Speaking Fast and Low: The Acoustics of Authoritarian Politics*

Marco Giani and Moses Shayo†

Preliminary Draft

February 28, 2026

Abstract

How do people speak where there is no freedom of speech? Using a speech corpus built for speech recognition, we compare Mandarin speakers from Mainland China and Taiwan reading sentences that are effectively randomly assigned and classified as political or non-political. In a difference-in-differences design with individual and sentence fixed effects, Mainland Chinese men speak 0.2 SD faster and 0.4 SD quieter when reading political content, an acoustic pattern associated with mild fear in Mandarin. No comparable effect appears among women, mirroring the gendered pattern of political repression. A placebo using American and British English speakers yields null results. The effect is more pronounced in provinces that experienced high fatalities during the Cultural Revolution and among older cohorts, but not in provinces more repressed today. This suggests an internalized legacy of fear embedded in speaking habits, not just rational responses to current surveillance.

Keywords: Authoritarian politics, China, free speech, audio data, gender politics

*We thank Maya Bar-Hillel, Simon Hix, Filip Kostelka, Brechtje Post, and audiences at the Cambridge Phonetics Laboratory, the Center for the Study of Rationality, and EUI comparative politics seminar, for valuable comments and suggestions. Financial support from King's College London is gratefully acknowledged.

†Giani: Department of Political Economy, King's College London. Shayo: Department of Economics, Hebrew University, and Department of Political Economy, King's College London.

1 Introduction

Appeals to "give citizens a voice" have long served as metaphors for political representation and empowerment, especially in authoritarian contexts (e.g., Hirschman 1970). We shift from metaphor to method and ask: does authoritarianism shape not only *what* citizens say but *how* they say it: how fast, how loud, how acute or grave they sound when talking about politics? Self-censorship is central to how authoritarian regimes function, yet it is inherently hard to measure. Survey responses may themselves be distorted by the very phenomenon one seeks to detect. Speech acoustics offer a different window. Lying partly beyond conscious control, and absorbing local norms and cultures of speech, they can reveal behavioral tendencies that standard instruments miss.

We test this proposition by comparing Mandarin speakers from Mainland China and Taiwan. This setting is uniquely suited to our question. The two populations share a language, allowing us to hold constant the many sentence-level characteristics that shape how text is spoken, yet inhabit starkly different institutional environments: Mainland China scores 9/100 on the Freedom House index (below Cuba and Iran), while Taiwan scores 94/100 (above the United States and the United Kingdom, Freedom House 2024). Our data come from the Common Voice Corpus (Mozilla 2024), an open-source dataset developed for speech recognition technology. Volunteers record short sentences displayed on their screen; crucially, sentences are effectively randomly assigned. We extract three acoustic measures—tempo, volume, and pitch—from approximately $300,000$ voice clips, and classify each sentence as political or non-political. Our difference-in-differences design, with both individual and sentence fixed effects, identifies the effect from within-person, within-sentence variation: how the *same* speaker's acoustics shift when reading a political rather than a non-political sentence, relative to the same shift under different political institutions.

We find that Mainland Chinese men speak *fast and low* when reading political sentences. Relative to Taiwanese men, they increase their tempo by 0.2 characters per second (approximately 0.2 standard deviations) and lower their volume by 2.3 decibels (approximately 0.4 standard deviations). No comparable effect appears among women. This acoustic pattern—faster and quieter speech—is associated with mild fear and stress in the Mandarin phonetics literature (Zhang, Ching, and Kong 2006, Li and Huang 2023, Chnag, Young, and Yuen 2010, Anolli et al. 2008). We do not find robust effects on pitch, consistent with the fact that tonal languages such as Mandarin constrain the use of pitch for emotional expression. The results are robust across alternative measures of the dependent variables, alternative definitions of political content, and alternative sampling restrictions. A placebo test comparing American and British English speakers—who share the language but face no comparable

institutional variation—yields null results for all outcomes and both genders. A randomization test confirms that the effects are unlikely to arise from any other, arbitrary, partition of sentences.

The concentration of the effect among men mirrors the gendered pattern of political repression in China: men constituted approximately 70% of those killed during the Cultural Revolution (MacFarquhar and Schoenhals 2006, p128) and remain the large majority of political prisoners today (Global Voices 2018). We explore whether the acoustic effect reflects rational caution in response to current surveillance, or an internalized legacy of historical repression. Exploiting within-China geographic variation, we find that the effect is driven by men from provinces with high rates of political killings during the Cultural Revolution (1966--71), not by men from provinces with high rates of political imprisonment today. Consistently, the effect is stronger among older cohorts. We term this the *acoustic legacy hypothesis*: vocal patterns around political speech reflect internalized dispositions—akin to Bourdieu's (1977, 1990) concept of *habitus*—shaped by historical repression and transmitted through family, schooling, and local norms, rather than deliberate risk calculation alone.

This paper makes three contributions. First, it provides direct quasi-experimental evidence that authoritarianism shapes how citizens speak about ordinary politics. If even low-stake political content triggers acoustic markers of fear, this may have implications for both preference revelation and information aggregation: citizens may distort the signals they send not only in surveys but in everyday discourse, compounding the epistemic gap that authoritarian regimes face in learning about policy needs (Kuran 1995, He and Warren 2011, Truex 2017, Kim et al. 2022). Second, it advances understanding of the gendered and historically rooted character of political behavior under authoritarianism, contributing to a growing literature on the legacies of repression (Rozenas and Zhukov 2019, Reisinger, Zaloznaya, and Woo 2023), the economics of culture and institutional persistence (Bisin and Verdier 2001, Nunn 2012, Alesina, Giuliano, and Nunn 2013, Giuliano and Nunn 2021), and gender and politics in China (Su 2006, Fan, Pan, and Zhang 2025). Third, it demonstrates that speech corpora, originally built for AI applications, can serve as a tool for social science, offering a lens on political behavior that is especially valuable in contexts where survey data are scarce or unreliable.

2 The Acoustics of Authoritarian Politics

This paper proposes and tests the proposition that human voices can reveal behavioral tendencies that are hard to gauge using standard survey and behavioral measures. Our central hypothesis is that threatened citizens under authoritarian rule, or those scarred by

political repression, display distinctive acoustic markers of stress when prompted to talk about politics—including ordinary politics. In the case of China, this refers overwhelmingly to men. Despite socialist claims of gender equality, women remain marginalized in politics—offline and online (e.g., Su 2006, Jiang and Zhou 2022, Fan, Pan, and Zhang 2025, Wagner, Gainous, and Abbott 2021). Politics is a male business, and so is political repression. Men made up approximately 70% of those killed in the late sixties during the Cultural Revolution (MacFarquhar and Schoenhals 2006, p128), and they are equally over-represented among those jailed for political crimes today (Global Voices 2018.

This hypothesis draws on two fields. First, comparative politics, which allows us to establish how the institution of free speech or lack thereof shape the characteristics of daily political discussion and self-censorship. Second, socio-phonetics, which has linked particular acoustic patterns of tempo, volume, and pitch to particular emotions such as anger, joy, or stress in general and for Mandarin in particular.

While authoritarian regimes seek to control the masses (Svolik 2012), contemporary China does not eliminate all forms of online or offline criticism emerging in everyday communication (e.g., King, Pan, and Roberts 2013, Qin, Strömberg, and Wu 2017, Chen and Xu 2017, Distelhorst and Hou 2017, Kloet and Poell 2025). This might allow for some genuine political discussion—and information aggregation—at least regarding mundane, day-to-day matters of public policies that do not directly threaten the regime (e.g., Wong and Liang 2023). Indeed, advocates of the Chinese ″deliberative authoritarianism″ model stress that the CCP does not simply tolerate but sometimes explicitly elicits constructive criticisms. This allows the regime to reduce the epistemic gap that authoritarian regimes typically face in learning about policy needs and preferences (He and Thøgersen 2010, He and Warren 2011, He and Wagenaar 2018, Pan and Xu 2018). In practice, however, the boundaries of permissible speech are opaque, which may lead citizens to avoid political discussion altogether (Kim et al. 2022, Wedeen 2022, Jiang 2021). In other words, even when this is unintended, the very nature of authoritarian rule may restrict political discourse not only on sensitive topics but also on routine policy issues.

Whether constraints on free speech silence casual political discussion (and not just dissent) is inherently hard to test. Some studies of contemporary China attempt to measure the extent of self-censorship by looking at patterns of overly positive answers or missing responses in surveys, but the evidence remains limited. Consider the 2018 World Values Survey. 82.2% of respondents from Mainland China agree that the government has the right to ″keep people under video surveillance in public areas″ , with less than 1% missing data, a similar rate to ″how often do you discuss politics″ . It is possible that citizens genuinely support this policy, as argued by Yang (2025). But it is also possible that citizens are reluctant

to say they do not support it.[1]

We suggest that the acoustic characteristics of political discussion provide an unexplored window into how institutional environments shape ordinary discussion. However, implementing this idea—i.e., isolating the effect of institutional constraints on speech acoustics—requires two critical sources of variation. First, we need institutional variation in political freedom between countries within the same language to control for how sentence characteristics (such as length and tonal requirements) inherently shape tempo, volume, and pitch across languages. Second, we need variation in topics of conversation spoken by the same individual, to control for individual speaking styles, since individuals naturally vary in how loud or fast they speak. Our research design satisfies both requirements. We test whether citizens from Taiwan and Mainland China display distinctive acoustic markers—suggestive of self-censorship—when speaking about ordinary political topics, compared to non-political topics. Specifically, the acoustics of authoritarianism should feature political speech that is both faster and quieter, a pattern that the Mandarin phonetics literature associates with stress and mild fear. (Zhang, Ching, and Kong 2006, Li and Huang 2023, Chnag, Young, and Yuen 2010, Anolli et al. 2008). Intuitively, people may speak quickly to finish the task, and softly to avoid notice. Importantly, however, this need not be a conscious or deliberate choice. It may reflect internalized local norms and cultures of speech.

