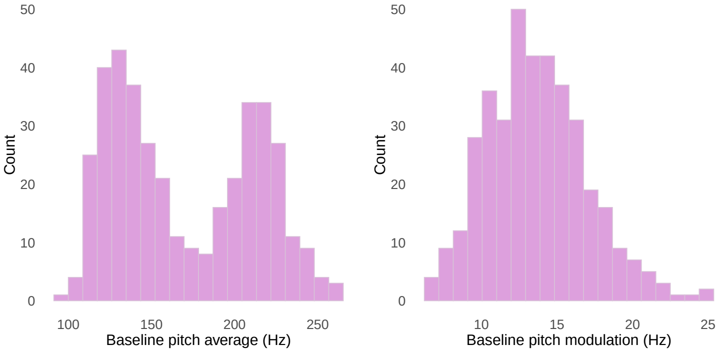
Table 6.3: Illustration of pitch measurement in a hypothetical speech with seven windows. The average is computed on only non-zero windows (3,4,5, and 7).

Standardization

After a speech-level estimate is computed for all speeches in the corpus, the measure is z-standardized by each legislator in the corpus. This maps each estimate to standard deviations below or above a legislator's average pitch. Higher values are indicative of greater emotional arousal (higher activation and lower composure) and lower values are indicative of less emotional arousal (lower activated and higher composure). I return to the validation of this in the next section. The mapping is done with the following equation:

$$\text{Std. pitch}_{ij} = \frac{\text{pitch}_{ij} - \mu_j}{\sigma_j} \tag{6.1}$$

where μ_j is the average of all speech-level average pitch measures for speaker j, and σ_j is the standard deviation of these averages. These quantities are computed across the entire corpus for each legislator. The distributions of the corpus-specific baselines are shown in Figure 6.2. The bimodal shape of Panel (a) shows the physiological and gendered nature of the vocal pitch whereas Panel (b) are more normally distributed.



(a) Distribution of baseline pitch averages.

(b) Distribution of baseline pitch modulations.

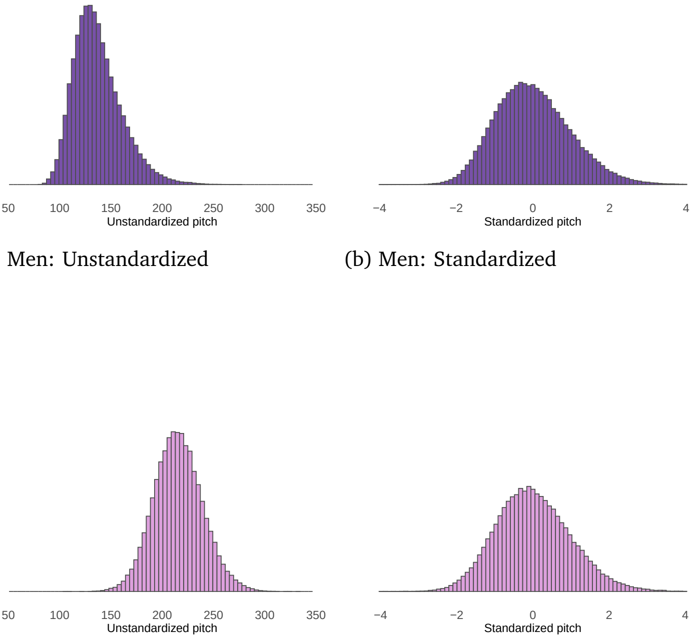
Figure 6.2: Distribution of legislators' speech-level pitch averages and modulations (i.e., standard deviations) measured in hertz (Hz).

Standardizing the vocal pitch solves two issues. First, it eliminates physiological differences between speakers that could confound the relationship between political roles and pitch. This effectively removes speaker heterogeneity and follows directly from the takeaways in Chapter 4, and aligns with existing work using vocal pitch as an indicator of emotional arousal in political science (Dietrich, Enos, & Sen, 2019; Dietrich, Hayes, &

O'Brien, 2019; Rittmann, 2024).[4] The effect is shown in Figure 6.3, which plots the unstandardized and standardized distributions of speech-level average pitch by gender. When comparing the unstandardized distribution for men and women (a + c), the distributions clearly center on different means, with men having a significantly lower speech-level average pitch than women. This difference is completely eliminated when comparing the standardized distributions (b + d). Second, it accounts for measurement errors in the computation of pitch. The calculation of vocal pitch is not a trivial task and may contain algorithmic dependency. Standardization effectively accounts for this "constant" by subtracting the legislator's mean and dividing by the standard deviation.

Standardization also affects the interpretation of vocal pitch in two ways. First, it heavily depends on the baseline. In this dissertation, a speaker's baseline is defined on the basis of parliamentary speeches. This means that positive deviations indicate higher than average arousal on the floor but not necessarily whether the legislator is more aroused in absolute terms. This illustrates an important insight. Standardization should be done within context, since the baseline in all likelihood changes. Otherwise, we might conflate emotional activation and composure with context effects. For example, politicians tend to speak with a lower pitch on the campaign trail (e.g., Touati, 1993), potentially shifting the baseline with more than a full standard deviation (Moez, 2024, p. 114). The same applies to comparative analyses. Levels of vocal pitch in the Danish Folketing and the U.K. House of Commons cannot be readily compared. Not because politicians are expected to have systematically differ-

[4] Standardizing essentially plays the role of speaker-specific HMM-classifiers used by Knox and Lucas (2021) to detect skepticism expressed by judges and the voice embeddings used by Rheault and Borwein (2019) to parse out speaker heterogeneity in classifying emotional arousal and valence from audio. In other words, standardization is similar to the use of fixed voice effects.



(a) Men: Unstandardized

(b) Men: Standardized

(c) Women: Unstandardized

(d) Women: Standardized

Figure 6.3: Distributions of unstandardized and standardized speech-level pitch estimates on the corpus used in the empirical analysis. Each plot only shows speech-level estimates with ±4 SDs for visualization purposes.

ent baselines, but because of the difference in debate style. The House of Commons is more agitated and aggressive, with speakers interrupting each other more relative to Scandinavian parliaments such as the Norwegian Storting (Søyland & Høyland, 2021) or the Danish Folketing. Second, standardization masks the absolute levels of arousal. This is not a drawback in itself, but it depends on the research question at hand. A relative measure is appropriate for this dissertation, as each hypothesis (see Table 3.1) is about pitch changes and not levels.

Robustness

An additional source of measurement error when analyzing audio signals is the act of recording itself.[5] Conditions such as microphone quality, placement, distance to source (i.e., the speaker), background noise, and room acoustics can all affect the digital signal's properties. The study of a single context, the Danish Parliament, eliminates some variability by holding recording conditions as constant as possible, but conditions might still change over time. While the span of the corpus used in the dissertation yields substantial benefits, the downside is a potential lack of robustness in the measurement.

To assess the robustness of pitch measures over time, I evaluate the extent to which it is unrelated to recording conditions. The sensitivity of vocal pitch is assessed by regressing the speech-level average and modulation of pitch and loudness on indicators of parliamentary terms. A low sensitivity is indicative of high robustness and a high sensitivity of low robustness. I analyze this using the adjusted R^2 model statistic. Recording conditions are expected to affect the pressure characteristics of the signal more than its periodicity. Loudness is a pressure characteristic, whereas pitch depends on the periodicity of the signal due to its relation to the autocorrelation function (Boersma et al., 1993). The results are reported in Table 6.4. As expected, pitch exhibits substantially less temporal dependency than loudness

[5] This type of measurement error differs from the one described in Chapter 4 where the analog-to-digital conversion may result in unrepresentative digital representations if the analog signal is not digitized properly with a sufficient sampling rate of bit encoding

and is entirely unrelated to the year of recording. In contrast, loudness shows strong temporal dependency as nearly 80% of the variation is explained by yearly information alone. This demonstrates that vocal pitch is robust to recording conditions, making it viable to study in a longitudinal corpus.

	RMSE	Adjusted R^2
Pitch average	44.0	0.00
Pitch modulation	9.28	0.02
Loudness average	2.91	0.80
Loudness modulation	0.59	0.41

Table 6.4: Model statistics from regressing pitch and loudness on parliamentary term indicators. Audio features are regressed using absolute levels: hertz (Hz) for pitch and decibel (dB) for loudness.

Validation

Vocal pitch as an indicator of emotional arousal has been extensively validated in psychology (Banse & Scherer, 1996; Bänziger & Scherer, 2005; Scherer et al., 2003). A selective set of those validations are replicated by Dietrich, Hayes, and O'Brien (2019) in the Supplemental Information in their study of the relationship between the gender of legislators and the engagement with which they talk about women's issues. However, the link has not been explicitly validated in the context of parliamentary speeches. In this section, I present a series of validation exercises investigating the extent to which vocal pitch, in the context of parliamentary speeches, is a (1) sufficient representation of the general sound of a speech, (2) reliable indicator of a speaker's arousal in a given speech (i.e., construct validity), and (3) methodologically trustworthy.

Content Validity

I first assess to what extent the vocal pitch adequately represents the sound of a speech. This can be thought of validating the content of the average pitch measure. Using only speech-level average vocal pitch (standardized by speaker) is, undoubtedly, a theoretically coarse and crude measurement strategy. To empirically assess this, I consider how speech-level average pitch correlates with speech-level modulation of pitch. Modulation captures variation in pitch and is found to be a strong perceptual screen for how voters evaluate the traits of political candidates (Damann et al., 2024). A high correlation indicates that speech-level average pitch is a strong representation of a speech's overall soundscape. I show the correlation using unstandardized vocal pitch in Figure 6.4, stratified by gender. The two features show a high correlation with Pearson correlation coefficients $\rho^{\text{men}} = 0.76$ and $\rho^{\text{women}} = 0.64$. The relationship is similar when using standardized measures (not shown here). This shows that while the speech-level average pitch is theoretically a coarse measure, it is empirically comprehensive as it encodes a great portion of the variance in the soundscape of a speech as well.

Construct Validity

Having established the content validity of using pitch to measure the overall sound of a speech, I then evaluate the construct validity of using speaker-standardized speech-level average pitch to measure a speaker's arousal in a given speech. For this task, I distinguish between convergent and discriminant validity (Adcock & Collier, 2001). Starting with convergent validity, I compare speaker-standardized speech-level average pitch estimates to manually labeled arousal scores in a randomly selected set of

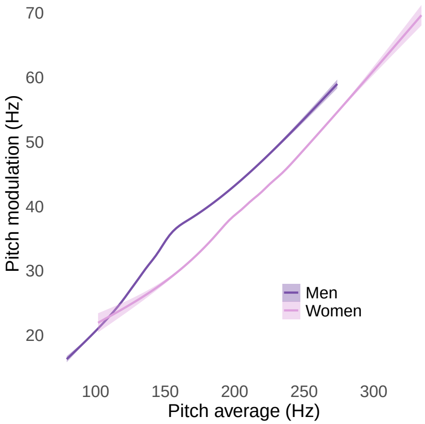


Figure 6.4: Relationship between speech-level pitch average and modulation measured in hertz (Hz) for men and women, respectively. The plot is shown for unstandardized values of pitch but standardized values are almost identical.

100 parliamentary speeches.[6] Following the procedure used by Cochrane et al. (2022), two coders were given the instruction:

> *On a scale from 0-10, where 0 indicates that the speaker was very subdued, 5 indicates that they were in a normal state of calm, and 10 indicates that the speaker was very animated, please indicate the emotional state of the speaker.*

The coders independently rated the emotional arousal expressed by the speaker in each of the 100 speeches in audio-only recordings. The reliability of manual arousal labels proved to be high, having an intraclass correlation coefficient (ICC) of 0.87.

[6] The validation exercise is conducted in Paper B and hence only uses dyadic exchanges. However, the exercise applies to all three substantive papers as this concerns the general link between pitch and arousal and not conflict-induced arousal per se.

