Hooked With Phonetics: The Strategic Use of Style-Shifting in Political Rhetoric

Markus Neumann The Pennsylvania State University

The study of non-policy representation has emphasized written or verbal communication by representatives, neglecting the crucial non-verbal component of symbolic representation. I argue that, in order to convince their constituents that they are "like them" and will act in their interests, politicians project likeness through the way they talk. When speaking to an audience of ordinary citizens, a politician will play the "average Joe;" when addressing a congressional committee, they will demonstrate sophistication and competence. Hence, political speakers shift their style according to the audience. I test this hypothesis using audio data from congressional and campaign speeches of U.S. senators uploaded to YouTube. I extract the acoustic properties of the audio signal and measure vowel space density, a concept developed in phonetics, to categorize the degree of articulation in speech. The results suggest that politicians do adjust their articulation to fit the needs of their audience.

Introduction

The representative-constituent relationship lies at the heart of every representative democracy: Citizens vote for candidates, who then turn their constituents' political attitudes into policy. As a result, discourse, and even more so, quantitative scholarship, on representation is frequently limited to policy congruence and responsiveness: Does the policy that is being implemented match the preferences of the constituents? However, representation is much broader than this.

As noted by Pitkin (1967), representation is as much about 'standing for' as it is about 'acting for' a constituency. The former is harder to define, but nevertheless important and falls under the label of 'symbolic representation' (Eulau and Karps, 1977). As opposed to the transactional aspects of representation, symbolic representation is more about gestures than palpable actions. Their goal is not necessarily to leave constituents with the impression that a representative acted in their interests, but more so that she is a "good person". Symbolic representation occupies a critical role in some of the most important theories of political representation, be it in Congress (Mayhew, 1974) or at home (Fenno, 1978).

If the non-policy, non-transactional aspects of representation are so important, then why has scholarship in the period since Mayhew (1974), Eulau and Karps (1977) and Fenno (1978) paid relatively little attention to it? Symbolic representation is a hard problem to

study. As opposed to roll-call votes on policy, the nuances of social interactions between constituents and their representatives are difficult to conceptualize and measure. Grimmer (2010) – who attempts to further develop Fenno's concept of home style – offers a solution by analyzing Senate press releases with the help of text as data methods. However, in pursuit of the goal of obtaining a quantitative measure of home style, Grimmer takes some liberties with Fenno's theory.

One of the important distinctions of Fenno (1978) is that he pays close attention to both verbal and non-verbal forms of communication. The latter, he argues, represents a more honest indicator of the true intentions of legislators and constituents therefore pay special attention to it. By contrast, Grimmer (2010) focuses only on the verbal. He dismisses the non-verbal forms of communication studied by Fenno as "folksy mannerisms". However, I argue that these mannerisms are important and should not be overlooked. These acts of symbolic representation are aimed at sending the following message: "You can trust me because we are like one another" (Fenno, 1977).

In this research, I expand on this idea and develop a theory of likeness. In the terms of the representative-constituent relationship, this means: "I am like you, therefore you should vote for me." This is a departure from the classic concept of representation, where the message is more akin to: "I want the same things as you and therefore you should vote for me." Likeness is a more diffuse concept than descriptive representation, which exists when representatives match a specific demographic qualifier of their constituents – i.e. African Americans politicians representing African Americans, women representing women, farmers representing farmers, and so on (Mansbridge, 1999). By contrast, likeness is about a *perceived* match in attributes between constituent and representative. Whether the two are objectively alike is not important, what matters is that the bond *feels* true. And as noted by both Mayhew (1974) and Fenno (1978), perception is absolutely critical for representation.

So how does a politician project likeness with her constituents? The strategy I focus on here is phonetic style-shifting. I argue that representatives rely on subtle rhetorical clues in order to signal likeness with their constituents. By modulating the degree of articulation in their speech, they convince their voters that they are one of them. There are a number of rhetorical devices representatives can use to accomplish this. Regional dialects and accents are one approach. There are also a number of informal pronunciations, such as "y'all" or g-dropping – the practice of omitting the /g/ from words ending in /-ing/. Sociolinguists have studied this phenomenon extensively and also observed politicians making use of it in certain contexts (Liberman, 2008; Nunberg, 2008; Liberman, 2011). I expect the extent of style-shifting to be dependent on the audience in front of which a politician performs: When politicians want to demonstrate competence and appeal

to elites – as in, when they are speaking in Congress – they will use a high degree of articulation. By contrast, when the primary goal is to demonstrate empathy and appeal to the "average Joe" – as would be the case in a campaign setting – representatives will use a lower degree of articulation. In my eyes, phonetic style-shifting thus represents a more faithful translation of the non-verbal component of Fenno's concept of representation of self than Grimmer's approach.

I rely on the U.S. Senate to test my theory. The upper chamber of the American legislature provides the ideal test case because senators represent diverse constituencies and command the national spotlight to an extent that provides sufficient data. In analyzing the audio of political speeches, I make use of a medium political science has largely ignored so far. My data is scraped from the YouTube channels of senators, most of whom maintain two channels - one for campaigns, and one for legislative activity. This division already suggests that my theory points in the right direction, and provides me with a convenient way of assessing the differences in speech patterns in low- (campaigns) versus high-brow (Congress) settings. My sample spans 19 senators who stood for re-election in 2018, and contains 1049 videos, which run for a combined 20 hours.

In keeping with prior research in phonetics (Sandoval et al., 2013; Story and Bunton, 2017), I rely on vowel space area, which approximates the extent to which the throat and mouth are used in the generation of sounds, to operationalize the concept of articulation. I compare the vowel space area of senators in two different settings - campaigns and Congress - and show that senators consistently make greater use of the vocal apparatus in the latter case. This provides evidence in favor of my argument that a variation in articulation, contingent on the audience, allows politicians to perform acts of symbolic representation through the projection of likeness.

This research has important implications for the concept of representation. The study of non-policy representation, while not as common as policy representation, is well-established, and as such follows well-tread paths – either through the use of anecdotes, as with Fenno, or the examination of explicitly verbal forms of communication, as demonstrated by Grimmer. My research shows that the more subtle and nebulous aspects of non-policy responsiveness can nevertheless be studied with quantitative evidence.

Furthermore, the strategy of phonetic style-shifting can provide some insights into populism. Populism rails against the establishment, against the government and against the elite. To varying degrees, this is something all politicians engage in. As Fenno (1978) famously notes, "Members of Congress run for Congress by running against Congress." But populism contains an inherent contradiction, as populists themselves are elites, and at the very least aim to form the government. Consequently they are faced with the challenge of convincing their voters – who hate elites – that they are one of them, and

not part of the elite. My theory of likeness allows for politicians to perform both roles simultaneously and to shift between them fluidly by varying the degree of articulation in their speech.

Literature

The Perception of Likeness

Like Mayhew (1974), I assume that politicians are office-seeking.[1] Mayhew (1974) posits that in order to pursue the goal of getting reelected, politicians engage in three activities: (1) position-taking, (2) advertising and (3) credit-claiming. Ultimately, all of these strategies accomplish their goal by building a specific image of the politician in the eyes of his or her constituents. This image, reduced to its core, is: "I am like you and therefore you should vote for me." In a way, this is a perversion (or in the words of Fenno (1977), "corruption") of the classical view (APSA, 1950; Downs, 1957; Schattschneider, 1960) of the electoral connection, in which politicians (or their parties) essentially claim: "I want the same policies as you and therefore you should vote for me." But as Mayhew (1974) observes, this does not actually happen. It is possible for members of Congress to vote against the interests of their constituents and get away with it because keeping track of what Congress does at all times and evaluating how its decisions fit into one's preference system (to the extent that it even exists (Converse, 1964)) requires citizens to pay enormous information costs (see also Arnold (1990). Rather than supporting politicians who represent their policy interests, voters simply tend to adopt the policy positions of their representatives (Lenz, 2009; Broockman and Butler, 2017).