While the verbal dimension of speech has received extensive attention in political science through text analysis, research on its nonverbal components remains very limited. At the intersection between political science and biology, the few papers on voice and politics share a research interest in gender and political leadership, a methodological focus on pitch, and an institutional focus on democracy and elections. These studies show that low pitch can affect political outcomes by conveying charisma, credibility, or authority across and within genders (Klofstad, Anderson, and Nowicki 2015, Klofstad 2016, Banai, Banai, and Bovan 2017, Searles et al. 2020, Klofstad 2017, Anderson and Klofstad 2018, Boussalis et al. 2021, Cinar and Kıbrıs 2024)—a result that is not specific to political leadership (e.g., Mayew, Parsons, and Venkatachalam 2013). Moving beyond how leaders' voices are

[1]Jiang and Yang (2016) show that expressed political support for the government surged in surveys following a political purge in Shanghai in 2006, suggesting extensive preference falsification. Meanwhile, Shen and Truex (2021) show that nonresponse rates are similar in authoritarian regimes and democracies, casting doubt about how sensitive discussion of politics is in the former. Carter, Carter, and Schick (2024) use two carefully designed survey experiments to find that institutional support for the CCP is grossly overestimated in public surveys: when asked about the same support in a list experiment, support drops spectacularly. In turn, such a result contradicts previous results by Lei and Lu (2017). Studying both support levels and nonresponse rates in China, Lei and Lu (2017) shows that randomly knowing whether a survey is commissioned by the CCP does not affect nonresponse rates, and can even decrease support. But directly asking citizens in Hong Kong about the extent of their self-censorship amid deteriorating democratic conditions following the 2020 crackdown reveal strong levels of self-censorship (Yuen and Lee 2025).

perceived, Dietrich, Hayes, and O′ brien (2019) study how variations in vocal pitch can themselves reveal underlying emotions. Analyzing Congressional floor speeches, they show that female Members of Congress speak with greater emotional intensity when talking about women as compared with both their male colleagues and their speech on other topics. While Dietrich, Hayes, and O′ brien (2019) uses individual-level FE, the inability to fully control for sentence characteristics—unavoidable given that Members of Congress rarely repeat the exact same text—leaves open the possibility that content differences drive the results. This is a powerful demonstration that audio data can convey political information inaccessible through text alone or other traditional means such as surveys.

We build on these insights to advance the literature in five important ways. First, we study authoritarian rather than democratic regimes. Unlike democracies, autocracies typically limit the collection and distribution of data, especially if suspected to have some political meaning. And as mentioned, traditional tools like surveys, interviews, and experiments may be less useful in autocracies. Second, we study citizens rather than leaders. Unlike leaders, citizens do not typically give political speeches. Analyzing their voices thus requires a novel data collection strategy. Third, unlike English Mandarin is a tonal language that highly constrains the use of pitch for expressivity. To address this, we go beyond pitch and employ a richer set of acoustic metrics, borrowed from the phonetics literature. Fourth, our design is comparative rather than based on a single country and institutional setting. Fifth, we are able to use not only individual but also sentence fixed effects, as different individuals may be assigned to read the exact same sentence. This allows for very tight internal validity. Essentially, we hold constant the numerous sentence characteristics that are hard to control for when comparing speech across sentences.

3 Data

We start by describing the construction of our main dataset. We then explain the acoustic metrics—tempo, volume and pitch—which form our main dependent variables. Finally, we describe how we code the political contents of sentences, which forms our main explanatory variable. Table A.1 provides descriptive statistics.

3.1 The Common Voice Project speech corpus

We use data from the November 2024 release of the Common Voice Corpus (Mozilla 2024), hereafterCVC. Launched in 2017, the CVC project aims to build the largest open-source, multilingual dataset of voice recordings for machine-learning applications in speech technol-

ogy (e.g., automatic speech recognition). While the CVC is generally a language rather than country-based dataset, it includes separate datasets for Mainland Chinese and Taiwanese speakers, the former relying on simplified written Mandarin and the latter on Traditional Mandarin.

The CVC is crowdsourced: contributors (i) record sentences in voice clips, (ii) submit additional sentences, and (iii) validate (i) and (ii). Most volunteers only record voice clips and most sentences are scraped from mainstream internet sources, especially Wikipedia.[2] Each of these voice clips is recorded as a separate MP3 file. Each MP3 has an ID that matches to the ID of each sentence/speaker couple. When speakers sign up to donate their voice, they are asked to report their vocal gender (male-masculine, female-feminine, nonbinary or transgender voice), their age (binned in intervals of ten years), and their accent. For accent, individuals can either click a regional option or write a short summary of their vocal influences.

The CVC is structured around standard ML subsets (train, unvalidated, test, development, other, and validated). The validated subset consists of voice clips and sentences that have been reviewed and confirmed by at least two volunteers as accurate transcriptions—meaning that the voice matches the sentences, avoiding people playing with the dataset and recording something else. We rely on this large, higher quality subset. The raw validated dataset covers, respectively, 186,502 and 83,118 voice clip recordings, of 50,676 and 17,204 sentences, by 5,582 and 1,732 Mainland Chinese and Taiwanese speakers.

We clean the dataset in four steps. First, we identify and remove poor quality recordings that escaped users' validation.[3] Second, we identify and drop complex IDs from individuals that record their voice in both the Chinese and Taiwanese dataset.[4] Third, we remove

[2]Due to sentence donation being insufficient, most sentences are scraped from Wikipedia and added to the sentence collector according to a specific protocol discussed in a dedicated github page here. After agreeing to donate their voice at commonvoice.mozilla.org, volunteers get random sentences on their screen and read them out aloud (see e.g. A.1).

[3]We can calculate the output variables for respectively 99.98%, 99.99% and 100% of MP3s. For 537 entries, either pitch or volume is missing. For 83 entries, both volume and pitch metric are missing. In those cases, the file is basically silent. It still has a tempo, because it has a beginning and an end (i.e. the MP3 is not ″corrupted″). Utterances that could be ascribed a volume but not a pitch are typically covered in noise (listen e.g. zh-CN_33489767) while in the opposite case the speaker is either whispering or recording from too far (listen e.g zh-CN_32981377). We remove rows with missing volume and/or pitch, typically of low quality, and our sample slightly decreases to 7,256 unique IDs and 269,083 utterances.

[4]This exclusion removes 53 IDs for a total of 10,302 entries that appear in both the Mainland Chinese and Taiwanese corpora, since individuals could in principle record their voice in multiple datasets, and a very small number of them did. This could reflect e.g. the fact that the Chinese corpus was first published in 6/11/2019, while the Taiwanese dataset began earlier in 2/25/2019, and hence some Mainland China speakers wanting to donate their voice had nowhere to go for a few months, or simply wanting to engage with multiple variations of written Mandarin. After this second cut, we have 7,203 unique individuals and 258,781 recorded MP3s.

one-word utterances and ensure that identical sentences are given the same ID.[5]

Our effective sample includes 239,576 voice clips from 6,114 Mandarin speakers covering 61,468 unique sentences. Of these, 5,582 Mainland Chinese speakers recorded 186,502 voice clips reading 50,676 sentences, while 1,732 Taiwanese speakers recorded 83,118 voice clips reading 17,204 sentences, 2,398 of which were assigned to both Mainland Chinese and Taiwanese speakers. We are able to compute the key acoustic measures for over 99% of all voice clips. Recordings started in 2019, however we have no information about the exact date of each recording.

3.2 Acoustic metrics

Speech is produced in two steps. Air is pushed out from the lungs and sets the vocal folds into vibration, generating sound waves—patterns of pressure variation in the surrounding air. Coordinated movements of the mouth and nasal cavity then shape these waves into recognizable speech sounds. These sound waves vary along three key dimensions: tempo, volume, and pitch. Tempo reflects how speech unfolds over time and captures the rate at which sounds are produced. Volume reflects the strength of pressure variations in the sound wave, and pitch reflects how fast the vocal folds vibrate. Our acoustic metrics are derived using the `parselmouth` library in Python (Jadoul, Thompson, and De Boer 2018), which is based on Praat, the standard phonetics software available at http://www.fon.hum.uva.nl/praat/. Before processing, each MP3 file is denoised using a specific Python package (`noisereduce`) and converted into a format Praat can read.

Tempo

"Tempo" is a shorter name for "speech rate", given by

$$Tempo = \frac{\text{number of characters in a sentence } (c)}{\text{duration of speech in seconds } (d)} \equiv \rho_{d,c}. \tag{1}$$

Higher tempo means a quicker reading of a sentence. We measure c using all characters excluding punctuation, and d by calculating the total length of the MP3. As d may include some silence before and after the speech starts and ends, we also report analyses using related

[5]This was not always the case in the original dataset due to minor spelling differences. Cutting sentences with one word drops 19,205 entries, leaving us with 239,576 entries for 6,114 individuals reading 64,157 unique sentences, with 2,407 of them assigned at least once to Mainland Chinese and Taiwanese speakers. One last cleaning reveals that some "unique" sentences only differ to others because of a space before or after the sentence, a comma, or a slightly different translation from simplified and traditional Mandarin to English. Examples include "I am good" vs "am good"; "The two *parties* are not in a good mood" vs "The two *sides* are not in a good mood" (emphasis added). After taking this into account we end up with 61,468 unique sentences, with 2,398 shared across the Mainland Chinese and Taiwanese "sentence collectors".

metrics such as time until the recording passes a certain threshold of volume or pitch (note that identifying precisely when speech begins and ends is not straightforward, as background noise can be difficult to distinguish from speech; see appendix B.1 for details). This allows us to explore whether differences in tempo may partly reflect differences in recording behavior, where silences are shorter because people begin speaking sooner after starting the recording or stop the recording more quickly after finishing certain sentences.

An individual with higher cognitive abilities or stronger respiratory muscles is, all else equal, able to read at a faster tempo. A similar point holds for sentences: for a given number of characters, a more complex sentence may take slightly longer to cognitively process, especially if including subordinate clauses, resulting in slower tempo. However, once we control for individual and sentence-level characteristics, tempo is an important marker of agency: speakers change their tempo to a greater degree than other acoustic metrics in response to emotional cues (Anolli et al. 2008, Brenner, Branscomb, and Schwartz 1979, Chnag, Young, and Yuen 2010, Dietrich and Abbott 2012).

Figure 1 shows the distribution across countries and genders. While men and women in our data talk at similar speeds, speakers from mainland China tend talk faster overall. This will be important when constructing our empirical strategy.