This indicates that coders consistently infer the same amount of arousal in a speech. To assess how manual arousal labels align with the speaker-standardized speech-level average vocal pitch, the relationship between the two is shown in Figure 6.5. The manual labels and the standardized estimates align almost perfectly and have correlation coefficients of $\rho^{\text{coder 1}} = 0.89$ and $\rho^{\text{coder 2}} = 0.87$, respectively.

Turning to discriminant validity, I show that speaker-standardized speech-level average vocal pitch is not merely capturing textual proxies of arousal. If pitch variation encodes a speaker's arousal in a given speech, one would expect pitch to correlate with the sentiment and emotionality of a speech, but only to a certain extent (see, for example, Cochrane et al. 2022). Sentiment is computed using the Danish tool, Sentida (Lauridsen et al., 2019), with higher values denoting more positive language and negative values denoting more negative language. Emotionality is computed using the word embedding approach developed by Gennaro and Ash (2022) and uses the AFINN dictionary (Nielsen, 2011) to construct poles of neutral language and emotive language. Values above one indicates that a speech uses more emotive language than neutral, and values lower than one indicates more neutral than emotive language.[7] Consistent with expectations, Figure 6.6 shows that the relationship is negative and weak for sentiment and positive for emotionality, suggesting that pitch captures more than what is merely captured in textual proxies of arousal.

Methodological Validity

As a last validation exercise, I investigate the measure's general methodological validity by replicating the studies conducted by Dietrich, Hayes, and O'Brien (2019) and Rittmann (2024). The

[7] I refer to Appendix H in Paper B for further details on the computation of sentiment and emotionality.

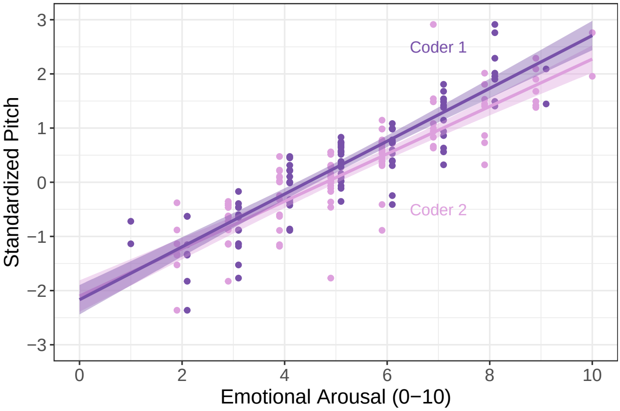
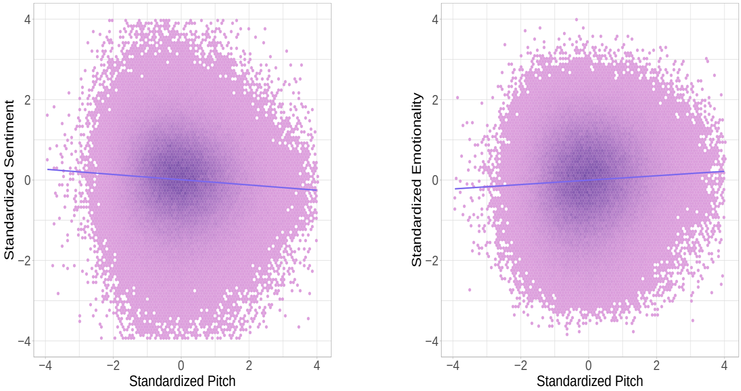


Figure 6.5: Correlation between manually labeled arousal scores for each of the two coders and speaker-standardized speech-level estimates of the average vocal pitch.



(a) Standardized Sentiment

(b) Standardized Emotionality

Figure 6.6: Correlation between standardized pitch and standardized sentiment and emotionality. Sentiment and emotionality are textual measures. The regression line is from simple bivariate linear regressions.

two articles use audio recordings of parliamentary speeches from the U.S. House (study 1) and German Bundestag (study 2) to study whether female legislators speak more emotionally when they mention women than when they talk about other policy issues and whether this differentiates them from male legislators. In line with their expectations, both studies find that women representatives speak more emotionally than men when referencing women relative to other issues.

This validation exercise re-analyzes the question raised in these studies but in the context of the Danish Folketing. Although this analysis provides substantive information about the descriptive-to-substantive representation link investigated in Paper C, the main focus is on the general methodological value of using pitch as an outcome in the context of parliamentary speeches.

To conduct the analysis, I follow the estimation strategy used by Rittmann (2024) and estimate legislator fixed effect regressions with robust standard errors. The outcome is speaker-standardized speech-level average vocal pitch. This model is estimated for speeches from the U.S. House and German Bundestag using publicly available replication data from Rittmann (2024).[8] For each of the two models, the predictors are a binary indicator of whether the speaker is a woman, whether the speaker mentions women using the dictionary of women-related terms developed by Pearson and Dancey (2011), and an interaction term. For the Danish Folketing model estimated in this dissertation, the second predictor is replaced with an indicator that captures whether a speech is mainly about "gender and equality", measured by an unsupervised Structural Topic Model

[8] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PAE6GB.

(STM) (Roberts et al., 2014).[9] This predictor is used instead of the women-related terms dictionary to make the result compatible with the findings reported in Chapter 7 from Paper C.

The results are reported in Figure 6.7. In accordance with both study 1 and study 2, the result indicates that women representatives in the Danish Folketing, on average, are more emotionally aroused, as indicated by the positive change in pitch, when talking about women's policy issues relative to when talking about other issues. Specifically, when the speech is mainly about "gender and equality", their pitch is 0.15 standard deviations higher relative to when talking about other policy issues. The same holds for men, presumably because of the polarization of the issue, but to a far less extent. The difference between the marginal effect also turns out to be almost identical: 0.094 (study 1), 0.073 (study 2), and 0.092 for the Folketing. Although also theoretically interesting, this replication further strengthens the validity and trust in using a pitch-based measure of emotional arousal.

6.5 Modeling: Legislator Fixed Effects

The empirical expectations developed in Chapter 3 and summarized in Table 3.1 involve causal claims about how political roles cause changes in a legislator's vocal pitch in one direction or the other. However, causal identification is ultimately limited by the fact that partisan polarization (Paper B), talking about a group's issue in a speech (Paper C), and governing positions (Paper D) are not randomly assigned. As a partial solution, I employ linear legislator fixed effects (FEs) regressions that adjust for unobserved, legislator-specific, and time-invariant con-

[9] A speech is defined as being mainly about "gender and equality" if this topic has the highest topic proportion. See the operationalization in Chapter 7 for more details.

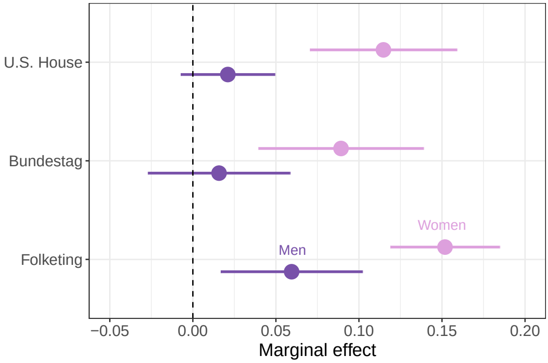


Figure 6.7: Marginal effect of legislator fixed effect models for the U.S. House (Dietrich, Hayes, & O'Brien, 2019), the German Bundestag (Rittmann, 2024), and the Danish Folketing (this dissertation). The outcome is speaker-standardized speech-level average vocal pitch.

founders (Imai & Kim, 2019). As the starting point, I define the following model:

$$\mathrm{y}_{ijt} = \alpha_j + \lambda_t + \rho D_{ij} + X'_{ijt}\beta + \epsilon_{ijt} \tag{6.2}$$

where y_{ijt} is the legislator-standardized vocal pitch of legislator j at time t in speech i, α_j and λ_t are legislator and time FEs, and X' is a vector of covariates used to parse out observed time-variant heterogeneity. Generally, the main predictor is D_{ijt}, one or a set of binary indicators denoting when a political role (partisan, representative, or governing) is switched on.[10] The coefficient ρ captures the effect of occupying a given role relative to not occupying a role on vocal pitch. Recall that pitch is standardized by speaker such that the difference denotes deviations from a speaker's own baseline. This model efficiently eliminates

[10] For Paper C, the indicators encode the seniority of a descriptive representative and not the role itself. In this paper, the role is switched on by the data partitioning where only descriptive representatives are included in the main analysis. See Chapter 7 for further information and the self-contained paper.

all unobserved time-invariant heterogeneity between legislators within the same time period. Each paper reports results from a model like (6.2) but with paper-specific variations in the specification of the time FEs, covariates, and the operationalization of D_{ij}. I elaborate on the paper-specific modeling choices when I report the results of each paper in Chapter 7. An overview of the main specifications are found in Table 6.5.

	Paper		
	B	C	D
Predictor	Outbloc target	Seniority of descriptive representative	Ministerial position
Level	Discrete	Continuous	Discrete
Values	{0, 1}	[0-20] years	{0, 1}
Modeling	{0, 1} indicator	{0, 1, . . . , 20} indicators	{-18, +9} relative indicators
Time FEs	Speech year	Parliamentary term	Speech year

Table 6.5: Model specifications.

Chapter 7
Core Findings

THIS CHAPTER PRESENT THE CORE EMPIRICAL FINDINGS of the dissertation. I assess each political role and present the core findings in each of the three substantive papers. The analyses of robustness and mechanisms and other auxiliary results are set aside in the dissertation report, but are found in the individual papers. Instead, the chapter focuses more on the general and core findings. Recall that unless otherwise specified, the outcome is the speaker-standardized speech-level average vocal pitch. For simplicity, I may refer to this as “vocal pitch” or “pitch”. Positive deviations indicate increased arousal, and negative deviations indicate lower arousal.

7.1 Partisan Role

Research Question

In Paper B, we ask whether legislators, in their role as partisans, speak with higher emotional arousal when faced with partisan conflict. Partisan conflict is investigated on two dimensions: polarization and policy. Specifically, we expect legislators to speak with a higher pitch in their speeches when targeting parties or legislators that are more polarized and with whom they disagree on specific policy proposals (see Table 3.1 in Chapter 3).

Sample and Operationalization

We study the research question by constructing speech dyads. A dyadic speech is defined as having a clear and unambiguous target in the form of parties or legislators. The target must be clear in that a speech is only dyadic if a single party or one or more legislators from the same party are mentioned. When multiple parties or legislators from two or more parties are mentioned, the speech is not classified as dyadic. This type of dyadic exchange is fairly common in legislative debates with 38.9% of speeches classified as dyadic. The analysis is conducted with this subset of dyadic speeches. A speech dyad is coded as polarized if the target is from a different bloc and as policy conflict if the target voted differently in the subsequent legislative vote.[1]

Model Specification

A model similar to (6.2) is estimated for each of the two in-

[1] Parties classified as left-leaning: the Red-Green Alliance (Ø), The Alternative (Å), The Green Left (F), the Social Democratic Party (A), and the Social Liberal Party (B). Parties classified as right-leaning: Danish People's Party (O), The Liberal Alliance (I), The Conservative Party (C), and the Liberal Party (V). This classification means that the ideology is captured with a binary indicator based on a party's bloc affiliation.

dependent variables with speech year as time FEs, controls for whether a legislator holds a ministerial position (in testing $\mathbf{H}_1^{\mathbf{B}}$) or is a member of a government party (in testing $\mathbf{H}_2^{\mathbf{B}}$) and the election cycle by coding whether a speech is given in an election year, and textual covariates for the number of words, sentiment, emotionality, and complexity as well as topic fixed effects.[2]

Main Results

The findings are reported in Figure 7.1. Starting with polarization, the estimate shows the difference in speaker-standardized vocal pitch when the target is from the outbloc relative to the inbloc. As expected, the difference is positive and statistically significant ($p < 0.00$). This shows that legislators speak with a higher pitch when they target the outbloc compared to when they target the inbloc. Notably, the relationship is largely invariant with respect to the inclusion of two-way FEs and covariates. When estimating a simple and bivariate linear regression with speaker-standardized vocal pitch as the outcome and an indicator of whether the target is an outbloc or inbloc, the estimate is $\hat{\rho}^{\texttt{naive}} = 0.18$ compared to the two-way FE estimate of $\hat{\rho}^{\texttt{twfe}} = 0.17$.[3] This indicates that vocal pitch conveys a different dimension of partisan polarization than what is captured by verbal measures alone.