Politicians also exploit the fact that "I am like you" is a shortcut for "I want the same policies as you." As noted by one of the politicians studied by Fenno (1977), as long as voters believe that their representative is "a nice fella," they are willing to make "presumptions in my favor." As a result, the way politicians present themselves to citizens is enormously important. Mayhew (1974) and Fenno (1977) have laid the groundwork here, as they have analyzed the lengths that representatives go to in order to cultivate their image in the public eye – both when they spend their time in Washington, as well as in their districts. To that end, Fenno points to their presentation of self as being absolutely critical. Fenno (1977) adapts this concept from the the sociologist and social psychologist Erwin Goffman. Goffman (1959) points out that when it comes to communication, both verbal and non-verbal expressions are critical. In fact, between the two, Goffman attributes

[1] That does not mean they do not have any other goals (or even that my findings on rhetoric have no other consequences for them), but for the purpose of my core argument, this is the only necessary assumption.

greater importance to non-verbal communication. Fenno (1977) translates this idea into the political realm: He points out that constituents cannot fully trust the performance of their representative: He may claim to be acting in their best interest, but his verbal assurances are no guarantee. As a result, they look towards his non-verbal behavior as a more honest indicator for his true intentions.

Grimmer (2013) has rekindled interest in this line of research, added innovative quantitative evidence and further developed the concept of representational style. Grimmer notes that there are differences in the way that representatives perform their role, depending on both the characteristics of the constituency as well as the legislator herself. He focuses on how legislators connect their Washington style – the way they spend their time and resources in the capital – to their specific home style. Grimmer uses text as data methods to show that the representational style of a legislator is revealed through what she communicates to her constituents with the help of press releases. One caveat of this approach is that it assumes a monolithic constituency. Furthermore, it focuses only on the verbal. And while Fenno is very important to Grimmer, he also dismisses his version of "home style" as too focused on "folksy mannerisms". The reason for that is because to Grimmer, home style is about information, and folksy mannerisms don't help to inform constituents. However, to Fenno (1977), home style isn't really about information at all. The goal that representatives, in his eyes, are pursuing, is to build trust with constituents. To this end, they face a number of challenges: On the one hand, the need to demonstrate they are qualified to do the job. On the other hand, they need to identify and empathize with their constituents. Acts of symbolic representation are aimed at doing precisely that.

My own argument then, building on Grimmer, is that politicians have more than one representational style. They adjust their behavior according to the needs of the situation, and specifically the audience. What politicians say matters - but how they say it is also important. Representation and presentation of self through press releases is explicitly overt. Projecting likeness with constituents through subtle changes to the way spoken language is used is exactly the opposite. This strategy enables politicians to navigate the difficulties of having to satisfy a diverse set of constituents.

One Representative, Many Constituencies

There are at least four reasons to assume that senators have diverse constituencies. First, constituencies are not monolithic. Many of the politicians described by Fenno (1977) have both urban and rural constituencies, and even though they generally gear themselves towards one of them, they try not to neglect the other entirely. But as shown by Cramer (2016), rural and urban constituencies can be very different, and a strategy that garners success with one leads to resentment from the other. This problem of diverse constituencies

is even more pressing for senators, who represent entire states as opposed to individual districts.

Second, in order to get themselves elected, politicians have to serve more than one master. Interest groups play a substantial role in elections, whether it is through campaign finance (Desmarais, La Raja and Kowal, 2015) or the provision of information (Hall and Deardorff, 2006). A slew of research on winner-take-all politics (Hacker and Pierson, 2010) and elite domination of representation (Bartels, 2009; Gilens and Page, 2014; Achen and Bartels, 2016) has shown that economic elites exert a considerable degree of control over politicians. Acting like a 'country boy' may play well with the 'folks back home', but economic elites, who hold fundamentally different attitudes (Page, Bartels and Seawright, 2013), are likely to be less impressed by such demeanor. Politicians need to prove themselves as reliable partners to these elites, and to this end, perception is just as important as it is with voters.

Third, politicians are elites themselves. As shown by Butler (2014), representing other elites comes naturally to politicians, and doesn't even necessitate conservative economic preferences - the presence of shared experiences is enough. Most members of Congress will likely have an easier time explaining to their constituents how to make use of school vouchers or expedite a passport, rather than how to track down a missing social security check. This is compounded by the dynamics of Washington itself. Ultimately, "this town" revolves around green rooms, fundraisers and "$100 haircuts" (Leibovich, 2014). If politicians employed their "folksy mannerisms" at swanky Georgetown parties, they would be laughed out of town.

Fourth, when politicians act as trustees rather than delegates (and most politicians wear both hats to at least some extent (Miller and Stokes, 1963; Saward, 2014)), voters value competence over ideological congruence (Fox and Shotts, 2009). The trustee model assumes that politicians inherently know better because they are elites - so in this case, acting like their constituents would be counterproductive for representatives.

The Shape-Shifting Representative

So politicians have some incentives to act like elites, and some incentives to act like the 'average Joe'. Which one do they choose? According to Saward (2014), both. Relying heavily on Machiavelli, the author argues that it would be foolish for politicians to play the same role at all times. The greatest strength of his shape-shifting representative, whom he pits against the delegate and trustee models (but the argument works just as well for my purposes) is his flexibility. Rather than filling a single role, the representative of Saward (2014):

> "positions him or herself as a subject with respect to constituents, supporters, or *listeners*; in other words, they adopt subject positions." [emphasis added]

Although not as developed, this theory can already be found in Mayhew (1974), who notes that "Handling discrete audiences in person requires simple agility," and cites a Congressman who defends his strategy of giving different speeches to pro- and anti-war crowds as following: "My positions are not inconsistent; I just approach different people differently."

Saward (2014) follows a similar approach and goes on to note that:

> "by shaping strategically (or having shaped) his persona and policy positions for certain constituencies and audiences"

the representative can have his cake and eat it too. Translated to my case, this means that depending on the audience, a politician will play either the shrewd statesman, or the 'average Joe'. The word 'play' is carefully chosen here: Saward (2014) does in fact see representation as a performance - "it is performed in the theatrical sense [...] and in the speech-act sense":

> "the actor offers himself as a representative by virtue of (a) substantive policy positions, then (b) on the basis of likeness or similarity to constituents, then (c) in terms of the champion of particular interests"

Policy is Not a Good Signal for Shape-Shifters

If we accept that in order to get (re-)elected, politicians attempt to prove to their various constituencies that they are like them and will act in their interest – how do they do so? As pointed out by Box-Steffensmeier et al. (2003), for the concept of representation, the legislator "being like me" is at least as important to constituents as policy. After all, a roll-call vote – which is a crude indicator of legislator preferences to begin with – that is well-received in one constituency may offend another. Once a politician has voted in favor of something, they can't take it back - every roll-call vote becomes part of their record, and can therefore be cited against them by their opponents. As noted by Maestas (2003), legislators need to take great care "to ensure that they do not cast votes or take positions that might return to haunt them." The need to balance multiple constituencies can even go so far that legislators with more diverse constituency opinions abstain on contentious votes altogether, and thus avoid having to take a position (Cohen and Noll, 1991; Rothenberg and Sanders, 2000; Jones, 2003). The weaknesses of policy-based representation when it comes to serving the needs of multiple constituencies demonstrates the need to pay greater attention to non-policy representation.

Rhetoric and Plausible Deniability

By contrast, it is considerably easier to say one thing to one constituency, and something else to another. However, even then, there is a chance that an errant comment by a politician makes it onto the record that they end up regretting later. For example, a particularly unfortunate rhetorical mishap was Romney's infamous 47% statement, in which he scorned (nearly) half of the voters in the country, accusing them of being unable to take care of themselves. The statement was made at a fundraiser and obviously aimed at the wealthy donors present at the event, who may have found it to be perfectly measured and reasonable. But it was an obvious slap in the face of the Republican party's considerable blue-collar base (at the time further strengthened by the Tea Party) as well as undecided voters.

There is however another, even more subtle element of communication which allows speakers to empathize with their audience. Voice[2] is much easier to modulate than any of the other means of symbolic representation described above. When speaking to a crowd of potential voters at a campaign rally or town hall event, politicians can lower their degree of articulation by dropping the g in -ing words, not releasing /t/, or speaking with a regional dialect. A Southern drawl can be code for "I am one of you" in the same way a derogatory reference to 'welfare queens' is, and it does so in a much less risky manner. When speaking to a room filled with business executives, the same politician might rely on carefully placed pauses and flawless pronunciation instead, signaling his likeness with this, very different, audience. Either way, rhetoric, generally a form of symbolic representation, thus becomes implied descriptive representation. This strategy is not just a passing fancy, but has been employed by politicians for thousands of years: Cicero (1986), the master orator of ancient Rome comments (disapprovingly) on this practice in one of his writings, *De Oratore*. He warns the budding rhetorician against the use of overly sophisticated language when in the presence of the common man, as the speaker should not wish to appear "so very wise among fools" that "though they very much approve his understanding, and admire his wisdom, yet should feel uneasy that they themselves are but idiots to him."