Volume

"Volume" captures the average strength of the speech signal within a sentence. It reflects the magnitude of pressure variations in the sound wave and is measured using the root mean square (RMS) amplitude of the wave.[6] Because sound intensities vary over a very large range, volume is measured in decibels (dB) which compresses this range using logarithms. Specifically, in any analysis window, we compute:

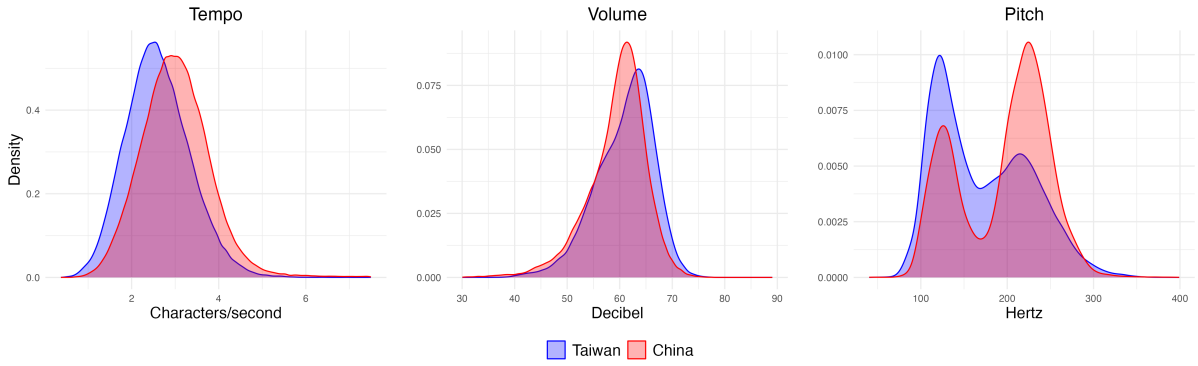
$$Volume = 20 \times \log_{10}\left(\frac{\text{RMS amplitude}}{\text{reference amplitude}}\right). \tag{2}$$

A whisper is around 30 dB, normal conversation around 60 dB, and shouting can reach 90 dB. Using default parameters, each analysis window is 6 milliseconds long. We exclude unvoiced parts from the calculation to avoid biasing the number down when the share of unvoiced part grows, and use standard ceilings to distinguish voice from noise, considering alternative settings in appendix (B.2). Sentence recordings in our data have an average of 60

[6]Specifically, let $p(t)$ be the sound pressure (in Pascals) at time t, measured as the deviation from ambient atmospheric pressure. Since $p(t)$ oscillates above and below zero—positive during compression and negative during rarefaction—its time average is approximately zero and cannot serve as a measure of magnitude. The *RMS amplitude* is then defined as $\sqrt{\frac{1}{T}\int_0^T p(t)^2\,dt}$ where T is the duration of the signal.

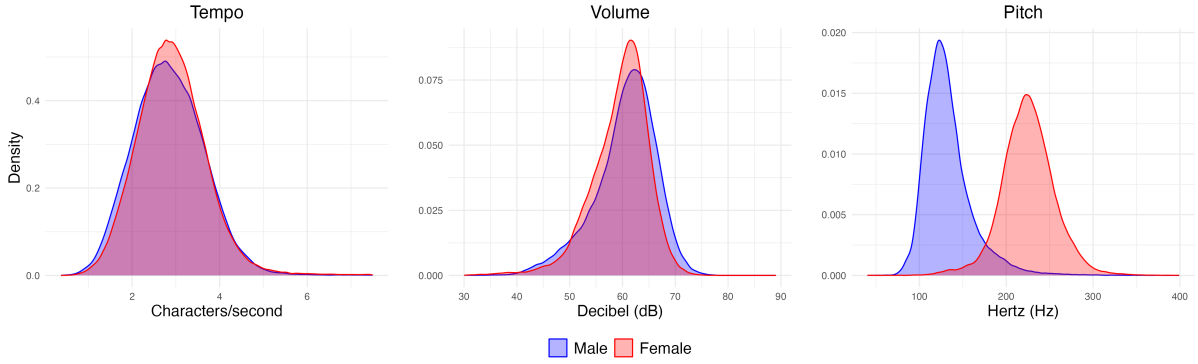
Figure 1: Speech Acoustics by Country and Gender

(a) China vs Taiwan



N. id: 6218 , N. unique sentences: 62424

(b) Male vs Female



N. id: 6194 , N. unique sentences: 62374

dB, displaying low variance as a result of the logarithmic transformation (see Appendix A). As Figure 1 illustrates, Taiwanese participants in our data talk slightly louder, as do men.

Volume is thought to be less affected by individual than by sentence-level characteristics. A particular combination of e.g. vocal fold thickness and head shape can make people sound louder, but sentence-level characteristics may play a greater role. At the micro level, this is because the extent to which we articulate phonemes by opening and closing the mouth and using nasal resonance has a large impact on the air pressure we release. At the macro level, this is because reading specific sentences clearly may or may not require stressing parts of the sentence by raising the volume.[7]

[7]Different languages affect how much the micro or macro levels induce volume levels. In Mandarin, sentences containing open vowel sounds such as /a/ in 爸 ("dad") tend to have a higher RMS amplitude than closed vowel sounds such as /i/ in 你 ("you"). The same applies to unaspirated stops such as /p/, /t/, and /k/, which exhibit lower intensity relative to aspirated stops as well as sonorants like /m/, /n/, and /l/. In English, sentences entailing open vowel sounds such as /a/ in "cat" lead to higher RMS amplitude than closed vowel sounds such as /i/ in "knee" . Hence, two sentences with similar contents, emotional

Pitch

Pitch is the frequency of vocal fold vibrations producing sound waves in the surrounding air and determining how "low" or "high" a person sounds when voicing. Denoting by T the time elapsed between two successive peaks of a sound wave, we have

$$Pitch = \frac{1}{T}. \tag{3}$$

Thus, faster vibrations produce a higher pitch. These vibrations are measured in Hertz (Hz). A "fundamental frequency"of 100 Hz captures 100 complete vocal folds vibrations per second. In our sample, men's average pitch is 133 Hz (SD=28.5), corresponding to $\♫ = C\sharp 3$, while women's is 9 semitones more acute, averaging 225 Hz (SD: 30.8), see Figure 1. This is inline with extant research (e.g., Chen, Huang, and Wu 2016). This number of vibrations is associated by our brain with a pitch. Pitch is measured using Praat′ s "fundamental frequency" ($f0$) algorithm. For each sentence, Praat estimates $f0$ at 10-millisecond intervals, producing a time series of pitch values. Each estimate is based on an analysis window of 60 milliseconds centered on the time step. We filter implausibly low and high $f0$ values using 40 and 400Hz as lower and upper bounds.[8]

Pitch is strongly affected by individual-level biological and sociodemographic characteristics. The minimum and maximum number of complete vocal folds vibrations per second one can choose to produce—one's "vocal range" —mostly depends on the size and tension of one's vocal folds, itself correlated with one's body type and strongly affected by one's gender and, to a lesser extent, age (Graddol and Swann 1983). The range subset people feel naturally comfortable with in expressing their intonation is considerably smaller. Sentence characteristics, including words and syntactic sentence characteristics, influence pitch. The extent to which pitch must preserve the lexical meaning of single words or the syntactic function of entire sentences depends on the language. In Mandarin, a tonal language, words are composed of meaningful monosyllables characterized by one out of five available tones. Hence, speaker pitch in Mandarin is highly constrained to preserve lexical meaning. In English, a non-tonal language, speakers are freer to alter their intonation to stress a particular word or convey the nature of a sentence (e.g., declarative vs interrogative).[9]

underpinning, and intonation can display different volume.

[8]Praat's default settings used e.g. in Dietrich, Hayes, and O′ brien (2019) range from 75Hz (floor for men) to 500 (ceiling for women). These seem appropriate for spontaneous speech and especially public speech which naturally shows much greater pitch variability, but for quiet reading they seem too high—men reading can sometimes use vocal fry, which they would not in a public speech, delving well below 75Hz, while a pitch of 500Hz is unusually high for a woman reading a sentence. That said, we later relax these assumptions in our robustness checks

[9]For instance, the distinction between a declarative and an interrogative sentence is primarily marked by

3.3 Coding political sentences

We use GPT, validated by human native speakers, to code sentences as either related to politics ($Political = 1$) or unrelated to politics ($Political = 0$). In the appendix we also use continuous measures. For our baseline analysis we employ the following procedure. We use GPT 4-o, and set the temperature to 0 to ensure conservative, deterministic and hence replicable coding. The prompt is designed to be very broad, so as to capture any sentence related to politics and not just sensitive topics:

> "Label each sentence as 1 if it is related to politics and 0 if it is unrelated. A sentence is related to politics if it refers to political institutions (for example parties, the government, politicians, the police, judicial court, and so on), political events (historical or current), political debates (on domestic or foreign policy, on economic, social, or cultural issues) and political beliefs (including ideology, public opinion, ethical viewpoints)" .

All coding is performed on the original text in Mandarin. Figure 2 shows a series of word clouds to visualize the words that appear in sentences coded as political versus non political, translated to English and using the entire dataset (use the playback buttons to move between word clouds). In appendix B.3, we report results using alternative (more or less specific) prompts. We also report results using an alternative GPT model and alternative levels of determinism in coding a sentence as political. Appendix Table B.3 shows the complete list of political sentences (according to the different coding schemes) in the subset of sentences used in our main specifications, i.e. those sentences assigned in both countries.

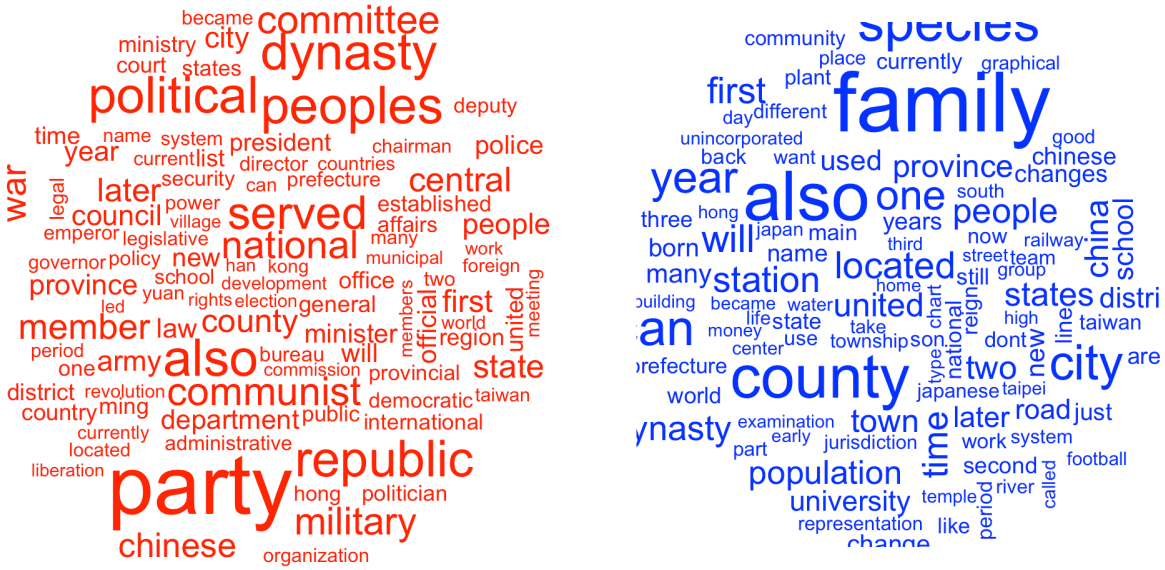
To validate the accuracy of our baseline coding, two native Mandarin speaking research assistants coded manually and independently a 10% sample of the sentences, that includes 25% political sentences. Accuracy rates (the fraction of correctly classified observations) is respectively 93.53 % and 87.23% for the two RAs. Recall rates (share of actual positives recovered) are 69.25% and 83.80%. In the overall sample, political sentences are 10.04% of the total. See Appendix A.1 for details.