[2] In the dissertation, I report results from 2 FE regressions for Paper B to unify the modeling strategy across the three substantive papers. I also add a set of different covariates to make the modeling as identical as possible. The results shown in Paper B are from simple linear regressions with only verbal covariates, but the results are substantially the same as those reported in the paper. I refer to the individual paper for details about the definition of specific variables. Note that the test of $\mathrm{H}_2^{\mathrm{B}}$ controls for the governing status of the legislator's party rather than the legislator's individual ministerial position since the data partition used to test the hypothesis only contains non-ministerial legislators.

[3] The naive estimate is computed from a bivariate linear regression with speaker-standardized vocal pitch as the outcome and an indicator of whether the target is an outbloc (= 1) or inbloc (= 0).

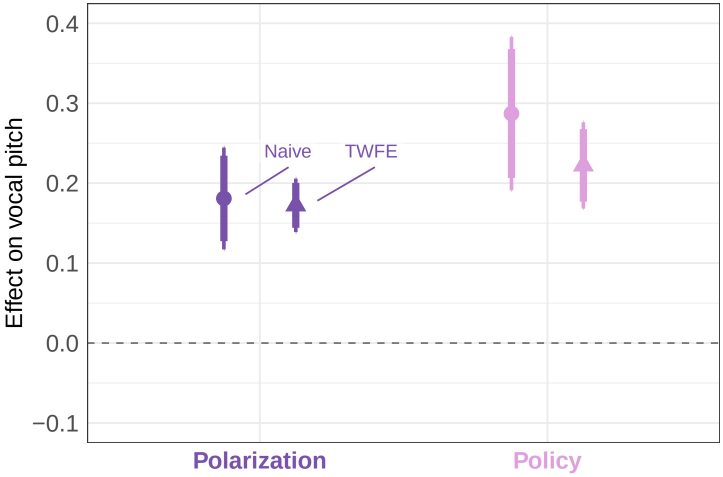


Figure 7.1: Marginal effect of partisan polarization and policy conflict on vocal pitch. Estimates are based on dyadic speeches in which the target party is outbloc (polarization, left) or voted differently in the upcoming legislative vote. Standard errors are clustered at the dyad level (speaker party ↔ target party). Thick and thin error bars are the model-specific 90 and 95 pct. confidence intervals, respectively.

Turning to policy, the estimate shows the difference in speaker-standardized vocal pitch when the target votes differently relative to the same as the speaker. As for polarization, the difference is positive as expected and statistically significant ($p < 0.00$). This illustrates that legislators raise their pitch when targeting parties or legislators from parties with whom they disagree on a policy. While not as invariant as the polarization estimate, the relationship is robust to the inclusion of two-way FEs and covariates. The naive estimate is $\hat{\rho}^{\texttt{naive}} = 0.29$ compared to the two-way FEs estimate $\hat{\rho}^{\texttt{twfe}} = 0.22$. Notably, this is not a lower-level replication of the partisan polarization. When including dyad fixed effects at the party level, the estimate drops from 0.22 to 0.15 but remains substantially and statistically significant. This indicates that vocal pitch conveys policy conflict, which is largely unrelated to ideological differences between parties.

Summary of Results

The results reported in Figure 7.1 strongly support the notion that legislators change their vocal pitch as a response to their partisan role when they find themselves in contexts of heightened polarization and policy conflict. This changes the sounds of legislators' voices. When the pitch is raised as a response to conflict arising by the nature of partisanship, the sounds of a voice also change, conveying the notions of being agitated and animated.

7.2 Representative Role

Research Question

In Paper C, I ask whether descriptive legislators, in their role as representatives, continue to be engaged throughout their parliamentary careers, as indicated by vocal pitch, when talking about the issues of their groups. Issue engagement, the emotional arousal with which a policy issue is raised, is contrasted with issue attention, the frequency with which a policy issue is raised, to disentangle the mechanism explaining the seemingly diminishing substantive value of descriptive representation as legislators spend time in parliaments. The theory spelled out in Chapter 3 generated two competing hypotheses. Both predict that issue attention declines over a legislator's career (H_1^C) but for different reasons. However, the accountability model posits that issue engagement also diminishes ($\mathbf{H}_2^{\mathbf{C}}$), and the motivational model posits that it remains unchanged ($\mathbf{H}_3^{\mathbf{C}}$).

Sample and Operationalization

The expectations are tested using two social groups: women and lower social class. For women, I construct an indicator taking the value 1 if the legislator is a woman and 0 if a man.[4] For lower social class, I construct an indicator taking the value 1 if the legislator is from a lower class, defined as one whose highest educational level is either primary school or vocational education, and 0 if not.[5] The data is partitioned into two samples, one consisting of female legislators and one of lower social class legislators. Group issues are then measured by the unsupervised Structural Topic Model (STM) (Roberts et al., 2014). The topics returned from the STM are manually labeled "gender and equality" and "workers and wages" being classified as women's and lower social class issues, respectively. Each topic share is standardized to make effect sizes comparable to vocal pitch. Career effects are measured with 20 binary indicators denoting whether a legislator has l years of seniority when giving speech i. The first year spent in parliament, $l = 0$, is the reference category.

Model Specification

A model similar to (6.2) is estimated for each of the two groups and for both issue attention and issue engagement (a total of four). The issue attention models have the (standardized) proportion of the groups' policy issues as the dependent variables and the (binary) seniority indicators as the predictors. The issue engagement models have the speaker-standardized vocal pitch

[4] On average, 36 to 45 pct. of the legislators taking the floor during the parliamentary terms covered by the corpus are women.

[5] On average, 9 to 22 pct. of the legislators taking the floor during the parliamentary terms covered by the corpus are from a lower social class.

as the outcome and the main predictors are interactions between binary seniority indicators and amount of attention devoted to the groups' policy issues. The models include election term indicators as time FEs and controls for whether a legislator is minister and election cycle by coding whether a speech is given in an election year.

Main Results

The findings are reported in Figure 7.2 as coefficient plots as a function of seniority. The important component is not the individual significance of the yearly estimates but the overall trends. Starting with issue attention, the upper panel shows the effect of spending l years in parliament relative to the first year on the attention devoted to issues related to women (left) or lower social class (right). As expected by both accounts, the attention devoted to women's and lower social class issues declines over time. A total of 7 and 15 coefficients are negative and statistically significant ($p < 0.05$ level) for women and lower social class, respectively. Most importantly, the groups follow similar trends with clear downward trajectories as legislators spend time in parliament. This shows that descriptive representatives pay less attention to the issues of their groups over time.

Turning to issue engagement, the main quantity of interest, the lower panel shows the effect of spending l years in parliament relative to the first year on the engagement devoted to issues pertaining to women (left) or lower social class (right). As expected by the motivational account, the issue engagement remains stable and invariant to the seniority of the legislators. As evident by the lack of trend in the coefficients, this holds for both women and lower social class. Although the estimates are positive and significant for most indicators of social class, this is entirely driven by the reference category. A precondition for the relevance of the lack of trend is that descriptive representatives are more engaged than non-descriptive representatives in the first place. Figure 6.7 in Chapter 6 shows that this is clearly the case for women representatives in the Danish Folketing. I refer to the individual paper for the result for lower social class.

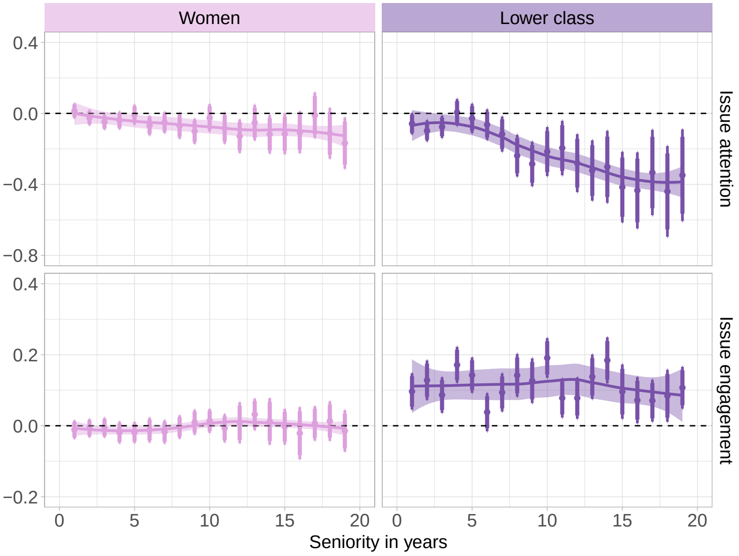


Figure 7.2: Marginal effects of spending $l \in \{1, \ldots, 20\}$ years in parliament relative to the first year ($l = 0$) on issue attention (upper row) and issue engagement (bottom row). The outcome for issue attention is the standardized share of group-related issues in a single speech. The outcome for issue engagement is the speaker-standardized speech-level average vocal pitch. A separate model is estimated for women and class for both attention and engagement (a total of four). Each model is estimated using only speeches given by descriptive representatives.

Summary of Results

The results reported in Figure 7.2 strongly support the notion that legislators change their vocal pitch in response to their representative role when addressing policy issues that are important to their constituents. When the vocal pitch is raised as a medium of representation, the sounds of a voice also change, conveying the notions of being engaged.

7.3 Governing Role

Research Question

In Paper D, I ask whether legislators lower the sounds of their voice when they hold a position of political power. I expect legislators to speak with a lower pitch when they assume a governing role relative to when they do not because they want to sound composed, signaling competence and dominance (see Table 3.1 in Chapter 3).

Sample and Operationalization

I analyze the research question by comparing the vocal pitch of legislators when they become ministers relative when they are not holding the position. The data is partitioned into a sample that includes a legislator's first spell as minister (if any) and not subsequent spells throughout their careers. From this partition, I construct relative indicators for the relative time since a legislator first entered government, with year zero denoting the first year as minister. These indicators vary from $\tau = -18$ to $\tau = +9$, meaning that a legislator is observed maximum 18 years before the first time serving as a minister in the corpus and maximum 9 consecutive years after becoming minister (see Figure 1 in Paper D for the distribution of the relative time indicators). The treatment year is defined as year zero, i.e., the first year a legislator assumes the position of minister.

Model Specification

A model similar to (6.2) is estimated with relative time indicators as the predictors with () and without () topic FEs (a total of two models). The topic FEs capture the fact that ministers are substantively constrained in which policy issues they address by comparing only within the same topic. The two models include the year of speech as time FEs and textual covariates for the number of words, sentiment, emotionality, and complexity of each speech. The reference category is the year prior to first entering a ministerial position ($\tau = -1$). This specification amounts to a fully dynamic difference-in-differences model, also called an event study, where the estimates prior to “treatment” (i.e., entering a ministerial position), are pre-trend coefficients with those after being the treatment effects.