To sum up, I argue that politicians project likeness to demonstrate to their voters that they are like them and can be trusted. This idea revolves around symbolic acts of representation as outlined by Mayhew (1974) and even more so, Fenno (1978). I adopt Fenno's argument on the 'presentation of self' and give precedence to his focus on nonverbal communication, over Grimmer (2013)'s information-centric interpretation of home style. Furthermore, I argue that since legislators have diverse constituencies,

[2]Note that throughout this paper, I use voice in the acoustic sense, as opposed to the ability of getting heard, employed by Schlozman, Verba and Brady (2012).

they need to be shape-shifting (Saward, 2014), changing their appeals conditional on the audience. Policy representation is difficult in this regard, as roll-call votes commit representatives to a set position (Maestas, 2003). By contrast, rhetoric, and even more specifically, phonetic style shifting enable representatives to be equivocal about their position and adjust their representational behavior as called for by the situation.

A Theory of Phonetic Style Shifting

The theoretical claim I advance in this paper is that politicians modulate their rhetorical style in order to gain their constituents' trust. Specifically, they adjust their degree of phonetic articulation to a level demanded by, and most appropriate in a given situation. When they find themselves in need of demonstrating warmth, as would be the case on the campaign trail, they project likeness by lowering their level of articulation down to that of their constituents. When their primary goal is to establish their competence, such as in Congress, they adjust their articulation upwards, towards a more formal manner of speech. In order for a politician to succeed, they need to be both technocrat and populist, and I argue that rhetoric helps them to navigate these conflicting roles.

The role of trust in my theory is critical, and adapted directly from Fenno (1977). For Fenno, trust is important because voters cannot take the words of a politician at face value. They need to evaluate whether he a) intends to follow through on the promises he made to them, and b) has the capacity to do so. Hence, Fenno splits trust into qualification, identification and empathy. Here, I simplify Fenno's theory, as these three concepts can effectively be mapped onto the commonly used dimensions of competence and warmth employed in political psychology (Fiske et al., 2002).

So why is it that projecting warmth is more important on the campaign trail, while competence is favored in Congress? (Laustsen and Bor, 2017) explore the first half of this question, as they seek to adjudicate between two competing theories regarding candidate evaluations: Do voters care more about warmth or competence when it comes to making a vote choice? The authors find that – consistent with social psychology, and contrary to political science – warmth is more important. The reason here is simply that affect comes before logic – amygdala before prefrontal cortex – so that vote choice depends more on hearts than minds.

My argument for the primacy of competence in Congress rests on the notion that voters associate a certain gravity with being a member of the U.S. government. When in Congress, its members evidently fulfill that role, and any deviations from the social norms associated with it would be viewed negatively. This shines through in the incivility literature, which shows that when members of Congress act in a way not befitting their stature,

trust in government suffers (Mutz and Reeves, 2005). Furthermore, the sociolinguistic literature has found competence-based styles to be associated with perception of greater performance in the speaker's job (Deprez-Sims and Morris, 2010). In this sense, projecting warmth would be the means of a politician getting the job, while competence is essential to doing it.

Finally, it should be noted that my theory is an integration of the models advanced by Fenno (1977) and Grimmer (2013). In Grimmer's words, Fenno's version of home style focuses on the "folksy mannerisms" used by politicians in order to demonstrate to voters that they are one of them. By contrast Grimmer's interpretation of home style is focused on what politicians do in order to explain their activities in Washington to their voters. These two approaches are not incompatible – they are two sides of the same coin – the concept of trust. Fenno focuses on the projection of warmth, whereas Grimmer emphasizes competence. My argument is that it merely depends on the situation which of these two approaches is applied.

Hypotheses

My primary hypothesis can be derived directly from the theoretical discussion above: When speaking in Congress, senators attempt to demonstrate to their constituents that they are sufficiently competent to represent them. On the campaign trail, their main goal is to project warmth. Consequently, a higher level of articulation is expected in Congress.

> **Hypothesis 1**: Articulation will be greater in congressional speeches than campaign speeches.

While the focus of this paper lies on within-legislator variation, between-legislator effects need to be considered as well. My expectations in this regard are derived from theory in both political science as well as sociolinguistics. The most basic expectation is on gender. Existing research has found that female politicians are held to a higher standard by others as well as themselves (Lawless and Fox, 2004; Pearson and McGhee, 2013; Kanthak and Woon, 2015). Consequently, they would have a need to demonstrate competence more so than warmth. Furthermore, the sociolinguistics literature has consistently found that women engage in low-articulation styles of speech to a much lower degree than their male counterparts (Fischer, 1958; Kiesling, 1998).

> **Hypothesis 2**: Articulation will be greater for female than for male senators.

Political and linguistic theory points towards additional hypotheses pertaining to differences in setting within Congress (floor vs. committee), the salience of the issue that

is being discussed, as well as the demographic, cultural and geographic origin of the senator and their constituency. I will test these hypotheses once I have expanded my sample to the full Senate, as there is currently not sufficient variance to investigate these propositions thoroughly.

Data & Methods

Data Source: YouTube

The speech data used in this paper comes from the senators themselves: All U.S. senators maintain at least one, and in most cases two YouTube accounts on which they publish videos of their activities. Most senators use two accounts – one for their election campaigns, and one to document their legislative activities. The former contain ads, speeches at rallies, videos shot at the senators' homes (i.e. "On the Farm with Sharla and John" is an example of a video series which portrays Sen. John Tester as just an ordinary farmer from Montana) or interviews with the media. The latter focus on the legislator's legislative activities, containing sections of video from floor speeches, committee hearings, lobbying for (or celebrating) the passage of bills prioritized by the senator. This division into a campaign and a Senate account is very convenient for me, because it divides a senator's speeches into low- and high-articulation settings, without the need for me to manually classify each video.

Consequently, I downloaded all videos from all senators running for reelection in 2018 and possess a campaign and a senate account.[3] This was facilitated by the command line-based program `youtube-dl`, which enables bulk downloading of all videos on an account. To speed up the procedure, I parallelized the process through the GNU program `xargs`. Along with the videos, I also downloaded metadata and closed captions.

The next step in this pipeline consists of converting the videos from video to audio. When downloading from YouTube, videos are generally either in .mp4, .mkv or .webm format, with an audio sampling rate of 44100Hz (The sampling rate determines at which intervals a sample is taken from a continuous soundwave. A sampling rate of 44100Hz corresponds to 44100 data points per second). Using the Python packages `librosa` and `soundfile`, I extract the audio from each video, resample it to 16000Hz (which is more than sufficient for my purposes) and a bitrate of 256Kb/s (the bitrate determines how many possible values each data point can assume) and save it in the .wav format. For a more thorough explanation of these concepts, see appendix 1.

[3] I exclude Bernie Sanders from my sample because in spite of the fact that he has multiple accounts, the content he puts on them is very different from other senators, largely relying on videos in which he is not the speaker himself.

At this point, I am faced with a more complicated problem. Not every second of a video consists of speech, and not every second of speech actually stems from the respective senator. The use of the senators' YouTube accounts alleviates this issue to a much greater extent than any other source, but it still exists nevertheless. To deal with this issue, I turn to two areas of machine learning and electrical/mechanical engineering: speech diarization and speaker recognition.

Speech Diarization

Speech diarization pertains to the segmentation of an audio stream into sections that actually contain speech, and the segmentation of that speech into different speakers. Importantly, speech diarization does not identify who those speakers are, it merely separates them and keeps track of how many they are. To this end, I rely on `AaltoASR`, a toolkit for acoustic modeling maintained by Aalto University. This program uses Hidden Markov Models (HMM), a commonly used tool of analysis in natural language processing (both speech as well as text), to process speech features in the form of mel frequency cepstral coefficients (MFCC). See the appendix for a detailed explanation of these concepts. The output from this kind of analysis is a list of start and end times of speech, along with a form of "anonymized" identifier for the speaker, i.e. speaker 1, 2, 3, etc. wherein the repeated occurrence of the same speaker is marked as such. In addition to segmenting the speakers, speech diarization also has the advantage of filtering out unwanted noise. Applause is a very common example for this category in my dataset, and if not removed, results in a considerable disturbance of the measures later extracted from the audiostream. Diarization either cuts these sections out entirely if they are not identified as human speech, or alternatively, assigns them to their own "speaker", which means they will be removed once speaker recognition is applied to them.