3.4 Addressing imbalances in sentence characteristics

As mentioned above, a major challenge in using speech to characterize attitudes norms and behavioral tendencies, is that many sentence characteristics may directly affect acoustic metrics, regardless of the lexical contents of the sentence. For example, some sentences may be more difficult to read, naturally reducing tempo. If the political sentences in our sample

swapping verb position and varying intonation in English, while it is marked by adding a question particle at the end of the sentence in Mandarin, with a much less marked increase in pitch.

Figure 2: Political (left,red) and Non-Political (right,blue) Sentences



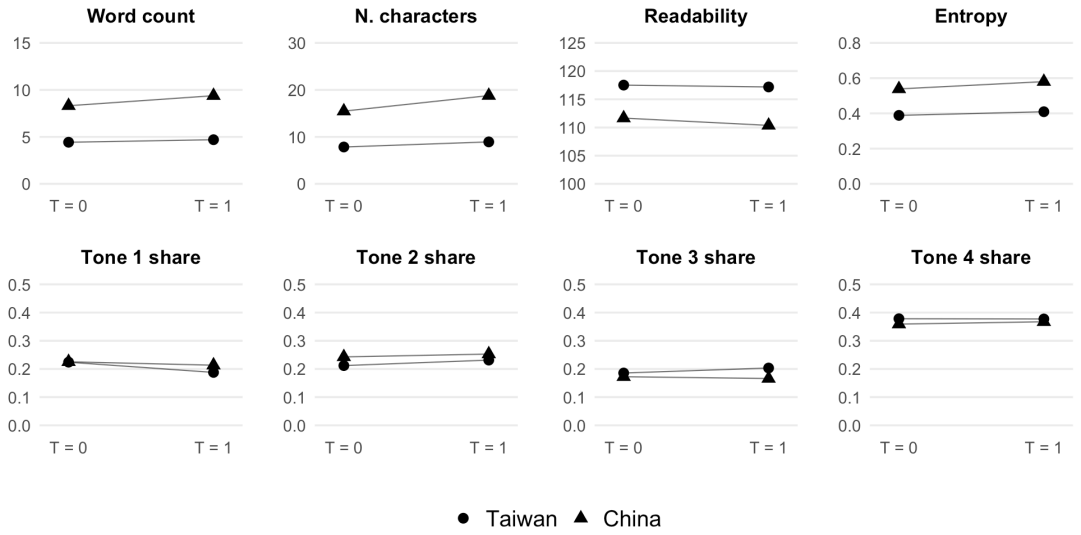
*Note:*Word clouds generated using the wordcloud package in R based on English translations of sentences in our dataset. Word size reflects frequency of occurrence, with larger words appearing more frequently in the respective subset. The left panel displays words from political sentences, while the right panel shows words from non-political sentences. Text preprocessing included conversion to lowercase, removal of punctuation, numbers, and common English stop words. Each cloud displays the 100 most frequent terms.

happen to be simpler than the non-political sentences, the former may display higher tempo even in the absence of a political treatment effect. Similarly, preserving the meaning of a sentence may place requirements on the intonation (especially in tonal languages), affecting pitch.

Figure 3 demonstrates this point empirically using the full sample of sentences in our data. For each country and sentence type—political (T=1) versus non-political (T=0)—we compute the sample average across unique sentences of a range of metrics computable from the written sentences alone. These include word and character counts, readability scores, and the share of the different Mandarin tones. As it turns out, Taiwanese speakers are assigned shorter and simpler sentences, on average. This may be because, as already mentioned, the sentence collector is country-specific. Furthermore, in both countries political sentences in our data tend to be somewhat longer and more semantically complex, e.g. longer by approximately one and a half words (average is 6.9) or four characters (average is 12.7). The imbalance in the prevalence of specific Mandarin tones is modest in magnitude both across countries and sentence type. (appendix B runs a balance test, reporting magnitudes and

coefficients).

Figure 3: Sentence Characteristics by Contents and Country



Notes. 'Word count' is simply the number of words in the original Mandarin sentence. 'Number of characters' is also self-explanatory. 'Readability' is proxied by the Flesch Reading Ease (FRE) index, a length–based metric combining average sentence length and average syllables per word: $\text{FRE} = 206.835 - 1.015\,\frac{\text{words}}{\text{sentences}} - 84.6\,\frac{\text{syllables}}{\text{words}}$. Here, "words" and "sentences" are computed via our segmenter; syllables are obtained by standard syllabification for the language. FRE ranges from 35.5 (hard) to 121.2 (easy). To interpret it as a *complexity* measure aligned with entropy (higher = more complex), we use the monotone transform $C_{\text{FRE}} = 100 - \text{FRE}$ when needed. 'Entropy' is another metric for sentence complexity based on the within–sentence token distribution. For a sentence s with token types $\{t_1, \ldots, t_K\}$, counts n_k, total $N = \sum_{k=1}^{K} n_k$, and empirical probabilities $p_k = n_k/N$, the entropy is $H(s) = -\sum_{k=1}^{K} p_k \log p_k$ with log base 2. We report the normalized entropy to make values comparable across sentences of differing vocabulary size, $H^*(s) = \frac{H(s)}{\log K} \in [0, 1]$. Higher values indicate a more uniform token distribution and greater lexical dispersion—hence greater complexity. Tokens are computed at the sentence level using our tokenizer (characters or words, as appropriate). As one can imagine, these first four metrics of sentence complexity are extremely correlated. Within the sample of common sentences between China and Taiwan, for example, the weakest correlation is that between number of characters and words and still is .86. For each sentence s, we segment the text into Mandarin syllables and assign each syllable a lexical tone $t \in \{1, 2, 3, 4\}$ (high level, rising, dipping, falling) from the transcription (Pinyin with diacritics or numeric tags). Let n_t be the count of syllables with tone t, and let $N = \sum_{t=1}^{4} n_t$ be the total number of tone–bearing syllables with tones 1–4. The share for tone t is $\text{share}_t(s) = n_t/N$. Neutral–tone syllables, punctuation and non-Chinese tokens (which are very sporadically included) are excluded from N (a separate neutral-tone share can be reported if needed).

Of course, the specific imbalances documented in Figure 3 can be controlled for in the analysis. However, other sentence characteristics may also be imbalanced, and fully controlling for all of them is not feasible. Yet such disparities could affect tempo, volume, or pitch. Two sentences with approximately the same length and topic classification may nonetheless have different speech acoustics not due to the speaker's agency and emotions, but because different words and syntax imply different tones.

To address this challenge, we exploit the fact that 2,398 unique sentences (out of 61,468

sentences in our data) were assigned to *both* Mainland Chinese and Taiwanese speakers. A sentence in this subset of identically-spelled sentences—in simplified or traditional characters—is assigned on average approximately 15 times to 2,292 individuals (of which, 1,538 in China). This yields a total of 33,919 MP3s that can be analyzed after controlling for both individual and sentence fixed effects to achieve the highest internal validity.

Within this subset, 30-50 sentences are classified as "political" by GPT, depending on GPT's temperature, prompt, or model.[10] See Appendix Table B.3 for the complete list. The vast majority of these sentences concern ordinary political topics, whether 'political' is defined broadly or narrowly. They seldom touch on sensitive issues. Instead, they focus on routine political content. Moreover, they often lack any context, and as a result can hardly be interpreted asymmetrically by Mainland Chinese and Taiwanese speakers (though we allow for this possibility). In additional analysis, we further sub-group political sentences according to what they express (opinion or fact), the level of the policy domain they refer to (foreign or domestic), and the tone (negative or neutral). We also perform a robustness check using a continuous variable to capture the political content of sentences, ranging between 0 and 1 and rising with the political sensitivity of the question (B.2). See appendix A for a full list of sentences and how they are coded in our main approach, according to some alternative prompts, and in manual labelling.

4 Empirical Strategy

Our empirical strategy seeks to identify the acoustic effects of political institutions on political speech. Specifically, consider the following equation, estimated for individual i reading sentence s:

$$y_{is} = \beta_0 + \beta_1 Political_s + \beta_2 China_i + \beta_3 Political_s \times China_i + \mu_i + \gamma_s + \epsilon_{ics} \tag{4}$$

where y_{is} is an acoustic metric (tempo, volume, or pitch); $Political_s$ is an indicator for a political sentence; and $China_i$ is an indicator for whether respondent i is from Mainland China ($China_i = 1$) or Taiwan ($China_i = 0$). The full specification also includes individual and sentence fixed effects (μ_i,γ_s) discussed later. ϵ_{ics} is an error term clustered at the sentence level.

Equation 4 allows for two types of acoustic differences that do not necessarily indicate an institutional acoustic effect. First, as noted above, it is possible that the political sentences in

[10]Political classification is consistent across Simplified and Traditional Chinese scripts for all but three sentences, which received different classifications depending on the script. We classify a sentence as "political" (treatment) only if it is coded as political in both scripts.

our sample have different (observed or unobserved) characteristics compared to non-political sentences, and that these differences affect how the sentences are spoken regardless of their political contents. Thus, β_1 may be nonzero even in the absence of an institutional acoustic effect. Second, it is possible that Chinese speakers from the mainland speak faster or louder than those from Taiwan for reasons unrelated to the different institutions. In this case, β_2 will be nonzero even in the absence of institutional acoustic effects. Our main interest is in β_3, which captures the differential effect of political contents in China versus Taiwan.[11]

While this baseline difference-in-differences specification allows for systematic cross-country differences, one might still be concerned that unobserved individual heterogeneity biases β_3. Speakers differ in baseline voice characteristics—such as natural speaking rate, vocal intensity, or articulatory habits—that may correlate with both country of origin and the share of political sentences they happen to record. To address this, the full specification also includes individual fixed effects μ_i. This captures any biological, socioeconomic, or other characteristics that affect person i's voice and manner of speaking.