Main Results

The findings are reported in Figure 7.3. The estimates show the difference in speaker-standardized vocal pitch between legislators holding and not holding a ministerial position at each time period relative to the reference period. As expected, there are no significant differences between legislators who have and do not have a ministerial position before entering government. This is evident by the fact that all pre-trends fluctuate around zero. This changes radically when legislators start serving in government with a clear and substantial lowering of the vocal pitch. The drop is sharp and abrupt upon transition and shows no signs of any learning effect regardless of whether topic FEs are included. The lack of sensitivity to speech topic suggests that the effect is unrelated to the difference in functional constraints faced by government members vis-à-vis opposition members. The average treatment effects on the treated (ATT) for the two models are $\mathtt{ATT} = -0.52$ and $\mathtt{ATT} = -0.55$, respectively.[6] This indicates that legislators substantially lower their vocal pitch to align with the power associated with their role.

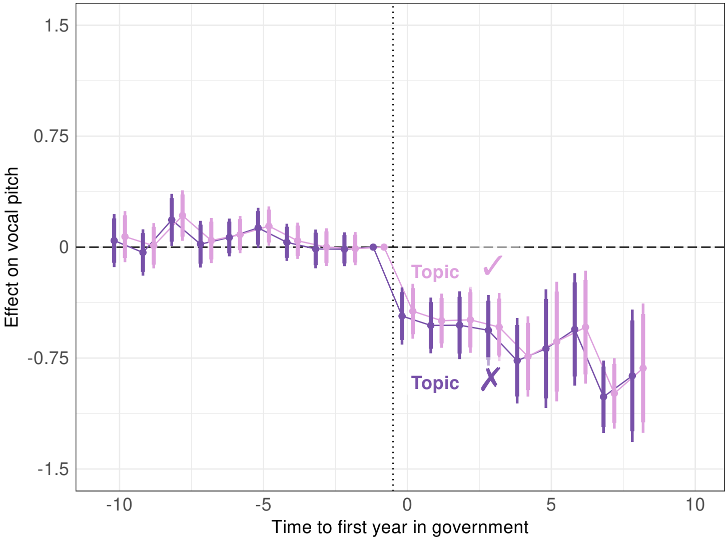


Figure 7.3: Event-study estimates by the relative time to first year assuming a governing role. The effects are shown using a fully dynamic difference-in-differences specification estimated with the TWFE estimator. The standard errors are clustered at the legislator-year level. Thick and thin error bars are the model-specific 90 and 95 pct. confidence intervals, respectively. Note that the x-axis is censored at $\tau = -10$ for presentational purposes.

[6] Note that this departs slightly from the ATTs reported in Paper D due to the inclusion of a different set of covariates in the modeling strategy for the findings presented in the summary report of the dissertation.

Summary of Results

The results reported in Figure 7.3 strongly support that legislators adjust their vocal pitch to align with the expectations that come with assuming a governing role. When the pitch is reduced in response to the traits expected from politicians who carry political power, the sounds of a voice also change, conveying the notions of being composed, calm, and controlled.

Chapter 8
Conclusion

THIS DISSERTATION HAS STUDIED THE SOUND OF POLITICS. Admittedly, this is an ambitious endeavor, yet I believe that the dissertation presents findings that strongly indicate that politics *does* has a sound. In a series of self-contained articles, I have shown that politicians vary their sound of their voices, as it manifests in their vocal pitch, systematically in ways that align with what we would expect and predict from existing political science theories. This indicates that politicians are not only strategic in what they say – that is what words are used – but also in how they say it – that is how words are spoken. In other words, to fully understand politicians and their behavior, it is not sufficient to focus solely on words, but also how those words are spoken. In this concluding chapter, I summarize the core findings of the dissertation, outline its core contributions and limitations, and discuss the outlook for future research.

8.1 Summary of Findings

The dissertation began with two research questions to answer the methodological and theoretical challenges of using audio data to study politics. The first question, *the extent to which political speech recordings can be automatically annotated*, pertains to the methodological challenge and is answered in Paper A. This article paves the way for the three substantive articles by developing an automated annotation pipeline that can annotate speech recordings with human-level accuracy at scale. The accuracy of the automated annotations enabled the compilation of the multimodal corpus used in the dissertation.

The second question, *when politicians change the sounds of their voices*, pertains to the theoretical challenge and is answered in three self-contained articles (B, C, and D). Each of these articles used speaker-standardized speech-level average vocal pitch as a measure of a speaker's emotional arousal to characterize the sounds of a speech as legislators assume core political roles in liberal democracies. In the following, I briefly summarize each article's core finding and discuss how it adds to the existing political science literature.

Partisan Role

In Paper B, we found that in their role as partisans, politicians speak with a higher than average vocal pitch when targeting parties or politicians from parties with whom they disagree ideologically or on specific legislation. This indicates that politicians change the sounds of their voices when facing partisan conflict to signal their agitation and disagreement. The article adds to and advances the existing literature on partisan and policy conflict by showing that polarization and bill-level conflict are conveyed not only in what politicians say but also in how they say things. In other words, what might appear as a deliberative and unifying debate when analyzing the text of speech may, in fact, be conflictual and filled with animosity when analyzing the audio of speech. This expands our thinking of how politicians communicate conflict to citizens and carries crucial lessons for how affective polarization should be studied. Likewise, since partisan conflict is conveyed in the sound of a politician's voice when targeting outpartisans, accounts of elite polarization should consider not only what politicians say to each other, but also how when characterizing the amount of conflict between politicians. The understanding of conflict as being transmitted in the sound of a politician's voice arguably aligns more closely with how humans perceive conflict than positional disagreement and is an intriguing topic for future research.

Representative Role

In Paper C, I found that in their role as representatives, politicians who share descriptive characteristics with social groups speak less frequently about the policy issues of their groups throughout their political careers, but that the vocal pitch remains the same when talking about the issues. In other words, while the issue attention declines over time, the issue engagement remains. This indicates that the decline in issue attention devoted by descriptive representatives is more due to structural and functional constraints faced throughout their careers than to an underlying change in the motivation, commitment, and engagement to represent their groups. The article adds to the existing literature on the link between descriptive and substantive representation, and more generally representational links, by showing that it matters deeply whether we understand substantive representation as issue attention – the frequency with which an issue is raised – issue engagement – the emotional intensity with which an issue is raised – or a combination. This expands our knowledge about the mechanisms of the descriptive-to-substantive representation link and suggests that more effort should be devoted to defining when descriptive representation holds representative value and hence when affirmative action policies work as intended. Specifically, addressing how voters value attention relative to engagement in the representational relationship and the amount of attention needed to compensate for a lack of engagement or vice versa are important topics for future work.

Governing Role

In Paper D, I found that when assuming governing roles, politicians speak with a lower than average vocal pitch independently of which policy issues they address in their speeches and how long they have assumed the role. This indicates that politicians change the sounds of their voices when they have political power to signal composure, a trait associated and expected from political leaders. The article provides insights to multiple literatures. In particular, it shows that the baseline vocal pitch of a politician is not the only way to display power. Specifically, the results indicate that the *same* politician varies the vocal pitch to change the impressions that the speaker leaves off, independently of the baseline. When in power, this manifests itself in a lowering of the pitch to increase perceptions of the politician's current power. The article also adds to the electoral and rhetorical "costs of governing" by arguing that lowering the vocal pitch may come with a rhetorical reward that might offset the costs associated with power. This expands our thinking on how politicians communicate power to citizens and potentially reconciliates seemingly contradictory laws of incumbency advantage and shrinking support (Cuzán, 2015). Specifically, while governments regularly lose votes when in power, the individual minister might be able to offset the costs of governing by lowering the pitch to sound composed, signaling valued traits such as competence and dominance. It is an intriguing avenue for future research to test this relationship and investigate the amount of pitch change needed to compensate for the costs.

8.2 Core Contributions

The dissertation consists of four self-contained articles that each have a set of contributions. This section highlights a selective set of contributions from each paper and the more general theoretical, empirical, and methodological contributions of the dissertation.

Theoretical Contribution

As paradoxical as it may seem, the purpose of theory becomes even clearer the more data is available. When working with large datasets as in this dissertation, developing falsifiable hypotheses becomes crucial in navigating the embarrassment of riches to distinguish between promises and perils. Using theory to develop hypotheses avoids the pitfalls of the "garden of forking paths" that come with the many degrees of freedom provided by large datasets (Gelman & Loken, 2013). I have handled this problem by developing a unified framework linking different political roles to variation in vocal pitch. This framework has allowed me to test foundational theories about political roles on a new data source. In doing so, I have also developed new theoretical concepts. For example, Paper C introduced the concepts of issue attention and issue engagement to disentangle strategic and motivational accounts of the descriptive-to-substantive representation link, while Paper D introduced the notion of "rewards of governing".

Empirical Contribution

Empirically, the dissertation has presented evidence consistent with the notion that to fully understand politicians and behavior, we need to study not only what they say but how. Specifically, I have presented findings showing that politicians vary their sound of their voices, as it manifests in their vocal pitch, systematically in ways that align with what we would expect and predict from existing political science theories. This has unlocked new knowledge about foundational theories and concepts of political science. For example, Paper B showed that partisan conflict is communicated in the sound of politicians' voices and not only in issue polarization.

Methodological Contribution

The biggest contribution of the dissertation is arguably methodological. The dissertation strengthens and maintains the foundation for using audio data to study political and social behavior as a larger research agenda. The sketches of this agenda have been set out by previous work that has employed audio recordings to study politically relevant behavior (e.g., Arnold & Küpfer, 2024; Dietrich, Enos, & Sen, 2019; Dietrich, Hayes, & O'Brien, 2019; Knox & Lucas, 2021; Neumann, 2019; Rheault & Borwein, 2019, 2022; Rittmann, 2024; Tarr et al., 2023). The dissertation contributes and expands this literature significantly. Not only does it show the general value of audio as data, it introduces the so far largest collection of aligned text-audio speech data used to study politics spanning 22 years and more than 10,000 hours. This data is collected with the automated annotation pipeline developed in Paper A, which other researchers can use to collect their own text-audio data when the software is made public.

8.3 Future Avenues

There is no shortage of future work and questions based on the dissertation. The research questions studied in future work are closely related to both the limitations and the challenges of working with audio data. In the following, I focus on the general prospects of using audio data rather than the more specific limitations that arise from the well-known and obvious shortcomings of using observational data to conduct causal inference (Angrist & Pischke, 2009) or using a single case – the Danish Parliament

– to draw wider generalizations. However, the dissertation obviously has drawbacks in both regard.

The Computational Cost of Audio Analysis

The main limitation of using audio data to study political speeches is the computational cost of analyzing the audio. Large-scale audio analysis requires substantially more computational power than large-scale text analysis due to the size of the data. A textual representation of a 30-second speech of 75 words amounts to 150 bytes, and the audio representation amounts to 1.3 million. This puts more computing demands on the user. For example, using a GPU makes automated annotation 12-15 times faster than using a CPU. This makes cloud computing or access to GPUs a key necessity for integrating audio in empirical research.[1] In the future, as computational power becomes a less scarce resource, the computational challenge is likely to be significantly reduced, but at the time of writing, it remains an obstacle to integrating audio data in empirical research unconditionally.

Moving Beyond Pitch

Another limitation of existing work using audio data, including this dissertation, is the narrow focus on pitch (for important exceptions, see e.g., Knox & Lucas, 2021; Neumann, 2019). In all three substantive papers, vocal pitch is the outcome. Although this mitigates the "garden of forking paths", and pitch is more robust to changes in recording conditions than spectral or pressure features (Vainio et al., 2023), it narrows what can be learned from audio data considerably. Future work on audio data should devote considerable attention to how other features, possibly in combination, can be used to study politically relevant behav-

[1] All analyses in the dissertation are conducted using the Ubuntu 20.04 operating system with NVIDIA RTX 5000 GPU with 16GB RAM.

ior to obtain more fine-grained measures, for example, separate measures of conflict-induced arousal, representational-induced arousal, and power-induced composure.