Speaker Recognition

The next step then consists of identifying whether any one speaker in an audio file is the respective senator or someone else. Speech by other people is quite common, for example, campaign ads, especially negative ones often have narrators and frequently contain testimonies by constituents as well. The videos of the senator in a legislative setting definitely contain a higher proportion of speech from them, but even here, there are the occasional interjections of other senators, or answers by witnesses in legislative hearings. Consequently I train a speech model for each senator using Gaussian Mixture Models (GMMs) and then compare that model with each speech sample. The training data was constructed as following: I hand-picked one video for each senator conforming

to the following constraints: 1) Duration. The video should be around 10 minutes long. 2) Setting. The video should be from a floor speech. 3) "Purity". The video should not contain any voice other than the senator's in question. I limited the pool of potential videos to floor speeches because they are a very controlled setting, with few intervening noises that might confuse the model. Furthermore, each senator has spoken in this capacity, which is important for comparability. If, for example, half of my training data came from floor speeches and the other half from campaign speeches, the classifier might inadvertently learn to pick up on this distinction instead. A sufficiently long sample is also important to ensure that word choice does not influence the model. After the models are trained, I compare each to a speech segment and then compare the log-likelihoods for each senator. The senator with the highest likelihood is identified as the speaker of this particular segment. All segments which are identified as stemming from the senator on whose account the video was published are marked as such and used in the subsequent analysis. Since the model is only capable of recognizing the senators, all segments of speech stemming from other people, such as constituents or committee witnesses will still be classified as senators, but generally not as the senator in question. Nevertheless, it is possible that this classification scheme involves some error, as a different person's voice might sound closer to the senator in question than to any other senator.

Dataset

As not every video contains a segment of speech by the given senator, a substantial amount of the audio files end up being discarded. Of the 8871 videos originally downloaded, 1049 are used in the analysis. These videos run for a combined 21 hours, about 1 hour of speech per senator. However, since audio is an extremely high-resolution form of data with - in my case - 16000 data points per second, the resulting sample still consists of about 1.2 billion data points in total.

Method of Analysis: Vowel Space Analysis

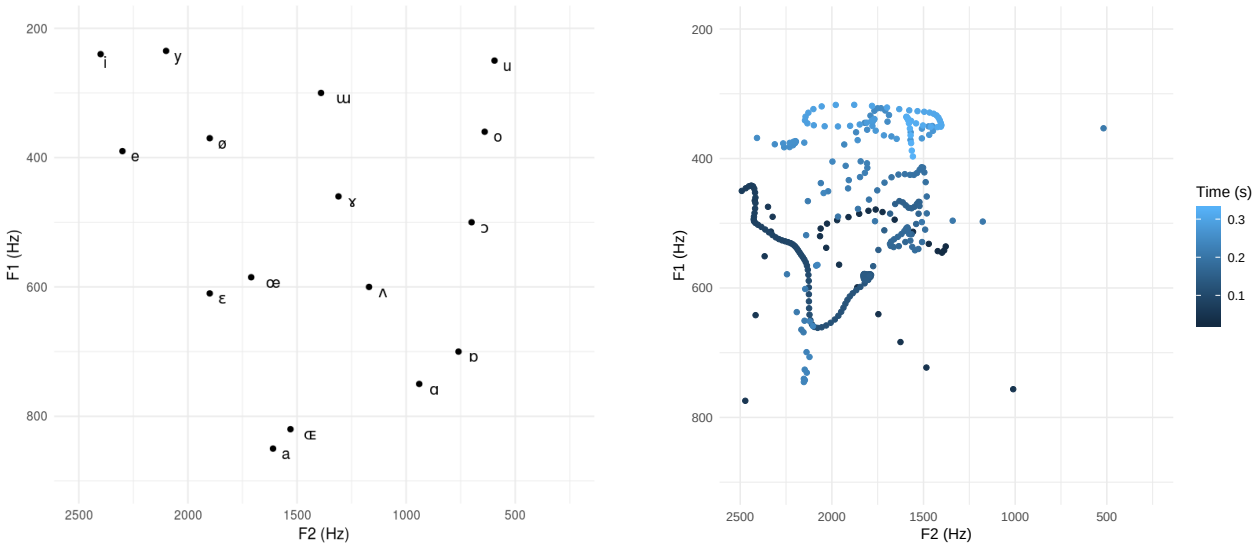
Given the theory of signaling likeness with constituents by shifting phonetic style, the purpose of this chapter is to test this theory in a broad sense. The goal, then, is to measure the performance of a speaker across an entire speech, or even a corpus of speeches – as opposed to specific instances of style-shifting for individual words. The linguistic concept that most closely captures the idea of sophistication and 'elite-ness' is articulation. Here, speech can be placed on a spectrum between hyperarticulation and hypoarticulation, the former pertaining to very clear, downright exaggerated forms of articulation (e.g., like an adult might talk to a little child) whereas the latter would correspond to unclear,

under-articulated speech, such as mumbling. It should be noted that my expectation is that political speech does not reach these extremes, but can be found in the continuum between them. My hypothesis, then, is that due to electoral concerns, politicians signal likeness with their constituents by modulating their overall degree of articulation in accordance with the audience. When addressing a more genteel audience, the need to signal poise and competence requires a greater degree of articulation, which means that it is closer to hyperarticulation. When addressing a more low-brow crowd, politicians need to lower their oratory sophistication, therefore moving closer to hypoarticulation. The unit of analysis here is the speech.

To measure the level of articulation in a speech, I break it down into its components – words, which themselves are comprised of phones. Between the two types of phones – consonants and vowels, the latter is far more informative of the way a speaker talks and sounds. Therefore, articulatory and acoustic phonetics largely deal with the analysis of vowels. To this end, formants are used to indicate which vocal organs are used in what way to produce a sound. The first formant (F1) corresponds to the pharynx – namely, the degree of jaw opening and the second (F2) to the oral cavity – specifically, the tongue position. For example, the vowel [i] corresponds to low F1 (tongue is high in the mouth) and high F2 (front vowel), whereas [a] yields a high F1 (tongue is low in the mouth) and low F2 (back vowel). Figure 1 (a) provides an overview of the average position of vowels - note that this can vary drastically both between and within speakers. Divergent levels of articulation will lead to different ways in which the vocal organs are used, and therefore different levels of F1 and F2. This form of analysis, the measurement of the vowel space area, is a commonly used approach in phonetics. To arrive at these measures, I follow the approach laid out by Story and Bunton (2017), measuring vowel space density.

This method largely relies on identifying how manipulation of the vocal tract leads to the production of different formants, mainly F1 (pharynx - jaw opening) and F2 (oral cavity - tongue position). The position and diversity of these formants allows researchers to make conclusions about vowel space area (VSA), with the idea that more articulate speech makes greater use of the entire VSA. In traditional VSA analysis, this is largely confined to identifying the corner vowels for very specific words, thus relying only on very small snapshots of speeches. This is useful for analyzing precisely how specific words are pronounced. Figure 1 (b) provides an example of Sen. Heidi Heitkamp pronouncing the word gettin'. The caveat of this approach is that it does not work as well across an entire speech. Consequently, Story and Bunton (2017) (see also Sandoval et al. (2013); Whitfield and Goberman (2014)) develop a measure they call vowel space density, plotting the use of F1 and F2 across an entire speech as a heatmap.

The first step, then, is to extract the formants. In my analysis, this is done through



(a) *IPA Vowel Chart* **(b)** *Example Vowel Space Plot*

Figure 1: *Figure (a) shows the International Phonetic Alphabet (IPA) vowel chart, which denominates at which F1 and F2 specific phonemes are produced on average. Figure (b) provides an example vowel space plot of the word gettin', pronounced by Sen. Heidi Heitkamp. In this pronunciation, the 'i' is dropped along with the g, which is visible as there is no activity in the upper left corner where 'i' is located. The other vowel, 'e' is visible in the line of points on the left. Note that the formant range varies between every person, so values in the two plots do not overlap completely.*

the program `Praat` and its R implementation `PraatR`. For this purpose, a ceiling to the formant search range of 5000Hz is set for male and 5500Hz for female speakers. First, the sound signal is down-sampled to twice that value. The audio signal is then divided into segments of 0.025s. The effective length of this window of analysis is 0.05s, because a Gaussian window is used, wherein another, tapered-off 0.0125s to each side of the central window includes signals below -120 dB. Pre-emphasis is applied, meaning that frequencies above 50Hz are amplified, wherein frequencies at 100Hz are amplified by 6dB, and another 6db for each additional 100Hz above. The purpose of this process is to enhance the signal in the noise and allow higher-frequency formants to be captured reliably. This process is illustrated in figure 4 in appendix 2. Then, the Burg LPC (Linear predictive coding) algorithm is used to calculate the frequencies for each formant. At the end of this process, a frequency value is produced for F1 as well as F2, for each 0.025s window.