More importantly, systematic differences in sentence characteristics can confound the comparison between political and non-political sentences. As documented above (Figure 3), political sentences in our data tend to be longer and more complex, and Taiwanese speakers are assigned shorter sentences on average. Such characteristics can directly affect acoustics—for example, longer sentences may mechanically reduce tempo, and specific tone sequences constrain pitch contours. Without sentence-level controls, β_3 could reflect such compositional differences rather than an institutional effect. To address this possibility, we include sentence fixed effects γ_s. These capture any sentence-specific characteristics that can affect the way a sentence is spoken, from the tonal requirements of a given syllable to the complexity of the sentence, and from the lexical meaning of single words to the syntactic function of the whole sentence.

Our identification rests on a parallel response assumption: absent institutional differences, Mainland Chinese and Taiwanese speakers would display similar acoustic responses to political versus non-political sentences. Note that Differential self-selection into the corpus likely works against finding an effect: individuals who volunteer to record their voice on a public platform in authoritarian China are plausibly less sensitive to political content than the general population, biasing our estimates toward zero. Finally, Appendix Table B.9 shows that observable speaker characteristics (including age, gender, and the total number

[11] This coefficient captures a difference-in-differences (conditional on included controls). To see this note that:

$$E(y|\textit{Political Sentence, China, controls}) - E(y|\textit{Non-political Sentence, China, controls}) = \beta_1 + \beta_3$$

$$E(y|\textit{Political Sentence, Taiwan, controls}) - E(y|\textit{Non-political Sentence, Taiwan, controls}) = \beta_1.$$

of sentences recorded) do not predict the assignment of political sentences in either country, consistent with effectively random sentence assignment.

5 Results

Table 1 reports our main findings. For each acoustic outcome — tempo (Panel A), volume (Panel B), and pitch (Panel C) — we present three specifications: a baseline OLS difference-in-differences, a model adding individual fixed effects, and our preferred specification with both individual and sentence fixed effects. Results are shown separately for men and women. Throughout, we refer to speakers who identify as men or as having a masculine voice as "men," and analogously for "women." In the preferred specification, individual fixed effects absorb $China_i$ and sentence fixed effects absorb $Political_s$, so the DiD coefficient β_3 is identified from variation that is within-individual and within-sentence.

Tempo. Panel A reports estimates with tempo (characters per second) as the dependent variable. In the Basic DiD for men (column 1), political sentences carry only a marginally higher baseline tempo than non-political ones ($\hat{\beta}_1 = 0.031$, SE = 0.018). Interestingly, however, Mainland Chinese speakers are substantially faster than Taiwanese at baseline ($\hat{\beta}_2 = 0.281$, SE = 0.007). The DiD coefficient is positive and significant at $p < .01$ ($\hat{\beta}_3 = 0.246$, SE = 0.025): Chinese men speed up on political content relative to Taiwanese men, compounding an already elevated speaking pace. Adding individual fixed effects in column (2) absorbs the China coefficient and reduces the DiD estimate by roughly one-third to 0.167 (SE = 0.022), still significant at $p < .01$. Individual fixed effects control for any time-invariant speaker characteristics — biological, socioeconomic, or related to recording equipment — that may correlate with both country of origin and acoustic delivery. The coefficient's decline suggests that part of the baseline DiD reflects compositional differences across speakers rather than a within-person response to political content.

The preferred "All FEs" specification in column (3) adds sentence fixed effects, restricting identification to sentences read by speakers in both countries. This is critical because, as documented in Section 3, political sentences tend to be longer and more complex, and such characteristics can directly affect tempo independently of content. The effective sample drops to 14,967 observations (770 individuals, 2,392 sentences) after dropping singletons — individuals or sentences that appear only once and contribute no within-group variation. Speaking about politics increases tempo among Chinese men by 0.201 characters per second (SE = 0.072, $p < .01$), approximately 7.3% of mean tempo and roughly 0.2 standard deviations. This corresponds to a reduction in recording duration of about 0.2 seconds, confirmed

Table 1: Speaking fast and low

	Men/Masculine voice			Women/Feminine voice		
	(1) Basic DiD	(2) Individual FE	(3) All FEs	(4) Basic DiD	(5) Individual FE	(6) All FEs
Panel A. Tempo (c/s)						
Political × China	.246***	.167***	.201***	.296***	.156***	-.065
	(.025)	(.022)	(.072)	(.031)	(.026)	(.075)
Political	.031*	.085***		-.040*	.101***	
	(.018)	(.016)		(.024)	(.019)	
China	.281***			.238***		
	(.007)			(.009)		
Mean dep. var.	2.762	2.762	2.756	2.811	2.811	2.746
N. obs	96,919	96,567	14,967	141,985	141,732	18,125
N. sentence	51,276	51,079	2,392	32,583	32,483	2,327
N. id	3,772	3,420	770	2,337	2,084	721
Panel B. Volume (dB)						
Political × China	.081	-.014	-2.270**	-.578***	-.245***	-.360
	(.149)	(.090)	(.892)	(.135)	(.085)	(.435)
Political	-.803***	-.641***		-.112	-.348***	
	(.122)	(.074)		(.121)	(.075)	
China	-1.707***			-.805***		
	(.042)			(.036)		
Mean dep. var.	60.560	60.560	61.737	59.584	59.584	60.473
N. obs	96,919	96,567	14,967	141,985	141,732	18,125
N. sentence	51,276	51,079	2,392	32,583	32,483	2,327
N. id	3,772	3,420	770	2,337	2,084	721
Panel C. Pitch (Hz)						
Political × China	2.742***	.584	4.733**	5.671***	2.880***	4.599
	(.679)	(.431)	(2.280)	(.991)	(.666)	(5.052)
Political	-2.503***	-.594		-6.710***	-3.586***	
	(.582)	(.377)		(.942)	(.633)	
China	-6.578***			.547**		
	(.196)			(.277)		
Mean dep. var.	133.187	133.187	136.550	225.172	225.172	230.361
N. obs	96,919	96,567	14,967	141,985	141,732	18,125
N. sentence	51,276	51,079	2,392	32,583	32,483	2,327
N. id	3,772	3,420	770	2,337	2,084	721
Individual FE	no	yes	yes	no	yes	yes
Sentence FE	no	no	yes	no	no	yes

Notes. Columns (1) and (4) present basic difference-in-differences specifications with controls for "Political" and "China" indicators. Columns (2) and (5) add individual fixed effects, absorbing time-invariant speaker characteristics—including whether speakers recorded as Mainland Chinese or Taiwanese, thus making "China" collinear. Columns (3) and (6) include both individual and sentence fixed effects, restricting the sample to sentences read by both Mainland Chinese and Taiwanese speakers. Stata's `reghdfe` command automatically drops singleton observations, which explains the reduction in observations and unique individuals/sentences in columns (3) and (6). Standard errors clustered at the sentence level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

when using MP3 length as the dependent variable.

For women (columns 4–6), the Basic DiD and Individual FE specifications produce estimates qualitatively similar to those for men. However, once sentence fixed effects are introduced in column (6), the DiD coefficient collapses to -0.065 (SE = 0.075) and loses statistical significance. Given that men are at higher risk of political repression, this gender contrast — a strong effect for men, a null for women — provides the first indication that our results may reflect a response to the fear of repression.

Volume. Panel B examines volume, measured in decibels. The pattern across specifications differs markedly from tempo and provides the clearest illustration of why sentence fixed effects matter. In the Basic DiD (column 1), political sentences are generally spoken more quietly ($\hat{\beta}_1 = -0.803$, SE = 0.122), and Mainland Chinese speakers are already substantially quieter at baseline ($\hat{\beta}_2 = -1.707$, SE = 0.042). Crucially, the DiD coefficient for men is small and statistically insignificant (0.081, SE = 0.149), while for women it is negative and significant (-0.578, SE = 0.135). Without further controls, one would conclude that Chinese women — not men — lower their volume on political content. This pattern survives the introduction of individual fixed effects in columns (2) and (5).

The preferred specification overturns this picture. Once sentence fixed effects absorb the systematic differences in content between political and non-political sentences — including phoneme composition, sentence length, and tonal structure, all of which directly affect measured volume — the coefficient for women shrinks to -0.360 (SE = 0.435) and loses statistical significance, while the coefficient for men becomes large, negative, and significant (-2.270, SE = 0.892, $p < .05$). This illustrates how sentence composition can generate misleading results: without comparing like-for-like content, the naive estimates point in the wrong direction for both genders. The column (3) estimate implies that Chinese men lower their volume by 2.3 dB when reading political sentences, approximately 3.7% of mean volume. Because decibels are logarithmic and volume has low cross-individual dispersion, this corresponds to roughly 0.4 standard deviations — a larger standardized effect than for tempo.

Pitch. Panel C reports estimates for pitch, measured in Hertz. In the Basic DiD (column 1), political sentences are associated with lower pitch for both genders, and Chinese men have substantially lower baseline pitch than Taiwanese men ($\hat{\beta}_2 = -6.578$, SE = 0.196). The DiD coefficient is positive and significant (2.742, SE = 0.679), but adding individual fixed effects in column (2) reduces it to 0.584 and it loses statistical significance. In the preferred specification (column 3), political content raises pitch among Chinese men by 4.7 Hz (SE = 2.280, $p < .05$), roughly 3.5% of mean pitch or 0.15 standard deviations — corresponding

Table 2: UK and US English speakers

	Men			Women		
	Tempo (c/s)	Volume (dB)	Pitch (Hz)	Tempo (c/s)	Volume (dB)	Pitch (Hz)
Political × US	-.041	.101	.679	.036	-.063	.163
	(.049)	(.115)	(.638)	(.125)	(.251)	(1.872)
Mean dep. var.	9.322	58.283	125.168	9.121	56.402	201.721
F-stat	.687	.773	1.133	.082	.063	.008
N. obs	151,262	151,262	151,187	35,123	35,123	35,114
N. sentence	16,231	16,231	16,231	8,377	8,377	8,377
N. id	3,608	3,608	3,608	914	914	913
Sentence FE	yes	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The specification is the same as the preferred "All FEs" specification in Table 1. The treatment indicator is constructed by replacing 'Chinese and Taiwanese politics' with 'British and American politics' in the prompt. Since the English-language dataset is much larger, we do not use gender imputation for missing data.

to slightly less than a semitone, moving the pitch from $\♫ = C\sharp 3$ to $\♫ = D3$. However, as we show in the robustness analysis (Section B), this result is not significant in most alternative specifications. We therefore do not emphasize it. While pitch is commonly associated with emotional arousal and stress in non-tonal languages (Section 2), in tonal languages such as Mandarin pitch contours carry lexical meaning, leaving speakers less freedom to vary pitch for emotional expression. Mandarin speakers may instead channel stress through tempo and volume.

The combined pattern — faster and quieter political speech among men, with no comparable effect among women — is consistent with a gendered response to the fear of repression. By including both individual and sentence fixed effects, the preferred specification absorbs all systematic variation linked to speaker- and text-level characteristics. The estimated differences therefore cannot be attributed to speakers' voice or recording equipment, nor to imbalance in sentence content, syntax, or tonal structure. Two additional tests further support this interpretation.