A promising way to move beyond pitch is to measure concepts that are not related to the emotionality of a speaker, such as accents and dialects. Dialects are not conveyed in speech transcripts but require analysis of speech audio as it is about how vowels and consonants are produced and pronounced by speakers. Dialects and accents have clear implications for politics, as they can serve as important identity markers and heuristic signals in multilingual or linguistically heterogeneous societies (Kanngieser, 2012; Ricks, 2020). Voters use accents as a cue to draw inferences about a candidate's policy positions and priorities (Amira et al., 2018; K. Ash et al., 2020). Dialects may even constitute a political cleavage due to their connection to the center-periphery cleavage.[2] Language cleavages are evident in the political landscape in multiple countries such as Switzerland, Belgium, and Canada (Dassonneville et al., 2022) and in contemporary support for radial right parties (Ziblatt et al., 2024).[3] Studying accents and dialects is a clear example of audio data offering insights that are completely lost in transcription and is a way to move beyond pitch.

Strategic or Sincere Sounds?

The question of whether politicians change the sounds of their voices strategically or sincerely is central to consider before dis-

[2] In their canonical writing, Lipset, Rokkan, et al. (1967) define the center-periphery cleavage as a "conflict between the central nation-building culture and the increasing resistance of the ethnically, linguistically, and religious subject population in the provinces and peripheries" (p. 14).

[3] In Germany, support for the radical right tends to be geographically clustered in historically peripheral areas where certain communities have persistent and visible lower-status cultural markers such as a nonstandard dialect that make residents more prone to feel "left behind" and harbor more anti-establishment attitudes and out-group discontent (Ziblatt et al., 2024, p. 1480).

cussing the democratic consequences and implications of the sound of politics. The core of this theoretical dispute revolves around whether humans can emulate physiological effects that have biological origins. As noted by Knox and Lucas (2021), it is highly likely that trained speakers, such as politicians, can imitate virtually any vocal behavior, including those that have biological origins. That trained speakers can emulate physiological effects is already evident by the fact that most of the research on vocal expressions of emotions is based on actor portrayals (Scherer et al., 2003, p. 232-233). This rules out the possibility that changes in pitch are completely sincere by definition (Knox & Lucas, 2021, p. 651).

However, the fact that politicians, as trained speakers, can intentionally change the sound of their voices does not imply that they always do so for strategic purposes. Deciphering the true motives of humans is inherently difficult, and so is it for politicians. Like humans, politicians are likely to change the sounds of their voices, some times for strategic purposes, sometimes of strict sincerity. This makes it a question of *when* politicians are expected to be strategic and sincere, not *if*. This moves the question from theory to design. To see this, consider the research design in Paper C. In this design, the same legislators are studied over time, meaning that the strategic incentives change as legislators enter new career stages. If issue engagement is more strategic, this should change as the incentives change, and if not, it is more indicative of sincere behavior. Likewise, the sound of politics as it unfolds in parliamentary debates is likely more indicative of politicians' sincere motives than in campaign debates, or at a minimum less strategic. Developing more sophisticated research designs that allow us to disentangle the question of strategy vs. sincerity is important in future work to understand the scope of using audio recordings to understand elite behavior.

Although deciphering the true motives of politicians is inherently difficult, motives also operate perceptually. This moves the focus from true motives to perceived motives. This dimension is crucial in a principal-agent relationship such as the constituency-representative linkage where the representative, the agent, is accountable to the constituents, the principal (Przeworski et al., 1999). In this relationship, voters delegate decision-making authority to the politicians, expecting them to act in ways that reflect the constituency's preferences and interests. However, due to information asymmetry and different incentives, politicians can sometimes pursue their own agendas or prioritize other interests, leading to potential agency problems such as reduced accountability and responsiveness (Gailmard, 2014). This principal-agent model of representation is essentially "promissory" in the sense that politicians make promises to voters, which the latter can sanction by rewarding or punishing politicians at the upcoming election for acting or not acting according to the promise (Mansbridge, 2003, p. 516).[4] Because voters have less information about whether politicians act on their promises, managing perceptions is central to the agent's effort to sustain the relationship.

Vocal signals such as vocal pitch are particularly efficient in this effort. In his seminal work on legislators' homes styles, Fenno (1978) draws on microsociologist Goffman (1959) and his notion of the "presentation of the self". While Goffman is interested in everyday encounters, the idea squares perfectly with the world of politics and the function of political speeches. When

[4] Perceptions also plays an important role in the "selection" model of representation (Mansbridge, 2011). In this model, the focus is not on how principals can *sanction* the agents but in how the voters can *select* politicians with aligned goals in the first place. The fact that voters generally prefer lower-pitched politicians (e.g., Klofstad, 2016) poses a threat to the selection model of representing to the extent that voice pitch does not signal true abilities (Klofstad & Anderson, 2018) and true intentions (Feinberg et al., 2018).

speaking, a presenter, that is, the politician, leaves off two different signals: “the expression that he *gives* and the expression that he *gives off*” (Goffman, 1959, p. 2). The former expression relies on verbal signals, and the latter on nonverbal signals. According to Fenno (1978), the nonverbal and more theatrical expressions are more important to the impressions that a presenter, that is, the politician, leaves off. This happens to be the case because the audience *thinks* that the verbal expressions are more controllable than the less controllable nonverbal expressions. Because of this, voters use nonverbal signs as a sanity check of the verbal, giving it a highly promissory dimension (Fenno, 1978, p. 54-55).

The interpretation of Goffman (1959) proposed by Fenno (1978) is highly supported by social psychologists in their work on attribution theory. When trying to filter out self-presentation, people rely on two rules. The first is called the “discounting rule” (Kelley, 1967) and states that behavior must not be taken “as a reflection of his or her true nature” (Kraut, 1978, p. 381). For political speeches, this means that the true policy positions and priorities of a politician cannot be recovered from the verbal content of speeches, such as the topic, choice of words, and speech length (Dietrich, Hayes, & O’Brien, 2019, p. 943). The second is called the “controllability rule” and builds on the work of Goffman (1959) to argue that individuals should form impressions based on the elements that the person is least able to deliberately and consciously control (Kraut, 1978, p. 381). When this rule is applied to political speeches, voters are expected to assign more weight to more uncontrollable signals, such as vocal pitch (Ekman et al., 1991, p. 134) relative to more controllable signals such as the verbal content of speeches and other nonverbal behaviors (Dietrich, Hayes, & O’Brien, 2019, p. 943).

When this is applied to the principal-agent relationship, the sound of a politician’s voice is likely more influential in evaluating whether the agent’s goals align with the principal than what the politician says, and often “become the basis for constituent judgment” (Fenno, 1978, p. 55). This makes changing the sound of their voices, for example, by lowering or heightening the vocal pitch an efficient tool, i.e., a strategy, for politicians to manage the constituency-representative linkage and voters’ perceptions of their motives.

Consequences of The Sound of Politics

The dissertation has considered the sound of politics at the elite level, but future avenues should consider how the mechanisms and concepts translate to the voter level. For example, future research can investigate whether voters evaluate partisan conflict mostly through text or audio (Paper B), how voters draw inferences about issue attention and issue engagement (Paper C), and whether a lowering of pitch can change perceptions of a politician’s traits or whether perceptions are sticky and the extent to which this depends on a speaker’s baseline (Paper D).

More generally, it is crucial for future work to study the wider democratic implications of the sound of politics in at least two ways. First, it is worth exploring the extent to which voters read, see, or hear politics to evaluate the overall effects of the sound of politics. This dissertation has studied the sound as it manifests in parliamentary speeches. Although most voters are certainly not exposed to parliamentary debates directly, speeches are reported in TV or radio news or posted on social media platforms. Second, the extent to which politicians can shape how they are perceived by changing the sound of their voices is essentially a double-edged sword. On the one hand, it can have unfavorable consequences for democratic accountability and representation if politicians, for example, can change invoke perceptions of issue engagement if they are not sincerely engaged. However, it can also have favorable consequences if voters are able to accurately detect which politicians are sincerely engaged in issues important to the voter. Getting a grasp of such questions is crucial in future avenues of the use of audio recordings in political science.

8.4 Concluding Remarks

This dissertation has broken new ground in studying political speeches and learning about the behavior of political elites. Although each individual paper adds to existing questions in political science in different ways, the outlook of the dissertation is broader in scope. The dissertation started as a motivation to investigate what can be achieved by using audio data to study politics. This quest proved to be more challenging than initially assumed, particularly due to computational and methodological challenges, but the dissertation has shown that political science *can* benefit, possibly and arguably significantly, from using audio recordings in empirical research. The jury is still out on the degree to which audio as data might transform research agendas in the same magnitude as text as data, but the dissertation has shown that valuable insights can be gleaned from audio recordings with simple tools.

Summary

The sounds of the human voice convey meaning and information that go beyond what is conveyed by words alone. This applies to all aspects of human and social interaction and likewise to politics. The way politicians modulate their voice transmits signals about their priorities and motivations independently of what they say. Given the importance function of speech and debate in politics, it is no surprise that political scientists have studied them in nearly every subfield of political science, from legislative studies to international relations, to understand, explain, and measure the behavior of political elites. However, quantitative speech studies have been conducted almost exclusively through transcripts, i.e., text, even when the corresponding recordings, i.e., audio, are available. Focusing solely on the text of speech, and hence what politicians say, inevitably strips away valuable information about politics and elite behavior.

Setting out to study the sound of politics as an outcome as it manifests in parliamentary speeches in the Danish Folketing, the dissertation takes as its starting point that *politics is difficult to understand in silence but benefits from hearing the sound of politicians as they speak*. To the extent that the sounds of politicians' voices transmit valuable information beyond the words themselves, this enables us to interpret, understand, and explain the behavior of political elites in new ways and from new perspectives. I present compelling evidence that politics has an independent soundscape transmitted solely in how politicians vary the sounds of their voices in an acoustic sense. The sound of

politics, as expressed in speaker-standardized speech-level average vocal pitch, is investigated along three core dimensions of liberal democracies and leverage an automated annotation pipeline developed in the dissertation to compile the multimodal dataset used in each analysis. The dissertation consists of four self-contained papers (three single-authored, one co-authored).

PAPER A is a methodological article in which I construct an automated annotation pipeline that facilitates large-scale annotation of political speech audio recordings without using human-annotated data. The pipeline relies entirely on pre-trained neural networks and a newly developed method based on fuzzy string matching to link speech segments in recordings to their corresponding transcript speeches. The automated annotations are fully on par with a human-level benchmark, showing that the pipeline can be used to annotate recordings at scale. The annotation tool is used to compile a multimodal text-audio corpus of parliamentary speeches in the Folketing from 2000-2022, the hitherto largest collection of natural speech audio of political speeches. This dataset is used in each of the three substantive papers, which all use speaker-standardized speech-level average vocal pitch as the outcome.

PAPER B, co-authored with Frederik Hjorth from the University of Copenhagen, investigates whether legislators, in their roles as partisans, change the sounds of their voices as they face heightened partisan conflict. We show that politicians speak with a higher than average vocal pitch when targeting parties or politicians from parties with whom they ideologically differ or disagree on legislation. This indicates that politicians change the sounds of their voices when facing partisan conflict to signal disagreement and agitation. The article adds to the existing literature on partisan conflict by showing that polarization and bill-level conflict are also conveyed by the sounds of politicians' voices and not only by words. This expands our thinking about how politicians communicate conflict to citizens and carries crucial lessons for how affective polarization should and can be studied and how we characterize the amount of conflict between politicians.

PAPER C investigates whether legislators, in their role as descriptive representatives of social groups, change the sounds of their voices to show their emotional engagement to representing their group members' interests and issues. I find that while descriptive representatives speak less frequently about their groups' policy issues throughout their political careers, they continue to be as emotionally engaged when they do talk about the issues. This suggests that legislators, who are descriptive representatives, change the sounds of their voices to signal their engagement in representing their groups. The article adds to the existing literature on the link between descriptive and substantive representation and more generally studies on representational linkages by showing that it matters deeply whether we understand substantive representation as issue attention, issue engagement, or a combination. This expands our knowledge about the mechanisms of the descriptive-to-substantive representative link and suggests that more effort should be devoted to defining when descriptive representation holds representative value.