The relative values of F1 and F2 pairs then provide information about which, and more importantly, how vowels were produced by the speaker. The result is converted to a two-dimensional kernel density. As described in Story and Bunton (2017), the density values are normalized to a range of [0,1] by dividing each value by the maximum value. This process ensures comparability between speeches and speakers. Finally, a convex

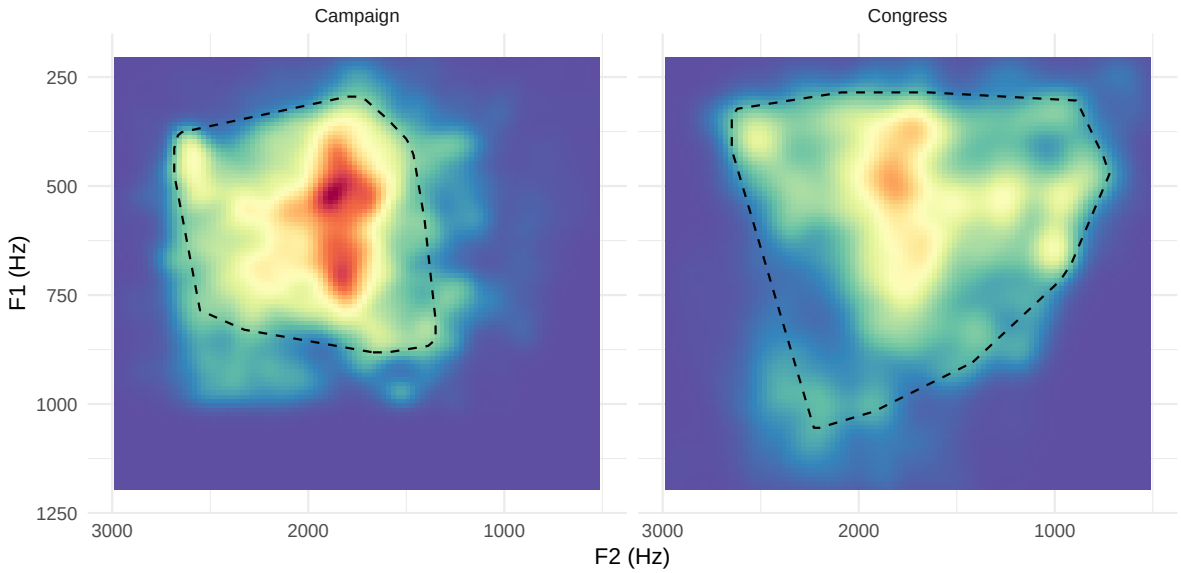


Figure 2: *Vowel space plot for two concatenated speeches of Sen. Heidi Heitkamp – one in a campaign context, and one in Congress. A larger area covered by the high-density areas (enclosed in the black dashed line) corresponds to a greater degree of articulation.*

hull is drawn around the area with a normalized density of 0.25 or higher, indicating the speaker's vowel space area across the speech. The area enclosed within this hull, calculated with the package `phonR`, is my indicator for the level of articulation of a speech, and can then be compared to other speeches. Since the area is large and varies greatly, I use the log of this value in the subsequent analysis. Figure 2 compares two such densities for speeches of Senator Heidi Heitkamp – one from a campaign ad, and the other from Congress. The black dashed line illustrates the convex hull, which encompasses a greater area in the congressional speech.

Results

With a measure for vowel space of each speech in hand, the main hypothesis can be put to the test. To reiterate, I expect greater informality in speeches given in a campaign context, compared to a legislative setting. Consequently, I conduct a two-sample t-test, in which I compare the (log of the) vowel space area of each speech on a campaign YouTube account, to each speech on a legislative account. The results indicate a significantly higher vowel space area for speeches given in Congress ($M = 13.01$, $SD = 0.55$), compared to the campaign setting ($M = 12.82$, $SD = 0.7$), $t(488.74) = 4.1803$, $p = 3.45e-05$. Consequently, I find support for my hypothesis.

A t-test has the advantage of being extremely simple and yet providing an appropriate

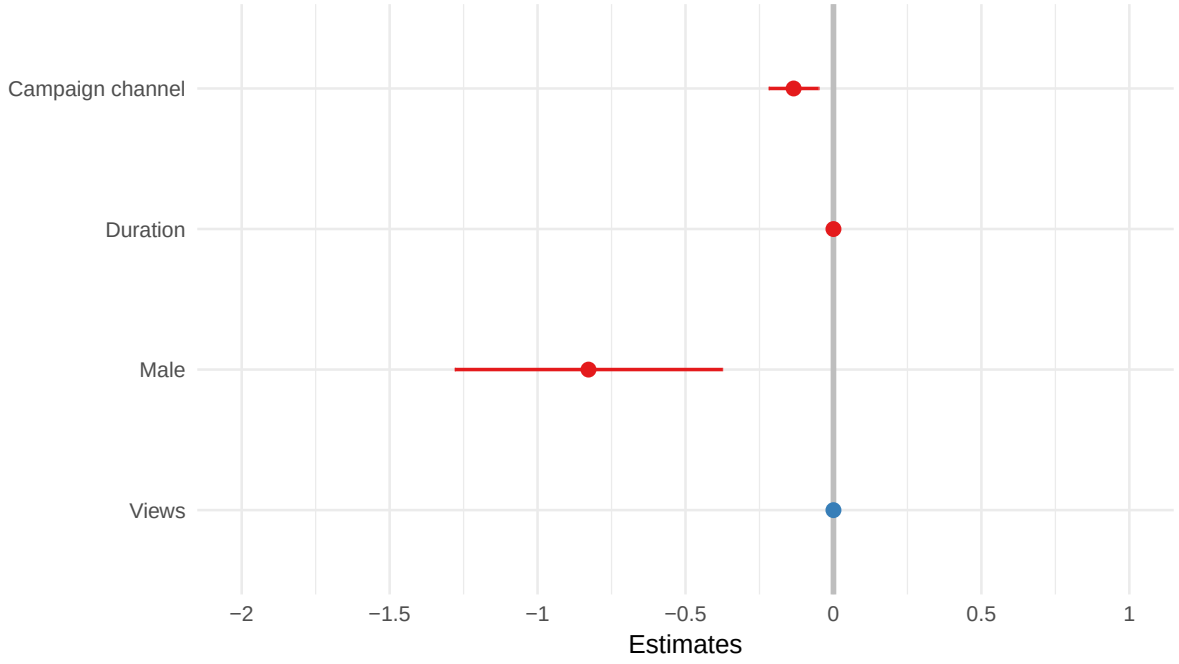


Figure 3: *Estimates from OLS. The figure shows that senators speak more colloquially in a campaign context. Furthermore, male senators speak with a lower degree of articulation.*

and sufficiently rigorous test for the question under investigation here. However, in the form conducted above, it ignores the fact that the individual speeches are clustered by senators, each of whom have their own individual style of speaking. A linear regression allows me to account for this problem and test for the influence of additional variables. As noted above, there are reasons to expect an effect of gender. Furthermore, I control for the duration of the video as well as the number of views it has received. Since I am interested in how the results differ within senators, I use senator fixed effects. Figure 3 shows the result (the regression table can be found in the appendix, see appendix 2, table 1). There is a clear negative effect of the campaign setting on vowel space area, again providing evidence in favor of hypothesis 1. Furthermore – and consistent with linguistic theory – male senators use a lower degree of articulation.