First, if speaking fast and low about politics reflects internalized institutional practices rather than linguistic drift, we should not observe the pattern when comparing two countries that share a language and similar levels of press freedom. We test this by replicating our analysis for US and UK English speakers, two countries scoring 83/100 and 91/100, respectively, on the 2024 Freedom House index. Table 2 reports the results. All Political × US coefficients are small and statistically insignificant, for both genders and all three outcomes.

Second, we implement a randomization (permutation) test. We randomly assign the

"political" label to 30 sentences in each of 1,000 replications (thus holding the share of treated sentences fixed) and re-estimate the preferred specification for each acoustic outcome each time. This addresses the concern that some unobserved sentence characteristic, rather than political content per se, might drive our results. However, beyond checking whether our specific political label matters, this effectively tests whether *any* arbitrary grouping of sentences of the same size could generate a comparable DiD estimate. Figure 4 plots the resulting distribution of DiD coefficients alongside our actual estimate. For tempo, only 1.2% of permuted estimates exceed the actual coefficient. For volume, 99.8% of estimates are larger (i.e., less negative). These results indicate that our effects are unlikely to arise from any arbitrary partition of sentences. In contrast, for pitch, 11.5% of randomized estimates exceed ours, suggesting that the pitch result is not reliably distinguishable from chance.

5.1 Robustness

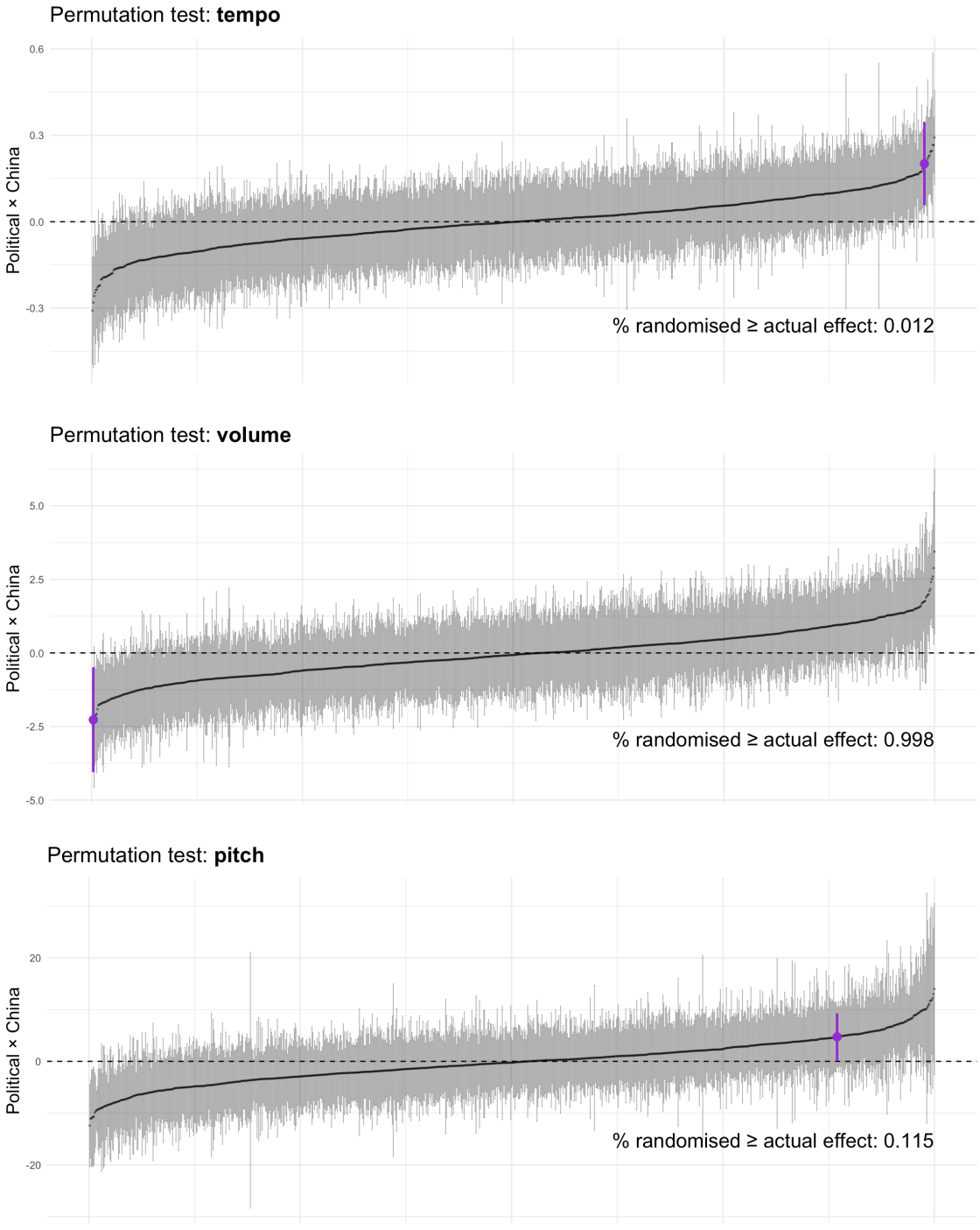
We now summarize the robustness of our findings to alternative definitions of the dependent variables, of the treatment, and to alternative specifications. Full results are reported in the appendix.

Dependent variables. Appendix B.1 re-estimates equation 4 using alternative measures of tempo, volume, and pitch. These include the log and standardized values of each dependent variable (used to inform the effect sizes reported in Section 5), the median rather than the mean of the within-clip pitch and volume, and tighter frequency and decibel bounds for recording valid pitch and volume entries. Across all alternatives, the DiD estimates for tempo and volume remain significant at least at the $p < .05$ level. By contrast, the DiD estimates for pitch are statistically insignificant in many specifications.

Treatment. Appendix B.2 tests whether our results depend on how sentences are classified as political. We consider two alternative GPT prompts (one that omits examples, another that leaves the definition of "politics"to the model), as well as a continuous treatment that assigns higher scores to sentences more directly relevant to contemporary Chinese or Taiwanese politics. We also raise the labeling temperature from 0 to 1, re-label all sentences using GPT 4o mini instead of 4-o, and allow the treatment indicator to differ across countries. Across all alternatives, the DiD estimate for tempo remains significant at least at the $p < .05$ level, and the estimate for volume at least at the $p < .10$ level. The DiD estimates for pitch are mostly statistically insignificant.

Specifications. Appendix B.3 tests robustness to alternative *sampling* restrictions (retaining only individuals assigned at least one political sentence, or only those reporting a Mainland Chinese or Taiwanese accent), *clustering* of standard errors (by individual, or by

Figure 4: Randomization test



Notes. We implement a randomization (permutation) test for the coefficient on Political × China. First, we estimate the model $Y \sim \text{Political} : \text{China}$ with two-way fixed effects (by id and sentence). Next, for 1,000 replications we randomly select exactly 30 sentences, set $\text{Political}_{\text{random}} = 1$ for those and 0 otherwise, re-estimate the same specification replacing the actual with the random political treatment, and store the point estimate and its 95% cluster-robust CI. The figure orders the permuted estimates (points with vertical CIs) and overlays the actual estimate and CI in violet; the dashed line marks zero. The percentage printed in the lower-right reports the share of permuted estimates $\geq$ the actual estimate (a one-sided randomization p-value). The procedure is repeated separately for each outcome (tempo, volume, pitch); main effects are omitted because they are absorbed by the fixed effects.

both individual and sentence), and controlling for the order of the sentence assigned to each speaker. The DiD estimates for tempo and volume remain significant at least at the $p < .05$ level across all alternatives. In this set of checks, the DiD estimates for pitch are also mostly significant, though we do not emphasize this result given its instability across the other robustness dimensions.

In sum, the finding that Chinese men speak faster and more quietly when reading political sentences is robust across all alternative dependent variables, treatment definitions, and specifications we consider. The estimated effect on pitch is not robust.

6 Mechanisms

6.1 Fear of repression

The most plausible interpretation of our findings is that they capture mild fear or stress triggered by political content. Several lines of evidence support this reading.

First, the acoustic pattern matches what phonetics research predicts for fear in Mandarin. Studies of emotional expression in Mandarin consistently link mild fear and stress to faster tempo and lower volume (Wang, Lee, and Ma 2018, Chnag, Young, and Yuen 2010, Anolli et al. 2008, Li and Huang 2023). In a cross-cultural comparison of Mandarin and English speakers, Wang, Lee, and Ma (2018) find that fear is associated with the fastest speech rate among all emotions tested, along with reduced volume. Because pitch carries lexical meaning in a tonal language such as Mandarin, it cannot be freely recruited for emotional expression (Wang, Lee, and Ma 2018), which explains why we observe robust effects on tempo and volume but not pitch. A meta-analysis of stress and vocal production (Giddens et al. 2013) confirms that stress generally increases tempo, though its effect on volume depends on arousal level: high arousal raises vocal intensity, whereas mild stress reduces it. Our findings—faster tempo combined with lower volume—are thus consistent with low-arousal stress rather than panic. We confirm this directly: we find no significant changes in jitter or shimmer, voice quality markers typically associated with high-arousal distress (see Appendix B).

Second, the gendered pattern in our acoustic results mirrors the gendered pattern of political repression in China. Speaking fast and low is concentrated among men (Table 1). Political repression in China has been predominantly male-targeted. During the Cultural Revolution (1966--71), approximately 70% of those killed were men. In the contemporary period, men constitute the large majority of political prisoners (Global Voices 2018). If the acoustic effect reflected a mechanism unrelated to repression—for example, general cultural

norms about avoiding political discussion—we would expect the effect among both genders. The fact that the gender asymmetry in acoustics tracks the gender asymmetry in repression is consistent with fear of repression as the underlying mechanism.

The placebo results are also consistent with this interpretation. Comparing British and American English speakers—who share a language but face no comparable institutional variation in political freedom—yields null results across all outcomes and for both genders (Table 2).

Together, these findings point to fear of political repression as the likely mechanism underlying speaking fast and low. The remaining question is what *kind* of fear drives the effect: a deliberate response to current surveillance, or an internalized legacy of historical repression?