PAPER D investigates whether legislators, when assuming a governing role, change the sounds of their voices to signal power and leadership abilities. I find compelling evidence consistent with this expectation, as legislators reduce their vocal pitch substantially when they become ministers, also independently of the topics they address in their speeches. This indicates that politicians change the sounds of their voices when they wear political power to show their composure, signaling their competencies and dominance. This expands our knowledge on the electoral and rhetorical costs and rewards that come with governing power and furthers our understanding of how politicians communicate power to citizens.

Dansk Resumé

Lydene fra den menneskelige stemme formidler betydning og information, der rækker ud over de ord, der bliver sagt. Det gælder i alle aspekter af menneskelig og social interaktion, og lige så for politik. Politikeres stemmeføring sender signaler om deres prioriteter og motiver uafhængigt af, hvad de siger. Givet vigtigheden af tale og debat i politik er det ikke overraskende, at politologer har analyseret politiske taler inden for næsten alle underdiscipliner af statskundskaben, fra lovgivningsstudier til internationale relationer, for at forstå, forklare og måle politikeres adfærd. Alligevel har kvantitative studier næsten udelukkende været baseret på transskriptioner, dvs. tekst, selv når de tilhørende optagelser, dvs. lyd, er tilgængelige. At fokusere alene på teksten og dermed hvad politikere siger, resulterer uundgåeligt i, at vigtige oplysninger om politik og eliternes adfærd bortfalder.

Med udgangspunkt i at studere lyden af politik som et outcome som den manifesterer sig i parlamentariske taler i Folketinget, bygger denne afhandling på tanken om, at *politik er svært at forstå i stilhed, men drager fordel af, at vi lytter til politikere, når de taler*. I det omfang at lydene af politikernes stemmer kommunikerer værdifuld information ud over ordene i sig selv, giver det os mulighed for at fortolke, forstå og forklare politiske eliters adfærd på nye måder og fra nye vinkler. Jeg præsenterer overbevisende beviser for, at politik har et lydlandskab, der kommunikeres via måden politikere ændrer lyden af stemmer i en akustisk forstand. Lyden af politik anal-

yseres, som den kommer til udtryk i variationer i politikeres standardiserede pitch langs tre centrale dimensioner af liberale demokratier, og anvender en automatiseret annotationspipeline, der er udviklet i denne afhandling, til at opbygge det multimodale datasæt, der anvendes i hver analyse. Afhandlingen består af fire selvstændige artikler (tre af mig som enkeltforfatter, én med en medforfatter).

PAPER A er en metodologisk artikel, hvor jeg udvikler en automatiseret annotationspipeline, der gør det muligt at annotere lydoptagelser af politiske taler i stor skala uden brug af tidligere menneskeskabte annotationsdata. Pipelinens metode bygger udelukkende på forudtrænede neurale netværk og en ny metode baseret på fuzzy string matching, som kobler talesegmenter fra optagelser til deres tilsvarende transskriptioner. De automatiserede annotationer er på niveau med menneskabte manuelle annoteringerog demonstrerer, at pipelinen kan anvendes til annotere store mængder lydoptagelser automatisk. Værktøjet bruges til at opbygge et multimodalt tekst-lyd-korpus af folketingsdebatter fra 2000-2022, den hidtil største samling af lydoptagelser af politiske taler brugt til at analysere politikeres adfærd. Datasættet anvendes i hver af de tre substantielle artikler, hvor pitch, standardiseret for hver politiker, bruges som den afhængige variabel.

PAPER B, som jeg har skrevet sammen med Frederik Hjorth fra Københavns Universitet, undersøger, om politikere i deres rolle som partipolitiske aktører ændrer lyden af deres stemmer, når de står over for øget partipolitisk konflikt, dvs. når partier er mere uenige. Vi viser, at politikere taler med en højere end gennemsnitlig pitch, når de adresserer partier eller politikere fra partier, de er ideologisk uenige med, eller når de er uenige om specifik lovgivning. Dette indikerer, at politikere ændrer lyden af deres stemmer, når de står over for partipolitisk konflikt, for at signalere deres uenigheder. Artiklen bidrager til den eksisterende forskning om partipolitisk konflikt ved at vise, at polarisering og konflikt på lovgivningsniveau også formidles gennem lydene af politikernes stemmer og ikke kun gennem ord. Dette udvider vores forståelse af, hvordan politikere kommunikerer konflikt til borgerne, og giver vigtige indsigter i, hvordan affektiv polarisering bør og kan studeres, samt hvordan vi karakteriserer mængden af konflikt mellem politikere.

PAPER C undersøger, om politikere i deres rolle som deskriptive repræsentanter for sociale grupper ændrer lyden af deres stemmer for at vise deres emotionelle engagement i at repræsentere deresgruppemedlemmers interesser og emner. Jeg finder, at selvom deskriptive repræsentanter taler mindre og mindre om deres gruppes politiske emner igennem deres politikere karriere, bliver de med at tale med lige så stor emotional engagement som i starten af deres karriere, når de taler om emnerne. Dette tyder på, at politikere, der er deskriptive repræsentanter, ændrer lyden af deres stemmer for at signalere deres engagement til at repræsentere deres gruppe. Artiklen bidrager til den eksisterende forskning om sammenhængen mellem deskriptiv og substantiv repræsentation og mere generelt studier om repræsentationsforbindelser ved at vise, at det er afgørende, om vi forstår substantiv repræsentation som emneopmærksomhed, emneengagement eller en kombination. Dette udvider vores viden om mekanismerne bag den deskriptive-til-substantive repræsentationsforbindelse og antyder, at der bør lægges større vægt på at definere, hvornår deskriptiv repræsentation har repræsentativ værdi.

PAPER D undersøger, om lovgivere, når de påtager sig en regeringsrolle, ændrer lyden af deres stemmer for at signalere deres magt og lederegenskaber. Jeg finder overbevisende beviser, der understøtter denne forventning, da lovgivere reducerer deres pitch markant, når de indtræder i en ministerpost, også uafhængigt af de emner, de taler om i deres taler. Dette indikerer, at politikere ændrer lyden af deres stemmer, når de har politisk magt, for at signalere kompetencer og dominans. Dette udvider vores viden om de valgmæssige og retoriske fordele og ulemper, der følger med regeringsmagten, og styrker vores forståelse af, hvordan politikere kommunikerer magt til borgerne.

Bibliography

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, *95*(3), 529–546.

Alim, H. S., & Smitherman, G. (2012). *Articulate while black: Barack obama, language, and race in the us*. Oxford University Press.

Amira, K., Cooper, C. A., Knotts, H. G., & Wofford, C. (2018). The southern accent as a heuristic in American campaigns and elections. *American Politics Research*, *46*(6), 1065–1093.

Anderson, R. C., & Klofstad, C. A. (2012). Preference for leaders with masculine voices holds in the case of feminine leadership roles. *PloS one*, *7*(12), e51216.

Andeweg, R. B. (2014a, May). 532cabinet ministers: Leaders, team players, followers? In R. A. W. Rhodes & P. Hart (Eds.), *The oxford handbook of political leadership*. Oxford University Press.

Andeweg, R. B. (2014b, June). Roles in legislatures. In S. Martin, T. Saalfeld, & K. W. Strøm (Eds.), *The oxford handbook of legislative studies* (pp. 267–285). Oxford University Press.

Andeweg, R. B., & Thomassen, J. (2011). Pathways to party unity: Sanctions, loyalty, homogeneity and division of labour in the dutch parliament. *Party Politics*, *17*(5), 655–672.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Aristotle. (350 B.C.E.). On the soul, translated by j. a. smith [Retrieved: 2024-11-22]. https://classics.mit.edu/Aristotle/soul.html

Arnold, C., & Küpfer, A. (2024). How alignment helps make the most of multimodal data.

Ash, E., Mikosch, H., Perakis, A., & Sarferaz, S. (2024). *Seeing and hearing is believing: The role of audiovisual communication in shaping inflation expectations* (tech. rep.). KOF Working Papers.

Ash, K., Johnson, W., Lagadinos, G., Simon, S., Thomas, J., Wright, E., & Gainous, J. (2020). Southern Accents and Partisan Stereotypes: Evaluating Political Candidates. *Social Science Quarterly*, *101*(5), 1951–1968.

Aung, T., & Puts, D. (2020). Voice pitch: A window into the communication of social power. *Current opinion in psychology*, *33*, 154–161.

Austen-Smith, D. (1990). Information transmission in debate. *American Journal of political science*, 124–152.

Bachorowski, J.-A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological science*, *6*(4), 219–224.

Bäck, H., & Debus, M. (2024). Legislatures and legislative behaviour. In A. Vatter & R. Freiburghaus (Eds.), *Handbook of comparative political institutions* (pp. 248–262). Edward Elgar Publishing.

Bäck, H., Debus, M., & Fernandes, J. M. (2021). *The politics of legislative debates*. Oxford University Press.

Bailer, S., Breunig, C., Giger, N., & Wüst, A. M. (2022). The diminishing value of representing the disadvantaged: Between group representation and individual career paths. *British Journal of Political Science*, *52*(2), 535–552.

Bale, T. (2003). Cinderella and her ugly sisters: the mainstream and extreme right in Europe's bipolarising party systems. *West European Politics*. https://doi.org/10.1080/01402380312331280598

Banai, B., Laustsen, L., Banai, I. P., & Bovan, K. (2018). Presidential, but not prime minister, candidates with lower pitched voices stand a better chance of winning the election in conservative countries. *Evolutionary Psychology*, *16*(2), 1474704918758736.

Banai, I. P., Banai, B., & Bovan, K. (2017). Vocal characteristics of presidential candidates can predict the outcome of actual elections. *Evolution and Human Behavior*, *38*(3), 309–314.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*(3), 614.

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech communication*, *46*(3-4), 252–267.

Barari, S., & Simko, T. (2023). Localview, a database of public meetings for the study of local politics and policymaking in the united states. *Scientific Data*, *10*(1), 135.

Blomgren, M., & Rozenberg, O. (2015). Legislative roles and legislative studies: The neo-institutionalist turning point? In *Parliamentary roles in modern legislatures* (pp. 8–36). Routledge.

Boersma, P., et al. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the institute of phonetic sciences*, *17*(1193), 97–110.

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, *82*(1), 55–59.

Bratton, K. A., & Ray, L. P. (2002). Descriptive representation, policy outcomes, and municipal day-care coverage in norway. *American Journal of Political Science*, 428–437.

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. *Proc. INTERSPEECH 2023*.

Bredin, H., & Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. *Proc. Interspeech 2021*.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). pyannote.audio: neural building blocks for speaker diarization. *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Broockman, D. E. (2013). Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives. *American Journal of Political Science*, *57*(3), 521–536.

Burgoon, J., Buller, D., & Woodall, W. (1996). *Nonverbal communication: The unspoken dialogue*. McGraw-Hill. https://books.google.dk/books?id=pG-xQgAACAAJ

Caldeira, G. A., Clark, J. A., & Patterson, S. C. (1993). Political respect in the legislature. *Legislative Studies Quarterly*, 3–28.

Caldeira, G. A., & Patterson, S. C. (1987). Political friendship in the legislature. *The Journal of Politics*, *49*(4), 953–975.

Camastra, F., & Vinciarelli, A. (2015). *Machine learning for audio, image and video analysis: theory and applications*. Springer.

Carnes, N. (2012). Does the numerical underrepresentation of the working class in Congress matter? *Legislative studies quarterly*, *37*(1), 5–34.

Carnes, N., & Lupu, N. (2015). Rethinking the comparative perspective on class and representation: Evidence from Latin America. *American Journal of Political Science*, *59*(1), 1–18.