Discussion

In this paper, I have presented evidence for the theory that politicians engage in phonetic style-shifting in order to project likeness with an audience. Electoral goals drive this phenomenon, as representatives find themselves in the position of having to appeal to multiple constituencies, with potentially conflicting interests and opinions. The analysis presented here shows that politicians do indeed speak in a more high-brow manner

when fulfilling their legislative role, as indicated by a greater vowel space area. It seems plausible that this behavior is driven by the need to demonstrate competence and impress the political sophisticates. By contrast, office-holders can demonstrate warmth by addressing voters in a more colloquial tone.

This phenomenon tells us something about representation. My findings match the conclusions of Fenno (1977) more closely than those of Grimmer (2013). Politician do not play either the 'statesman' or the 'appropriator', at the detriment of the other. Rather, as predicted by Fenno, politicians adopt a specific style in accordance with their audience, and they are capable of switching between these styles fluidly, as posited by Saward (2014). Grimmer (2013) does touch on the notion that some senators with multiple constituencies, such as Hillary Clinton, focus on two areas (in this case pork and policy). But as discussed above, this carries significant risks because it often means going on the record with positions that, while pleasing one constituency, will offend another. It is perhaps no coincidence that the most recent losers in presidential elections (Clinton and Romney most prominently, but to a lesser extent also McCain and Kerry) were known for attempting this balancing act, but failing to do so convincingly, thus coming off as insincere "flip-floppers". The approach analyzed in this paper – phonetic style-shifting – affords the practitioner much greater plausible deniability, and, thanks to how finely calibrated humans are to communicative subtleties, might be just as effective.

The research carried out in this paper represents one of the first attempts in political science to leverage the information contained in the audio of speeches[4]. A correlate of this novelty is that it necessarily cannot address every potential issue and therefore comes with a number of caveats:

The dichotomy of campaign versus legislative YouTube accounts is helpful in practice, but it may be overly simplistic. One, not every senator keeps to this practice. A few senators use only one account, and post both campaign and congressional activities on it. Manual coding of these videos into either category would allow for a moderate increase in sample size.

There are also a few cases in which I used an old campaign account, because a newer one was not available. This can happen if a senator comes from an electorally secure state – they still have to build up their name recognition the first time around, but after that, they do not face the danger of being unseated and therefore forego the trouble and expense of making campaign videos.

By not hand-coding anything about the setting of the video, I am also unable to determine which exact context it took place in. For example, I do not have any information on whether a video steams from a campaign ad, a rally or a town hall. Similarly, it might

[4]See Dietrich, Enos and Sen (2017); Knox and Lucas (2018) for other examples.

be of interest whether a campaign video was an ad which might have been broadcast in an area with a specific socio-demographic profile.

The theory and results presented in this paper raise further questions which I plan to address in future research. While vowel space area is a frequently used measure from the phonetics literature which neatly captures the concept I am trying to measure, it is not the most intuitive approach. To non-phoneticians, a vowel space area of, say, 60,000, does not really mean anything. Furthermore, I have largely motivated my research with common examples of "folksy mannerisms" such as g-dropping. Consequently I plan to develop a method for the detection of this kind of phrasing and test whether it relates to the target audience in a similar manner. Research done by (Yuan and Liberman, 2011) has already approached the issue of measuring g-dropping directly, and I intend to build on this literature.

Another aspect of this research, which is directly related to my theory, is the question of topic. For example, issues such as unemployment, crime or immigration have a very direct appeal to ordinary citizens. By contrast, other issues, such as foreign policy are further removed from the common man and therefore carry more interest for political sophisticates (this also touches on the debate originating from, among others, (Miller and Stokes, 1963). It follows that depending on the issue they are currently discussing, politicians would assume either a low- or high-brow style of talking. Consequently I intend to measure the within-speech variance in articulation, conditional on the topic. This also connects my research to the other area of natural language processing which is far more prominent in political science - text.

Finally, all of this research *assumes* that politicians modulate the degree of articulation in their speech because different forms of speaking have different effects on the audience. However, causal evidence for this relationship is another matter. Through a survey experiment in Thailand, Ricks (2018) has shown that different styles of speaking do indeed have the expected effect on listeners: While informal and local language cause respondents to rate politicians higher on likability and kinship, formal language is seen as a signal for competence. I plan to carry out a similar experiment in the U.S.

This paper then represents the first exploration of this topic - it establishes a theory of phonetic style-shifting in the service of symbolic and implied descriptive representation. Furthermore, it tests that theory by developing a measure for the primary theoretical concept in question - articulation. This measure – vowel space area – permits me to test the theory in the broadest sense possible, but it also sets me up for investigating its subtler aspects. The evidence presented here supports the theory of rhetorical style shifting, and I plan to expand on it in future work.

References

Achen, Christopher and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton: Princeton University Press.

APSA. 1950. "Part I. The Need for Greater Party Responsibility." *The American Political Science Review* 44(3):15–36.

Arnold, R. Douglas. 1990. *The Logic of Congressional Action*. Yale University Press.

Bartels, Larry M. 2009. *Unequal Democracy: The Political Economy of the New Gilded Age*. Princeton: Princeton University Press.

Box-Steffensmeier, Janet M., David C. Kimball, Scott R. Meinke and Katherine Tate. 2003. "The Effects of Political Representation on the Electoral Advantages of House Incumbents." *Political Research Quarterly* 56(3):259–270.

Broockman, David E. and Daniel M. Butler. 2017. "The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication." *American Journal of Political Science* 61(1):208–221.

Butler, Daniel M. 2014. *Representing the advantaged: How politicians reinforce inequality*. Cambridge University Press.

Cicero, Marcus Tullius. 1986. *De Oratore*. J.S. Watson (Translator).

Cohen, Linda R. and Roger G. Noll. 1991. "How to vote, whether to vote: Strategies for voting and abstaining on congressional roll calls." *Political Behavior* 13(2):97–127.

Converse, Philip E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. David Apter. University of Michigan.

Cramer, Katherine J. 2016. *The politics of resentment: Rural consciousness in Wisconsin and the rise of Scott Walker*. University of Chicago Press.

Deprez-Sims, Anne Sophie and Scott B. Morris. 2010. "Accents in the workplace: Their effects during a job interview." *International Journal of Psychology* 45(6):417–426.

Desmarais, Bruce, Raymond J. La Raja and Michael S. Kowal. 2015. "The fates of challengers in U.S. house elections: The role of extended party networks in supporting candidates and shaping electoral outcomes." *American Journal of Political Science* 59(1):194–211.

Dietrich, Bryce J, Ryan D Enos and Maya Sen. 2017. "Gender Dynamics in Elite Political Contexts: Evidence from Supreme Court Oral Arguments.".

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy.".

Eulau, Heinz and Paul D. Karps. 1977. "The Puzzle of Representation: Specifying Components of Responsiveness." *Legislative Studies Quarterly* 2(3):233–254.

Fenno, Richard. 1978. *Home style: House members in their districts*. New York: Pearson College Division.

Fenno, Richard F. 1977. "U.S. House Members in Their Constituencies: An Exploration." *American Political Science Review* 71(3):883–917.

Fischer, John L. 1958. "Social Influences on the Choice of a Linguistic Variant." *WORD* 14(1):47–56.

Fiske, Susan T., Amy J.C. Cuddy, Peter Glick and Jun Xu. 2002. "A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition." *Journal of Personality and Social Psychology* 82(6):878–902.

Fox, Justin and Kenneth W. Shotts. 2009. "Delegates or Trustees? A Theory of Political Accountability." *The Journal of Politics* 71(4):1225.

Gilens, Martin and Benjamin I. Page. 2014. "Testing Theories of American Politics: Elites , Interest Groups , and Average Citizens Martin Gilens Benjamin I . Page forthcoming Fall 2014 in Perspectives on Politics." *Perspecties in Politics* 12(3):564–581.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. New York: Doubleday.

Grimmer, Justin. 2013. *Representational Style in Congress: What Legislators Say and Why It Matters*. Cambridge University Press.

Grimmer, Justin Ryan. 2010. "Representational Style: The Central Role of Communication in Representation." *PhD Thesis* .

Hacker, Jacob and Paul Pierson. 2010. *Winner-take-all politics: How Washington made the rich richer - and turned its back on the middle class*. Simon and Schuster.

Hall, Richard L and Alan V Deardorff. 2006. "Lobbying as Legislative Subsidy." *American Political Science Review* 100(1):69–84.

Jones, David R. 2003. "Position Taking and Position Avoidance in the U.S. Senate." *The Journal of Politics* 65(3):851–863.