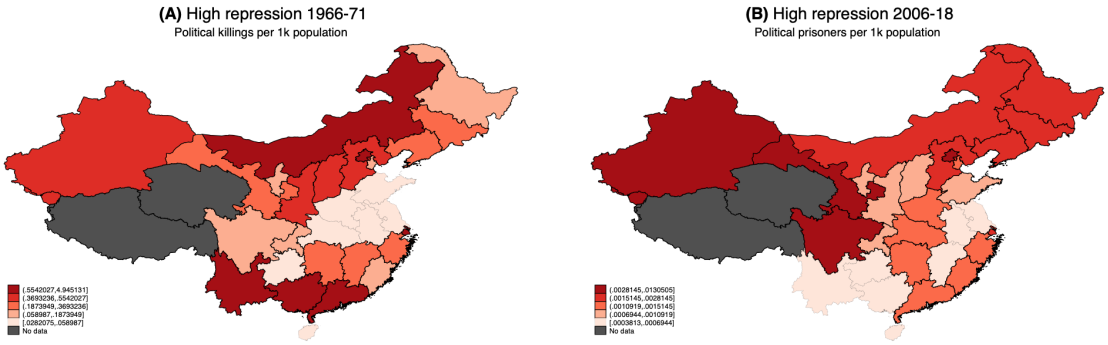
6.2 Current caution or internalized legacy?

Two distinct mechanisms could generate fear-driven acoustic patterns, and they yield different empirical predictions. The first can be referred to as *caution* given current threats. That is, speakers rationally adapt their vocal behavior to the present risk of persecution. This predicts that the effect should be strongest in provinces where the state currently imprisons the most political dissidents, and among younger cohorts, who are over-represented among contemporary political prisoners (Global Voices 2018). A second possibility, however is an *internalized legacy* mechanism. According to this account, vocal habits were shaped by historical repression and transmitted culturally through family socialization, institutional memory, and local norms. Because the effect we find appears even for routine, factual political sentences recorded in a private setting, it is arguably too pervasive to reflect only deliberate caution. The legacy account predicts that the effect should be strongest in provinces where the state killed the most during the Cultural Revolution, and among older cohorts who carry that legacy more directly. It may also be weaker in provinces that were economically liberalized early.

These mechanisms are not mutually exclusive. Exercising caution when talking about politics is a plausible response to current repression, regardless of internalized norms. However, our data allows us to explore which of these mechanisms better explains our results.

Geographic heterogeneity. To distinguish between these channels, we exploit the fact that the geographic pattern of repression differs across the two periods. Panel (A) of Figure 5 shows provincial variation in political killings per 1,000 population in 1966–71, during the Cultural Revolution, based on archival data compiled by Walder (2014) and made public by Wang (2021). This period saw mass factional violence, purges, and locally organized

Figure 5: Political Repression, historically and now



campaigns in which ordinary citizens participated in identifying, denouncing, and often executing alleged political enemies. The repression was both ideologically driven and intensely localized, resulting in sharp variation across provinces. Panel (B) shows provincial variation in political prisoners per 1,000 population between 2006 and 2018, based on the China Political Prisoner Dataset compiled by the US Congressional-Executive Commission on China, as employed by Xu (2021).[12] Provincial rates of historical and current repression are only modestly correlated ($\rho = .215$, $p < .01$), providing scope to separate their effects.

To perform this analysis, we focus on the subsample of Mainland Chinese men who reported an accent, from which we can infer their province of origin (see Appendix A.1).[13] We then replicate the baseline DiD specification within Mainland China, comparing speakers from provinces with above- versus below-median levels of historical or current repression. Because we restrict to Mainland Chinese men who reported an accent, and to sentences assigned in both high- and low-repression provinces, the sample shrinks to approximately 1,000 recordings.

The results in the first two columns of Table 3 favor the legacy account. Column 1 shows that men from provinces with historically higher repression speak at significantly faster tempo when reading political rather than non-political sentences (coefficient: 0.203, SE: 0.095), adding roughly 7% over the unconditional mean of 2.93. In contrast, current repression does not predict speaking fast and low. The tempo coefficient is negative and

[12]The Congressional-Executive Commission on China's Political Prisoner Database (CPPD) is widely regarded as the most comprehensive public record of China's political prisoners, documenting over 10,000 cases since the early 1980s. Nonetheless, the dataset is prone to systematic under-reporting biases: well-known urban dissidents are more likely to appear than marginalized or less-connected individuals, and many victims of repression are held in extrajudicial confinement and may escape inclusion in prison records (Steinert 2024).

[13]Speakers typically report their province of birth when self-reporting their accent, so this coding entails little measurement error. See Appendix A.2 for details. Of course people's speaking practices might have changed over time as a result of e.g. internal migration; still, those practices would have most likely been shaped in the coded province, and speakers are on average young.

insignificant (column 2). The point estimate on volume is negative for historical repression and positive for contemporary repression, but both are very noisy and neither estimate is statistically different from zero.

Table 3: The geography of speaking fast and low

	High repression 1966-71	High repression 2006-18	Special zone	Close to Beijing
A: Tempo (Hz)				
Political × Area	.203**	-.177	-.159*	-.288
	(.095)	(.122)	(.091)	(.196)
Output mean	2.928	2.845	2.808	2.887
F-stat	4.599	2.113	3.044	2.159
N. obs	985	985	1,020	1,007
N. sentence	488	488	505	499
N. id	99	99	103	102
B: Volume (dB)				
Political × Area	-.506	.557	.861	.942
	(.889)	(.821)	(.955)	(1.203)
Output mean	59.878	60.983	60.204	60.595
F-stat	.325	.461	.812	.613
N. obs	985	985	1,020	1,007
N. sentence	488	488	505	499
N. id	99	99	103	102
Common modelling assumptions				
Sentence FE	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. All geographic subsets are based on respondents reporting their accent. In the second column, we use data from Andy Walder (2014) on Cultural Revolution violence for the last colum. Provinces with higher than average number of deaths per thousand in 1966-1971 are considered here. In the third column, we use data from Xu (2021). 'Special economic zones' include respondents that reported accents from SEZ provinces, which were areas granted preferential policies—such as tax incentives and regulatory exemptions—to attract foreign investment and promote export-oriented growth. In China, the SEZs established were Shenzhen, Zhuhai, Shantou, and Xiamen in 1980, Hainan in 1988, and Pudong (Shanghai) in 1990. 'Close to Beijing' means lower than effective sample median distance in km from Beijing. The third column 'High repression 2006-2018' uses data from Xu (2020), originally released by the Congressional-Executive Commission on China (CECC), established in 2000 by the U.S. Congress to monitor human rights and the rule of law in the People's Republic of China which releases the most detailed, publicly accessible records of detained and imprisoned individuals in China.

Column 3 provides another angle on the possibility that historical legacies leave traces in speech patterns. Speakers from Special Economic Zones—areas integrated into global markets since the 1980s and culturally closer to the West—display significantly *slower* tempo on political sentences compared to other provinces (−0.159, SE: 0.091). This is consistent with economic liberalization and sustained international contact reducing the salience of political

threats. Distance from Beijing, which might be associated with current state monitoring and political control, does not affect the results (column 4).

Cohort heterogeneity. If speaking fast and low reflects rational caution in response to current surveillance, we would expect it to be most pronounced among younger cohorts, who are disproportionately targeted today. In contrast, the internalized-legacy account suggests the effect should be stronger among older adults. Table 4 replicates the analysis from Table zzz, (for men, full specification) subsetting by age cohort. Columns 1–2 report the estimates for all men who provided a valid age. The results replicate those for the full sample: tempo is 0.2 c/s faster, and volume is about 2.5 dB lower for men in China reading political sentences. Columns 3–4 restrict the sample to young men (ages 15–30) and columns 5–6 to older men (ages 30 and above). The effect of political content on both tempo and volume is driven by the older men. For younger men the effects are smaller and statistically insignificant.[14]

Table 4: Speaking Fast and Low across generations

	All men		Young men (15-30)		Adult men (30+)	
	Tempo (c/s)	Volume (dB)	Tempo (c/s)	Volume (dB)	Tempo (c/s)	Volume (dB)
Political × China	.207**	-2.517**	.036	-1.568	.318**	-3.995***
	(.095)	(1.013)	(.087)	(.954)	(.137)	(1.537)
Output mean	2.796	61.641	2.820	62.431	2.766	60.675
F-stat	4.736	6.170	.174	2.699	5.369	6.752
N. obs	11,901	11,901	6,463	6,463	5,438	5,438
N. sentence	2,400	2,400	2,293	2,293	2,330	2,330
N. id	549	549	394	394	155	155
Sentence FE	yes	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes	yes

Notes. Standard errors in parentheses. For the first part of the analysis (all men) we only use respondents that provided a valid age. As such, the total sample in the first and second colums corresponds to the sum of the effective samples in columns (3) and (5) or (4) and (6). To emphasize this equivalence, we exceptionally retain singletons, which reghfe by default drops. If we didn't do that, then the effective samples in columns (3) plus (5) or (4) plus (6) would not exactly match those in the first two columns, because a higher number of simpletons is dropped when the sample is split. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Two important caveats. First, there are many reasons why different cohorts might vary in their response to political prompts, besides their exposure to historical and current repression. Second, our age comparison is limited. Only 4.5% of respondents with valid age information are 40 or above, and only 5.5% are teenagers. We are thus mostly comparing speakers in their twenties with speakers in their thirties. These results should be viewed as suggestive and exploratory: theory-building rather than theory-testing.

[14]That said, the volume point estimate among younger men is negative, suggesting that even newer generations may internalize the tendency to some degree.

Sentence type Our central hypothesis is that Mainland Chinese speakers display some distinctive speech acoustics when prompted to talk about politics. While our sentences capture mostly statements about ordinary, low stake political life (see appendix A for a full list), we can explore whether heterogeneity within the political domain mediates our estimated effects. Doing so reveals that speaking fast and low is more pronounced for sentences including political opinions, rather than facts; sentences with a negative tone rather than neutral; and sentences related to foreign policy. See Appendix B.2 for details.

Acoustic Legacy Hypothesis Three pieces of evidence—the gendered pattern mirroring gendered repression, the role of historical rather than current repression, and the age gradient—together support what we term the *acoustic legacy hypothesis*: vocal patterns reflect an internalized legacy of fear, embedded in speaking habits, and not just a rational response to current surveillance. The concept of "habitus" (Bourdieu 1977, Bourdieu 1990) captures this logic: individuals internalize dispositions from their family, schooling, and broader social environment, and then deploy them unconsciously. In this sense, a history of political repression can shape vocal behavior even absent an immediate threat of punishment. Importantly, we are not arguing against rational caution as a mechanism. We are saying that the evidence also points to an internalized component that cannot be reduced to deliberate risk calculation alone.