Carney, D. R., Hall, J. A., & LeBeau, L. S. (2005). Beliefs about the nonverbal expression of social power. *Journal of nonverbal behavior*, *29*, 105–123.

Carroll, S. J., et al. (1994). *Women as candidates in American politics*. Indiana University Press.

Celis, K., Childs, S., Kantola, J., & Krook, M. L. (2008). Rethinking women's substantive representation. *Representation*, *44*(2), 99–110.

Celis, K., & Wauters, B. (2013). Pinning the butterfly: Women, blue-collar and ethnic minority mps vis-à-vis parliamentary norms and the parliamentary role of the group representative. In S. M. Rai (Ed.), *Ceremony and ritual in parliament* (1st, pp. 97–110). Routledge.

Cheng, J. T., Tracy, J. L., Ho, S., & Henrich, J. (2016). Listen, follow me: Dynamic vocal signals of dominance predict emergent social rank in humans. *Journal of Experimental Psychology: General*, *145*(5), 536.

Childs, S., & Krook, M. L. (2009). Analysing women's substantive representation: From critical mass to critical actors. *Government and opposition*, *44*(2), 125–145.

Christiansen, F. J. (2021). The polarization of legislative party votes: Comparative illustrations from denmark and portugal. *Parliamentary Affairs*, *74*(3), 741–759.

Cochrane, C., Rheault, L., Godbout, J.-F., Whyte, T., Wong, M. W.-C., & Borwein, S. (2022). The automatic analysis of emotion in political speech based on transcripts. *Political Communication*, *39*(1), 98–121.

Collins, S. A. (2000). Men's voices and women's choices. *Animal behaviour*, *60*(6), 773–780.

Coria, J. M., Bredin, H., Ghannay, S., & Rosset, S. (2020). A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification. In L. Espinosa-Anke, C. Martín-Vide, & I. Spasić (Eds.), *Statistical language and speech processing* (pp. 137–148). Springer International Publishing.

Cuzán, A. G. (2015). Five laws of politics. *PS: Political Science & Politics*, *48*(3), 415–419.

Dalton, R. J. (2008). The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences. *Comparative Political Studies*, *41*(7), 899–920.

Dalton, R. J. (2017). Party representation across multiple issue dimensions. *Party Politics*, *23*(6), 609–622.

Damann, T. J., Knox, D., & Lucas, C. (2024). A framework for studying causal effects of speech style: Application to u.s. presidential campaigns. https://christopherlucas.org/files/PDFs/more_than_words.pdf

Darwin, C. (1872). *The expression of the emotions in man and animals*. University of Chicago press.

Dassonneville, R., Fréchet, N., & Liang, B. (2022). Linguistic cleavages in public opinion. In C. D. Anderson & M. Turgeon (Eds.), *Comparative public opinion* (1st). Routledge.

Davidson, R. H. (1969). *The role of the congressman*. New York: Pegasus.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189.

Deutsch, M. (1973). *The resolution of conflict: Constructive and destructive processes*. Yale University Press.

Dietrich, B. J. (2021). Using motion detection to measure social polarization in the us house of representatives. *Political Analysis*, *29*(2), 250–259.

Dietrich, B. J., Enos, R. D., & Sen, M. (2019). Emotional arousal predicts voting on the us supreme court. *Political Analysis*, *27*(2), 237–243.

Dietrich, B. J., Hayes, M., & O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *American Political Science Review*, *113*(4), 941–962.

Druckman, J. N., Peterson, E., & Slothuus, R. (2013). How elite partisan polarization affects public opinion formation. *American Political Science Review*, *107*(1), 57–79.

Ekman, P., O'Sullivan, M., Friesen, W. V., & Scherer, K. R. (1991). Invited article: Face, voice, and body in detecting deceit. *Journal of nonverbal behavior*, *15*(2), 125–135.

Feinberg, D. R., Jones, B. C., & Armstrong, M. M. (2018). Sensory exploitation, sexual dimorphism, and human voice pitch. *Trends in ecology & evolution*, *33*(12), 901–903.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal behaviour*, *69*(3), 561–568.

Fenno, R. F. (1978). Home style: House members in their districts. *Little & Brown*.

Fernandes, J. M., Debus, M., & Bäck, H. (2021). Unpacking the politics of legislative debates. *European Journal of Political Research*.

Fish, K., Rothermich, K., & Pell, M. D. (2017). The sound of (in) sincerity. *Journal of Pragmatics*, *121*, 147–161.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading: Addison-Wesley.

Fiske, S. T., & Berdahl, J. (2007). Social power. *Social psychology: Handbook of basic principles*, *2*, 678–692.

Gailmard, S. (2014, May). 90accountability and principal–agent theory. In *The oxford handbook of public accountability*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0016

Galeotti, F., & Zizzo, D. J. (2018). Identifying voter preferences: The trade-off between honesty and competence. *European Economic Review*, *105*, 27–50.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.

Gennaro, G., & Ash, E. (2022). Emotion and reason in political language. *The Economic Journal*, *132*(643), 1037–1059.

Gennaro, G., & Ash, E. (2023). Televised debates and emotional appeals in politics: Evidence from c-span. *Center for Law & Economics Working Paper Series*, *2023*(01).

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, *87*(4), 1307–1340.

Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, *102*(1), 33–48.

Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to audio analysis: A matlab®*. Academic Press.

Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday.

Green-Pedersen, C. (2010). Bringing parties into parliament: The development of parliamentary activities in Western Europe. *Party Politics*, *16*(3), 347–369.

Green-Pedersen, C., & Thomsen, L. H. (2005). Bloc politics vs. broad cooperation? the functioning of danish minority parliamentarism. *The Journal of Legislative Studies*, *11*(2), 153–169.

Gregory, S. W. (1994). Sounds of power and deference: Acoustic analysis of macro social constraints on micro interaction. *Sociological Perspectives*, *37*(4), 497–526.

Gregory Jr, S. W., & Gallagher, T. J. (2002). Spectral analysis of candidates' nonverbal vocal communication: Predicting US presidential election outcomes. *Social Psychology Quarterly*, 298–308.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267–297.

Grose, C. R. (2005). Disentangling constituency and legislator effects in legislative representation: Black legislators or black districts? *Social Science Quarterly*, *86*(2), 427–443.

Guyer, J. J., Briñol, P., Vaughan-Johnston, T. I., Fabrigar, L. R., Moreno, L., & Petty, R. E. (2021). Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of Nonverbal Behavior*, 1–26.

Hammond, J. K., & White, P. R. (1996). The analysis of nonstationary signals using time-frequency methods. *Journal of Sound and vibration*, *190*(3), 419–447.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behavior*, *22*(3), 165–196.

Hill, K. Q., & Hurley, P. A. (2002). Symbolic speeches in the US Senate and their representational implications. *Journal of Politics*, *64*(1), 219–231.

Hjorth, F. (2024). Losing touch: The rhetorical cost of governing.

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, *63*(2), 467–490.

Jensen, H. (2002). *Partigrupperne i folketinget*. Jurist-og Økonomforbundets Forlag Copenhagen.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. *Handbook of emotions*, *2*, 220–235.

Kalkhoff, W., Thye, S. R., & Gregory Jr, S. W. (2017). Nonverbal vocal adaptation and audience perceptions of dominance and prestige. *Social Psychology Quarterly*, *80*(4), 342–354.

Kanngieser, A. (2012). A sonic geography of voice: Towards an affective politics. *Progress in Human Geography*, *36*(3), 336–353.

Kappos, C. (2024). *The way eu make me feel: Measuring anxiety in the brexit negotiations using text and audio* [Doctoral dissertation, University of California, Los Angeles].

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (pp. 192–238, Vol. 15). University of Nebraska Press.

Kirkham, S., & Moore, E. (2016). Constructing social meaning in political discourse: Phonetic variation and verb processes in ed miliband's speeches. *Language in Society*, *45*(1), 87–111.

Klint, T., Evert, A. S., Kjær, U., Pedersen, M. N., & Hjorth, F. (2023). The danish legislators database. *Electoral Studies*, *84*, 102624.

Klofstad, C. A. (2016). Candidate voice pitch influences election outcomes. *Political psychology*, *37*(5), 725–738.

Klofstad, C. A., & Anderson, R. C. (2018). Voice pitch predicts electability, but does not signal leadership ability. *Evolution and human behavior*, *39*(3), 349–354.

Klofstad, C. A., Anderson, R. C., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PloS one, 10*(8), e0133779.

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences, 279*(1738), 2698–2704.

Klofstad, C. A., Nowicki, S., & Anderson, R. C. (2016). How voice pitch influences our choice of leaders: when candidates speak, their vocal characteristics–as well as their words–influence voters' attitudes toward them. *American Scientist, 104*(5), 282–288.

Knox, D., & Lucas, C. (2021). A dynamic model of speech for the social sciences. *American Political Science Review, 115*(2), 649–666.

Kosiara-Pedersen, K., & Kurrild-Klitgaard, P. (2018). Change and stability in the Danish party system. In *Party system change, the european crisis and the state of democracy* (pp. 63–79). Routledge.

Kraut, R. E. (1978). Verbal and nonverbal cues in the perception of lying. *Journal of personality and social psychology, 36*(4), 380.

Lauridsen, G. A., Dalsgaard, J. A., & Svendsen, L. K. B. (2019). SENTIDA: A new tool for sentiment analysis in Danish. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift, 4*(1), 38–53.

Laustsen, L., & Petersen, M. B. (2015). Does a competent leader make a good friend? conflict, ideology and the psychologies of friendship and followership. *Evolution and Human Behavior, 36*(4), 286–293.

Laustsen, L., Petersen, M. B., & Klofstad, C. A. (2015). Vote choice, ideology, and social dominance orientation influence preferences for lower pitched voices in political candidates. *Evolutionary Psychology, 13*(3), 1474704915600576.

Laver, M. (2021). Analysing the politics of legislative debate. In H. Bäck, M. Debus, & J. M. Fernandes (Eds.), *The politics of legislative debates*. Oxford University Press.

Leinonen, L., Hiltunen, T., Linnankoski, I., & Laakso, M.-L. (1997). Expression of emotional–motivational connotations with a one-word utterance. *The Journal of the Acoustical society of America*, *102*(3), 1853–1863.

Levendusky, M. S. (2010). Clearer cues, more consistent voters: A benefit of elite polarization. *Political Behavior*, *32*, 111–131.

Lind, A. V., Hallsson, B. G., & Morton, T. A. (2023). Polarization within consensus? an audience segmentation model of politically dependent climate attitudes in denmark. *Journal of Environmental Psychology*, *89*, 102054.

Lipset, S. M., Rokkan, S., et al. (1967). *Cleavage structures, party systems, and voter alignments: An introduction* (Vol. 2). Free Press New York.

Louwerse, T., Sieberer, U., Tuttnauer, O., & Andeweg, R. B. (2021). Opposition in times of crisis: Covid-19 in parliamentary debates. *West European Politics*, *44*(5-6), 1025–1051.

Mansbridge, J. (1999). Should blacks represent blacks and women represent women? A contingent" yes". *The Journal of politics*, *61*(3), 628–657.

Mansbridge, J. (2003). Rethinking representation. *American political science review*, *97*(4), 515–528.

Mansbridge, J. (2011). Clarifying the concept of representation. *American political science review*, *105*(3), 621–630.

March, J. G., & Olsen, J. P. (1989). *Rediscovering institutions: The organizational basis of politics.* New York: Free Press.

Martin, S., Saalfeld, T., & Strøm, K. W. (2014, June). 1introduction. In S. Martin, T. Saalfeld, & K. W. Strøm (Eds.), *The oxford handbook of legislative studies* (pp. 1–26). Oxford University Press.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and emotion*, *23*(2), 209–237.

Mayhew, D. R. (1974). *Congress: The electoral connection*. Yale university press.