Jurafsky, Daniel and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River: Pearson/Prentice Hall.

Kanthak, Kristin and Jonathan Woon. 2015. "Women Don't Run? Election Aversion and Candidate Entry." *American Journal of Political Science* 59(3):595–612.

Kiesling, Scott Fabius. 1998. "Men's identities and sociolinguistic variation: The case of fraternity men." *Journal of Sociolinguistics* 2(1):69–99.

Knox, Dean and Christopher Lucas. 2018. "A Dynamic Model of Speech for the Social Sciences.".

Laustsen, Lasse and Alexander Bor. 2017. "The relative weight of character traits in political candidate evaluations: Warmth is more important than competence, leadership and integrity." *Electoral Studies* 49:96–107.
URL: *http://dx.doi.org/10.1016/j.electstud.2017.08.001*

Lawless, Jennifer L and Richard L Fox. 2004. "Why Don't Women Run for Office ?" *Brown Policy Report* .

Leibovich, Mark. 2014. *This Town*. New York: Penguin Group.

Lenz, Gabriel S. 2009. "Learning and opinion change, not priming: Reconsidering the priming hypothesis." *American Journal of Political Science* 53(4):821–837.

Liberman, Mark. 2008. "Empathetic -in' October.".
URL: *http://languagelog.ldc.upenn.edu/nll/?p=732*

Liberman, Mark. 2011. "Pawlenty's linguistic "southern strategy"?".
URL: *http://languagelog.ldc.upenn.edu/nll/?p=3032*

Maestas, Cherie. 2003. "The Incentive to Listen: Progressive Ambition, Resources, and Opinion Monitoring among State Legislators." *The Journal of Politics* 65(2):439–456.

Mansbridge, Jane. 1999. "Should Blacks Represent Blacks and Women Represent Women? A Contingent "Yes"." *The Journal of Politics* 61(03):628.

Mayhew, David. 1974. *Congress: The Electoral Connection*. Yale University Press.

Miller, Warren E. and Donald E. Stokes. 1963. "Constituency Influence in Congress." *The American Political Science Review* 57(1):45–56.

Mutz, Diana C. and Byron Reeves. 2005. "The New Videomalaise: Effects of Televised Incivility on Political Trust." *American Political Science Review* 99(1):1–15.

Nunberg, Geoff. 2008. "Palin's tactical g-lessness.".
URL: *http://languagelog.ldc.upenn.edu/nll/?p=733*

Page, Benjamin I., Larry M. Bartels and Jason Seawright. 2013. "Democracy and the policy preferences of wealthy Americans." *Perspectives on Politics* 11(1):51–73.

Pearson, Kathryn and Eric McGhee. 2013. "Should Women Win More Often than Men? The Roots of Electoral Success and Gender Bias in U.S. House Elections." *SSRN* .

Ricks, Jacob I. 2018. "The Effect of Language on Political Appeal: Results from a Survey Experiment in Thailand." *Political Behavior* pp. 1–22.

Rothenberg, Lawrence S. and Mitchell S. Sanders. 2000. "Severing the Electoral Connection: Shirking in the Contemporary Congress." *American Journal of Political Science* 44(2):316–325.

Sandoval, Steven, Visar Berisha, Rene L. Utianski, Julie M. Liss and Andreas Spanias. 2013. "Automatic assessment of vowel space area." *The Journal of the Acoustical Society of America* 134(5):EL477–EL483.

Saward, Michael. 2014. "Shape-shifting representation." *American Political Science Review* 108(4):723–736.

Schattschneider, E.E. 1960. The Scope and Bias of the Pressure System. In *The Semi-Sovereign People: A Realist's View of Democracy in America*, ed. E.E. Schattschneider. Hynsdale, NY: The Dryden Press.

Schlozman, Kay L., Sidney Verba and Henry E. Brady. 2012. *The Unheavenly Chorus. Unequal Political Voice and the Broken Promise of American Democracy*. Princeton, NJ: Princeton University Press.

Story, Brad H. and Kate Bunton. 2017. "Vowel space density as an indicator of speech performance." *The Journal of the Acoustical Society of America* 141(5):EL458–EL464.

Whitfield, Jason A. and Alexander M. Goberman. 2014. "Articulatory-acoustic vowel space: Application to clear speech in individuals with Parkinson's disease." *Journal of Communication Disorders* 51:19–28.
URL: *http://dx.doi.org/10.1016/j.jcomdis.2014.06.005*

Yuan, Jiahong and Mark Liberman. 2011. "Automatic detection of "g-dropping" in American English using forced alignment." *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings* pp. 490–493.

Appendix 1: Technical Appendix

This section of the appendix attempts to explain commonly used techniques from natural language processing, mechanical/electrical engineering and phonetics for a political science audience. As such, it cannot cover every aspect in full detail. The reader is encouraged to refer to the relevant literature, such as Jurafsky and Martin (2008), for further reference.

Audio Data

At a fundamental, physical level, sound is a wave traveling through and disturbing a medium, such as the air. When undisturbed, the medium is in equilibrium. When a wave propagates through it, the **sound pressure** causes a corresponding disturbance, also referred to as **amplitude**. The high points (i.e. maximum distance to the equilibrium) of the wave are referred to as **crests** and the low points as **troughs**. When traveling through air, a sound wave moves at a speed of about 343m/s (depending on the temperature). The **frequency** f of a wave describes the number of cycles it undergoes per time period T, usually per second - which is then described as hertz, or Hz.

$$f = \frac{1}{T}$$

This means that for a wave traveling at $20Hz$, $0.05s$ pass between two crests (i.e. a **wavelength**).

The method used here for representing an analog signal in digital form is pulse code modulation. To do so, the amplitude of the signal is sampled at regular intervals. The number of these intervals within a given time frame (i.e. the frequency) corresponds to the **sampling rate**. Each cycle of a wave needs to be represented by at least two data points, one for the positive and one for the negative section of the wave. Sound perceptible to humans occurs between 20 Hz and 20,000 Hz. Hence, a sampling rate of 40,000 Hz (i.e. 40,000 data points per second) would be necessary to adequately represent this information. The audio tracks for most of the videos used in this analysis have a sampling rate of 44,100 Hz. Human speech generally only occurs at frequencies below 10,000 Hz, and the formants (see below) which constitute my most important form of data occur below 5,000 Hz for males and 5,500 Hz for females (young children can go up to 8,000 Hz, but this is of no concern to this analysis). Consequently, encoding audio at such high quality is not strictly necessary for my purposes.

Furthermore, the **bit depth** determines the number of possible values at each interval.

For example, a sampling rate of 16000 Hz means there are 16000 samples per second, and a bit depth of 16bit means that each sample has a resolution of 2^{16} possible values. This means that integers between -32768 and 32767 can be represented, entailing a fairly high level granularity.

Another concept of some relevance here is the number of channels. Most audio, including the videos scraped from YouTube in this paper, is in stereo format, meaning two channels. For the analyses in this paper, this is entirely irrelevant, so I convert all of my audio data to mono.

Hence, both sampling rate and bit depth determine the audio quality. These concepts, along with the number of channels, determine the bitrate as following:

$$\text{Bit rate} = \text{sampling rate} * \text{bit depth} * \text{channels}$$

As an example, uncompressed .wav files usually store data at 16-bit, 44.1 kHz and consist of two audio channels (i.e. stereo), so 1 hour of audio requires 635.04 MB of storage (about the size of a CD). Even a compressed[5] MP3 file with a bit rate of 128kB/s (i.e. the number of bits used for each second of audio) still weights in at 57.6 MB. It follows that the storage demands for this project are rather large. Since, as discussed above, human speech does not fill the entire spectrum of human aural perception, I address this problem by using a lower sampling rate: In this paper, I use .wav files with a sampling rate of 16000 Hz, a bit depth of 16bit, and one channel, resulting in a bitrate of 256Kbit/s.

The Fourier transform

The Fourier transform is an extremely commonly used technique in the processing and analysis of audio data. In this paper, it is used at several steps, such as the extraction of formants, speech diarization and speaker recognition. It is used to convert a signal from the time to the frequency domain.

There are two types of the Fourier transform - one for continuous and one for discrete data. Since audio data in the digital domain is in discrete form, I generally rely on the latter.