Moez, C. (2024). *Political disaffection and the decline of the centre: Quantitative text analysis approaches* [Doctoral dissertation, University of Toronto (Canada)].

Mollin, S. (2018). The use of face-threatening acts in the construction of in-and out-group identities in british parliamentary debates. In B. Bös, S. Kleinke, S. Mollin, & N. Hernández (Eds.), *The discursive construction of identities on-and offline: Personal-group-collective* (pp. 205–226). John Benjamins Publishing Company.

Moore, C. (2013). *Margaret Thatcher: From Grantham to the Falklands*. Vintage.

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855–869.

Neumann, M. (2019). Hooked with phonetics: The strategic use of style-shifting in political rhetoric. *Annual Meeting of the American Political Science Association. Washington, DC*.

Niebuhr, O., Tegtmeier, S., & Brem, A. (2017). Advancing research and practice in entrepreneurship through speech analysis: From descriptive rhetorical terms to phonetically informed acoustic charisma metrics. *Journal of Speech Sciences*, *6*(1), 3–26.

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Novák-Tót, E., Niebuhr, O., & Chen, A. (2017). A gender bias in the acoustic-melodic features of charismatic speech? *Proceedings of the International Conference on Spoken Language Processing*, 2248–2252.

O'Connor, J. J., & Barclay, P. (2017). The influence of voice pitch on perceptions of trustworthiness across social contexts. *Evolution and human behavior*, *38*(4), 506–512.

O'Grady, T. (2019). Careerists versus coal-miners: Welfare reforms and the substantive representation of social groups in the British Labour party. *Comparative Political Studies*, *52*(4), 544–578.

Oguchi, T., & Kikuchi, H. (1997). Voice and interpersonal attraction. *Japanese Psychological Research*, *39*(1), 56–61.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of f of voice. *Phonetica*, *41*(1), 1–16.

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic bulletin & review*, *24*(3), 856–862.

Oppenheim, A. V. (1999). *Discrete-time signal processing*. Pearson Education India.

O'Shaughnessy, D. (1987). *Speech communication: Human and machine*. Addison-Wesley Publishing Company. https://books.google.dk/books?id=mHFQAAAAMAAJ

Owren, M. J., & Bachorowski, J.-A. (2007). Measuring emotion-related vocal acoustics. *Handbook of emotion elicitation and assessment*, 239–266.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317.

Pearson, K., & Dancey, L. (2011). Speaking for the underrepresented in the house of representatives: Voicing women's interests in a partisan era. *Politics & Gender*, *7*(4), 493–519.

Peterson, A., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, *26*(1), 120–128.

Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, 825–850.

Phillips, A. (1998). *The politics of presence*. OUP Oxford.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: a window into the origins of human vocal control? *Trends in cognitive sciences*, *20*(4), 304–318.

Pisanski, K., Mora, E. C., Pisanski, A., Reby, D., Sorokowski, P., Frackowiak, T., & Feinberg, D. R. (2016). Volitional exaggeration of body size through fundamental and formant frequency modulation in humans. *Scientific reports*, *6*(1), 1–8.

Pisanski, K., Oleszkiewicz, A., Plachetka, J., Gmiterek, M., & Reby, D. (2018). Voice pitch modulation in human mate choice. *Proceedings of the Royal Society B*, *285*(1893), 20181634.

Pitkin, H. F. (1967). *The concept of representation*. University of California Press.

Podesva, R. J., Reynolds, J., Callier, P., & Baptiste, J. (2015). Constraints on the social meaning of released/t: A production and perception study of US politicians. *Language Variation and Change*, *27*(1), 59–87.

Preuhs, R. R. (2006). The conditional effects of minority descriptive representation: Black legislators and policy influence in the american states. *The Journal of Politics*, *68*(3), 585–599.

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, *44*(1), 97–131.

Proksch, S.-O., & Slapin, J. B. (2012). Institutional foundations of legislative speech. *American Journal of Political Science*, *56*(3), 520–537.

Proksch, S.-O., Wratil, C., & Wäckerle, J. (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, *27*(3), 339–359.

Przeworski, A., Stokes, S. C., & Manin, B. (1999). *Democracy, accountability, and representation* (Vol. 2). Cambridge University Press.

Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and human behavior*, *27*(4), 283–296.

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 340–344.

Rabiner, L. R., & Schafer, R. W. (2011). *Theory and applications of digital speech processing*. Pearson.

Rauh, C., & Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.

Rheault, L., & Borwein, S. (2019). *Multimodal techniques for the study of affect in political videos* (tech. rep.). Working Paper.

Rheault, L., & Borwein, S. (2022). *Audio as Data*. Edward Elgar Publishing.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112–133.

Ricks, J. I. (2020). The effect of language on political appeal: Results from a survey experiment in thailand. *Political Behavior*, *42*(1), 83–104.

Rittmann, O. (2024). Legislators' emotional engagement with women's issues: Gendered patterns of vocal pitch in the german bundestag. *British Journal of Political Science*, *54*(3), 937–945.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, *58*(4), 1064–1082.

Rocca, M. S., & Sanchez, G. R. (2008). The effect of race and ethnicity on bill sponsorship and cosponsorship in congress. *American Politics Research*, *36*(1), 130–152.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of personality and social psychology*, *76*(5), 805.

Saalfeld, T. (2014). Parliamentary questions as instruments of substantive representation: Visible minorities in the uk house of commons, 2005–10. In *The roles and function of parliamentary questions* (pp. 13–31). Routledge.

Saalfeld, T., & Bischof, D. (2013). Minority-ethnic mps and the substantive representation of minority interests in the house of commons, 2005–2011. *Parliamentary Affairs*, *66*(2), 305–328.

Saalfeld, T., & Strøm, K. W. (2014, June). Political parties and legislators. In S. Martin, T. Saalfeld, & K. W. Strøm (Eds.), *The oxford handbook of legislative studies* (pp. 371–398). Oxford University Press.

Schattschneider, E. (1960). *The semisovereign people*. Holt, Rinehart; Winston.

Scherer, K. R. (1993). Interpersonal expectations, social influence, and emotion transfer. In P. D. Blanck (Ed.), *Interpersonal expectations: Theory, research and applications* (pp. 316–334). Cambridge University Press.

Scherer, K. R. (2018). Acoustic patterning of emotion vocalizations. In S. Frühholz & P. Belin (Eds.), *The oxford handbook of voice perception* (pp. 61–91). Oxford University Press.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). *Vocal expression of emotion.* (Vol. 40). Oxford University Press.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. (1984). Vocal cues to speaker affect: Testing two models. *The Journal of the Acoustical Society of America*, *76*(5), 1346–1356.

Schild, C., Stern, J., & Zettler, I. (2020). Linking men's voice pitch to actual and perceived trustworthiness across domains. *Behavioral Ecology*, *31*(1), 164–175.

Schirmer, A., Chiu, M. H., Lo, C., Feng, Y.-J., & Penney, T. B. (2020). Angry, old, male–and trustworthy? How expressive and person voice characteristics shape listener trust. *Plos one*, *15*(5), e0232431.

Schwarz, D., Traber, D., & Benoit, K. (2017). Estimating intraparty preferences: comparing speeches to votes. *Political Science Research and Methods*, *5*(2), 379–396.

Searing, D. D. (1994). *Westminster's world: understanding political roles*. Harvard University Press.

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, *210*(4471), 801–803.

Signorello, R. (2019). Voice in Charismatic Leadership. In *The oxford handbook of voice studies*.

Signorello, R. (2021). The Vocal Attractiveness of Charismatic Leaders. In *Voice attractiveness* (pp. 41–54). Springer.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and linguistics compass*, *3*(2), 621–640.

Sneddon, I. N. (1995). *Fourier transforms*. Courier Corporation.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5329–5333.

Snyder Jr, J. M., & Groseclose, T. (2000). Estimating party influence in congressional roll-call voting. *American Journal of Political Science*, 193–211.

Søyland, M., & Høyland, B. (2021). Norway: Committee-membership matters, party loyalty decides. In H. Bäck, M. Debus, & J. M. Fernandes (Eds.), *The politics of legislative debates* (pp. 633–650). Oxford University Press.

Stæhr Harder, M. M. (2022). Nordatlantiske mandater i folketinget. *Politica: Tidsskrift for Politisk Videnskab*, *54*(1).

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, *8*(3), 185–190.

Strøm, K. (1997). Rules, reasons and routines: Legislative roles in parliamentary democracies. *The Journal of Legislative Studies*, *3*(1), 155–174.

Surawski, M. K., & Ossoff, E. P. (2006). The effects of physical and vocal attractiveness on impression formation of politicians. *Current Psychology*, *25*(1), 15–27.

Tarr, A., Hwang, J., & Imai, K. (2023). Automated coding of political campaign advertisement videos: An empirical validation study. *Political Analysis*, *31*(4), 554–574.

Thomassen, J. J. (1994). Empirical research into political representation: Failing democracy or failing models? In *Elections at home and abroad: Essays in honor of warren miller* (pp. 237–265). University of Michigan.

Tigue, C. C., Borak, D. J., O'Connor, J. J., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, *33*(3), 210–216.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707.

Titze, I. R. (1994). *Principles of voice production*. Prentice Hall. https://books.google.dk/books?id=m48JAQAAMAAJ

Touati, P. (1993). Prosodic aspects of political rhetoric. *ESCA workshop on prosody*.

Tumminia, J., Kuznecov, A., Tsilerides, S., Weinstein, I., McFee, B., Picheny, M., & Kaufman, A. R. (2021). Diarization of Legal Proceedings. Identifying and Transcribing Judicial Speech from Recorded Court Audio. *arXiv preprint arXiv:2104.01304*.

Vainio, M., Suni, A., Šimko, J., & Kakouros, S. (2023). The power of prosody and prosody of power: An acoustic analysis of finnish parliamentary speech. *arXiv preprint arXiv:2305.16040*.

Wahlke, J. C. (1962). Theory: A framework for analysis. In J. C. Wahlke, H. Eulau, W. Buchanan, & L. C. Fergusson (Eds.), *The legislative system: Explorations in legislative behaviour* (pp. 3–28). New York: John Wiley; Sons.

Wahlke, J. C., Eulau, H., Buchanan, W., & Fergusson, L. C. (1962). *The legislative system: Explorations in legislative behaviour*. New York: John Wiley; Sons.

Walton, J. H., & Orlikoff, R. F. (1994). Speaker race identification from acoustic cues in the vocal signal. *Journal* *of Speech, Language, and Hearing Research*, *37*(4), 738–745.

Wängnerud, L. (2009). Women in parliaments: Descriptive and substantive representation. *Annual Review of Political Science*, *12*, 51–69.

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*(1), 529–544.

Willumsen, D. M. (2021). Open-list pr and parliamentary speech. In H. Bäck, M. Debus, & J. M. Fernandes (Eds.), *The politics of legislative debates*. Oxford University Press.

Wilson, W. C. (2010). Descriptive representation and latino interest bill sponsorship in congress. *Social Science Quarterly*, *91*(4), 1043–1062.

Wüst, A. M. (2014). A lasting impact? on the legislative activities of immigrant-origin parliamentarians in germany. *The Journal of Legislative Studies*, *20*(4), 495–515.

Zárate, M. G., Quezada-Llanes, E., & Armenta, A. D. (2024). Se habla español: Spanish-language appeals and candidate evaluations in the united states. *American Political Science Review*, *118*(1), 363–379.

Ziblatt, D., Hilbig, H., & Bischof, D. (2024). Wealth of tongues: Why peripheral regions vote for the radical right in germany. *American Political Science Review*, *118*(3), 1480–1496.

Zuckerman, M., & Driver, R. E. (1989). What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of nonverbal behavior*, *13*(2), 67–82.

Zuckerman, M., & Miyake, K. (1993). The attractive voice: What makes it so? *Journal of nonverbal behavior*, *17*(2), 119–135.