Continuous Fourier transform. Applied to an analog signal of potentially infinite length.

$$X(F) = \int_{-\infty}^{\infty} x(t) \mathrm{e}^{-j2\pi Ft} \, \mathrm{d}t \tag{1}$$

Discrete Fourier transform. The discrete Fourier transform is applied to a windowed

[5] Audio compression is based on filtering out information that is imperceptible to the human ear and is therefore redundant.

portion $x[n]...x[m]$ of a signal.

$$X_k = \sum_{n=0}^{N-1} x_n \mathrm{e}^{-\frac{j2\pi kn}{N}} \tag{2}$$

The output X_k, represents the k-th frequency bin. Since the audio signal is sampled at discrete points, this means that an amplitude is being calculated at each frequency. There are n such frequency bins. For example, if a signal is sampled 8 times, n will be $n = [0, 1, 2, ...7]$. The points k at which the frequencies are sampled are all multiples k of the fundamental frequency $\frac{2\pi}{T}$ in the continuous case, and $\frac{2\pi}{N}$ in the discrete case.

Mel-frequency cepstral coefficient (MFCC)

Mel-frequency cepstral coefficients (MFCC) are a commonly used feature in machine learning applied to audio tasks. MFCCs denominate how much energy exists in regions of the frequency domain. This matters because human hearing cannot perceive very small differences in frequencies, but is generally better at doing so for lower frequencies. Consequently the filterbank determines the increasing distances of the 'bins' into which frequencies are grouped.

To this end, frequencies are converted to mels according to the following equation[6]:

$$M(f) = 1125 \ln(1 + \frac{f}{700}) \tag{3}$$

To convert back to frequency, the following equation is used:

$$f = 700(\exp \frac{m}{1125} - 1) \tag{4}$$

To calculate the MFCCs, the audio signal is portioned into a set of windows, each of which consists of a number of frames. Then, the discrete fourier transform, as outlined above, is applied to such a window. Then, the periodogram power spectral density estimate is calculated by squaring the absolute value of the output of the DFT and dividing by the number of samples in the window. Then, the mel-spaced filterbank is applied to this periodogram, yielding 26 coefficients. After taking the log of these numbers and applying the discrete cosine transform, I am left with 26 cepstral coefficents for each frame. In this paper, MFCCs are used for both speech diarization and speaker recognition. For the formant extraction, Linear Prediction Coefficients (LPCs), a similar type of feature, are used. The speaker recognition program also makes use of delta MFCCs, which denominate the change in MFCCs between frames. For further reference,

[6]Note that there is no ubiquitous equation for this. The constant in front of the log is not always exactly the same.

see Jurafsky and Martin (2008).

Gaussian Mixture Models (GMMs)

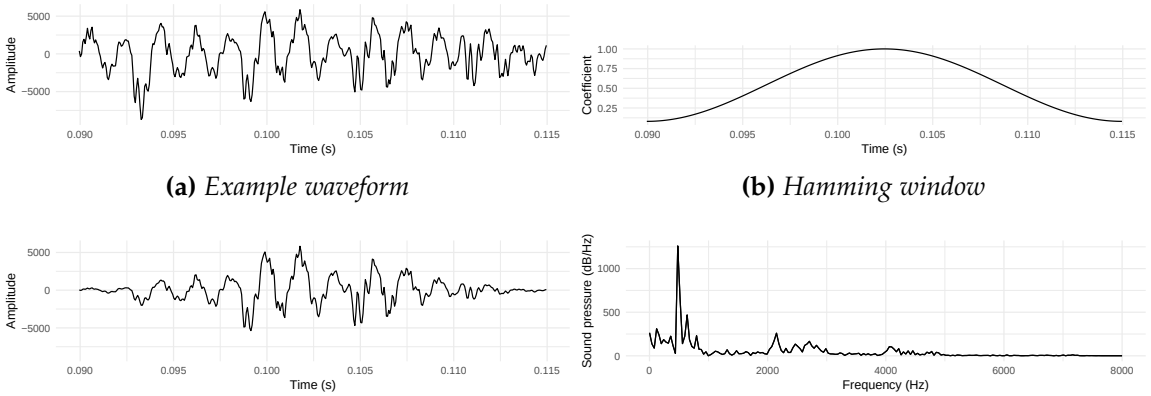
Gaussian mixture models are generative models describing the distribution of data. They are frequently used as a probabilistic clustering model, which addresses some of the shortcomings of k-means clustering that some political scientists might be familiar with. GMMs assume a multivariate Gaussian distribution, which gives them more flexibility than, for example, k-means clustering. The analogue of clusters in this model are its components. The data-generating process assumes the selection of a component, and the subsequent generation of a data point from a normal distribution according to the model parameters. These parameters are learned via Expectation Maximization (EM).

In the case of speaker maximization, I use a GMM with 16 components and train one model for each of the senators in the sample. Essentially, the GMM learns, in an unsupervised manner, how the MFCCs for the speaker's training data were generated from a multivariate normal distribution. At test time, the model computes, for each frame and for each component, the log-likelihood that this test data was generated, given the model parameters of each speaker model. These log-likelihoods are then summed up, and the speaker model with the highest log-likelihoods is predicted to be the actual speaker of the sample.

Appendix 2: Tables & Figures

Table 1: *Linear regression. The effect of setting (campaign/Congress) on vowel space area. The table shows that senators speak more colloquially in a campaign context.*

	Dependent variable:
	Vowel space area
Campaign channel	-0.135^{***}
	(0.043)
Male	-0.828^{***}
	(0.231)
View count	0.00000
	(0.00000)
Video duration	-0.0001
	(0.001)
Senator fixed effects	
Observations	1,060
R^2	0.249
Adjusted R^2	0.233
Residual Std. Error	0.531 (df = 1036)
F Statistic	14.964^{***} (df = 23; 1036)
Note:	*p<0.1; **p<0.05; ***p<0.01



(a) *Example waveform*

(b) *Hamming window*

(c) *Waveform after the hamming window has been applied*

(d) *Spectrum, after the Fourier transform has been applied to the waveform*

Figure 4: *These four figures show the process of transforming a basic waveform into a spectrum. This spectrum will then, in the next step, be turned into a spectrogram, from which formants can be identified.*

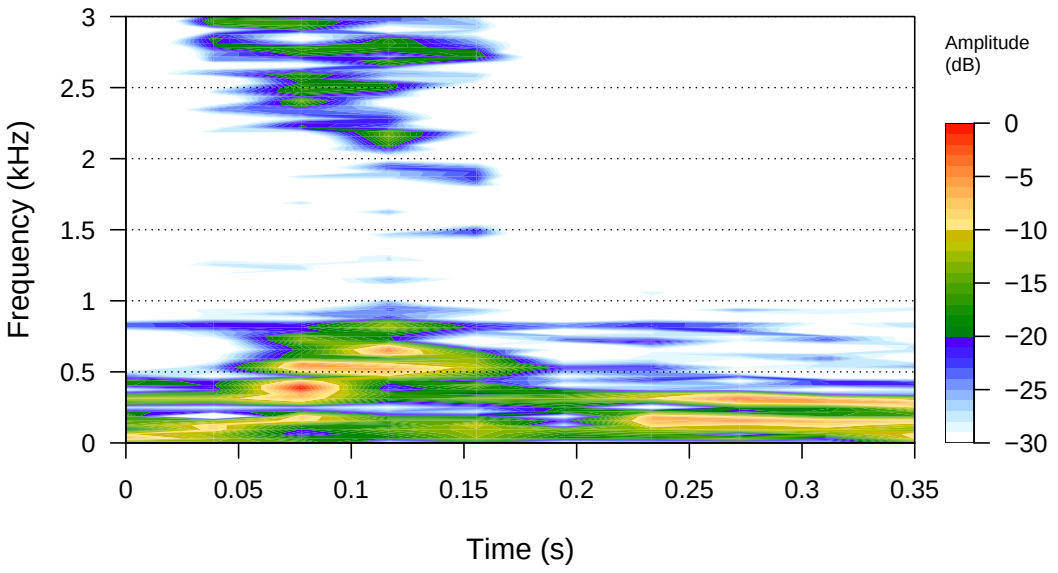


Figure 5: *Spectrogram of a single word. The spectrogram is essentially a spectrum over time, plotted as a heatmap. The high-density areas that go on for some time correspond to the formants, from the bottom up. So F1 corresponds to the bottom "line", at around 0.25 khZ, F2 the next between 0.4-0.6 kHz, and so on.*