There are both strengths and limitations to this analysis. On the one hand, we maintain high internal validity by retaining sentence and individual fixed effects in all mechanism analyses within China. Moreover, while the province is not the most granular level of analysis, an ANOVA decomposition reveals that between-province variation in Cultural Revolution death rates accounts for approximately 42% of total variation, nearly as large as within-province variation across prefectures.[15] On the other hand, the mechanism sample is much smaller because many speakers do not report their accent ($N \approx 1{,}000$). Furthermore, because China does not release official data on repression, both the historical and current repression measures are imperfect proxies. While the sign and magnitude of the DiD estimates are stable across alternative grouping criteria (median, mean, or quartile splits), statistical significance varies with the cutoff used. We discuss and rule out several alternative explanations related to gender differences in socialization, attrition, and biology in Appendix B.5.

[15]The between-province component (model sum of squares = 290.13) explains about 42% of the total variation in the share of deaths per 1,000 by prefecture. The F-statistic of 6.44 ($p < 0.001$) confirms that provincial means differ systematically.

7 Conclusion

Using speech corpora intended for AI speech-to-text recognition thus escaping authoritarian control, we show that when exogenously assigned sentences with political content in a private online environment, Mandarin speakers from Mainland China change the acoustics of their speech relative to Mandarin speakers from Taiwan. DiD estimates, controlling for both sentence and individual fixed effects. We find that when prompted with mundane political sentences, male speakers in Mainland China consistently speak faster and more quietly than comparable speakers in Taiwan reading the exact same sentences, while women show no systematic change. We interpret this as evidence that authoritarianism is not only written in laws, memories, and institutions—it is also spoken, in hurried and hushed tones, every time politics comes up.

This is the first paper showing that authoritarianism causally affects how people speak, delivering insights for some conversations among political scientists interested in authoritarianism from political theory, gender, and methodological perspectives.

First, our findings highlights the limits of political discussion in authoritarian regimes. Authoritarian regimes lack the most obvious competitive mechanism disclosing and aggregating citizens' policy preferences. One remedy envisaged by the Chinese Communist Party and taken seriously by scholars consists in consulting citizens at the local or federal level through a suitable combination of online and offline consultations about their preferences (He and Thøgersen 2010, He and Warren 2011). A necessary condition for authoritarian regimes to limit their epistemic gap and gain a deliberative character is that, across such consultations, they manage to maximize constructive and minimize contentious political discussion—a task complexified by the fact that citizens' risk aversion with respect to the ambiguity of the government with respect to what can and cannot be said may lead to generalized self-censorship (Truex 2017, Kim et al. 2022, Teets 2024). That Mainland Chinese citizens are uncomfortable even to talk about mundane politics in a safe environment suggests such necessary condition may not be met: citizens may self-censor even for *constructive* political discussion, hence consultations promoted by the Chinese Communist Party are unlikely to work.

Second, self-censorship on mundane political discussion is gendered. Our mechanism hint at fear of repression, rather than generic socialization patterns stressing assertiveness or accommodating communication among boys and girls, and contribute to a growing literature seeking to understand gendered patterns in authoritarian context (Su 2006, Elsayed et al. 2025, Fan, Pan, and Zhang 2025). The authoritarian acoustics are stronger in provinces that experienced intense repression during the Cultural Revolution, rather than in those facing high contemporary repression, suggesting that the effect reflects internalized habits rather

than rational caution (Bourdieu 1977, Phelps 1972). Hence, in seeking to explain gendered patterns, we develop a new acoustic legacy hypothesis that can be of interest for scholars in historical political economy who are particularly interested in measures of deep-rooted cultural variation. Case studies drawing on vocal data from regions who share the language but experienced distinct historical trajectories of political cultures, such as Eastern and Western Germany, could further contextualize how embodied political behavior responds to variation in institutional memory.

Finally, we are the first to document that ordinary voice recordings—collected for unrelated technological purposes—carry the traces of political regimes. This may be methodologically appealing. As countries become increasingly reluctant to participate to social surveys, speech corpora offer a new window into how political institutions affect human expression. Future research could explore similar patterns across diasporic communities, where linguistic habits may persist or shift based on local institutions and exposure to surveillance. Comparative studies in post-authoritarian settings—such as Eastern Europe or Latin America—or in formerly colonized regions like Algeria or South Korea, could test whether similar vocal traces of repression endure.

References

Alesina, Alberto, Paola Giuliano, and Nathan Nunn (2013). "On the Origins of Gender Roles: Women and the Plough." In: *Quarterly Journal of Economics* 128.2, pp. 469–530.

Anderson, Rindy C. and Casey A. Klofstad (2018). "Voice pitch predicts electability, but does not signal leadership ability." In: *Evolution and Human Behavior* 39.3, pp. 147–153. DOI: 10.1016/j.evolhumbehav.2018.02.007.

Anolli, Luigi et al. (2008). "The voice of emotion in Chinese and Italian young adults." In: *Journal of Cross-Cultural Psychology* 39.5, pp. 565–598.

Baevski, Alexei et al. (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations." In: *Advances in neural information processing systems* 33, pp. 12449–12460.

Banai, Irena Pavela, Benjamin Banai, and Kosta Bovan (2017). "Vocal characteristics of presidential candidates can predict the outcome of actual elections." In: *Evolution and Human Behavior* 38.3, pp. 309–314.

Bisin, Alberto and Thierry Verdier (2001). "The Economics of Cultural Transmission and the Dynamics of Preferences." In: *Journal of Economic Theory* 97.2, pp. 298–319.

Bourdieu, Pierre (1977). *Outline of a Theory of Practice*. Trans. Richard Nice. Cambridge: Cambridge University Press.

Bourdieu, Pierre (1990). *The Logic of Practice*. Trans. Richard Nice. Stanford: Stanford University Press.

Boussalis, Constantine et al. (2021). "Gender, candidate emotional expression, and voter reactions during televised debates." In: *American Political Science Review* 115.4, pp. 1242–1257.

Brenner, Malcolm, Harvie H Branscomb, and Gary E Schwartz (1979). "Psychological stress evaluator—two tests of a vocal measure." In: *Psychophysiology* 16.4, pp. 351–357.

Carter, Erin Baggott, Brett L Carter, and Stephen Schick (2024). "Do Chinese citizens conceal opposition to the CCP in surveys? Evidence from two experiments." In: *The China Quarterly* 259, pp. 804–813.

Chen, Jidong and Yiqing Xu (2017). "Why do authoritarian regimes allow citizens to voice opinions publicly?" In: *The Journal of Politics* 79.3, pp. 792–803.

Chen, Jing, Qiang Huang, and Xihong Wu (2016). "Frequency importance function of the speech intelligibility index for Mandarin Chinese." In: *Speech communication* 83, pp. 94–103.

Chnag, HS, Shuenn-Tsong Young, and Kevin Yuen (2010). "Effects of the Acoustic Characteristics on the Emotional Tones of Voice of Mandarin Tones." In: *20th International Congress on Acoustics, ICA*. Vol. 2010.

Cinar, Asli Ceren and Özgür Kıbrıs (2024). "Persistence of voice pitch bias against policy differences." In: *Political Science Research and Methods* 12.3, pp. 563–580. DOI: 10.1017/psrm.2023.51.

Dietrich, Bryce J, Matthew Hayes, and Diana Z O′ brien (2019). "Pitch perfect: Vocal pitch and the emotional intensity of congressional speech." In: *American Political Science Review* 113.4, pp. 941–962.

Dietrich, Maria and Katherine Verdolini Abbott (2012). "Vocal function in introverts and extraverts during a psychological stress reactivity protocol." In.

Distelhorst, Greg and Yue Hou (2017). "Constituency service under nondemocratic rule: Evidence from China." In: *The Journal of Politics* 79.3, pp. 1024–1040.

Elsayed, Nourhan A et al. (2025). *The gendered persistence of authoritarian indoctrination*.

Fan, Yingjie, Jennifer Pan, and Tongtong Zhang (2025). "Gender and Political Compliance Under Authoritarian Rule." In: *Comparative Political Studies* 58.9, pp. 1890–1929.

Fouquet, Meddy et al. (2016). "Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood." In: *Royal Society open science* 3.10, p. 160395.

Freedom House (2024). *Country and Territory Scores*. URL: https://freedomhouse.org/country/scores?sort=fiw&sort_order=asc (visited on 02/26/2026).

Giddens, Cheryl L et al. (2013). "Vocal indices of stress: a review." In: *Journal of voice* 27.3, 390–e21.

Giuliano, Paola and Nathan Nunn (2021). "Understanding Cultural Persistence and Change." In: *Review of Economic Studies* 88.4, pp. 1541–1581.

Global Voices (June 2, 2018). *Who Are China′ s Political Prisoners? A Human Rights Assessment, 29 Years After Tiananmen*. URL: https://globalvoices.org/2018/06/02/who-are-chinas-political-prisoners-a-human-rights-assessment-29-years-after-tiananmen/ (visited on 02/26/2026).

Graddol, David and Joan Swann (1983). "Speaking fundamental frequency: Some physical and social correlates." In: *Language and Speech* 26.4, pp. 351–366.

He, Baogang and Stig Thøgersen (2010). "Giving the people a voice? Experiments with consultative authoritarian institutions in China." In: *Journal of contemporary China* 19.66, pp. 675–692.

He, Baogang and Hendrik Wagenaar (2018). "Authoritarian deliberation revisited." In: *Japanese Journal of Political Science* 19.4, pp. 622–629.

He, Baogang and Mark E Warren (2011). "Authoritarian deliberation: The deliberative turn in Chinese political development." In: *Perspectives on politics* 9.2, pp. 269–289.

Hirschman, Albert O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press.

Jadoul, Yannick, Bill Thompson, and Bart De Boer (2018). "Introducing parselmouth: A python interface to praat." In: *Journal of Phonetics* 71, pp. 1–15.

Jiang, Jue (2021). "The eyes and ears of the authoritarian regime: Mass reporting in China." In: *Journal of Contemporary Asia* 51.5, pp. 828–847.

Jiang, Junyan and Dali L Yang (2016). "Lying or believing? Measuring preference falsification from a political purge in China." In: *Comparative Political Studies* 49.5, pp. 600–634.

Jiang, Xinhui and Yunyun Zhou (2022). "Coalition-based gender lobbying: Revisiting women's substantive representation in China's authoritarian governance." In: *Politics & Gender* 18.4, pp. 978–1010.

Kim, Jieun et al. (2022). "At your own risk: A model of delegation with ambiguous guidelines." In: *Journal of Political Institutions and Political Economy* 2.4, pp. 483–508.

King, Gary, Jennifer Pan, and Margaret E. Roberts (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression." In: *American Political Science Review* 107.2, pp. 326–343.

Kloet, Jeroen de and Thomas Poell (2025). "Why the everyday is essential: Navigating censorship and surveillance on WeChat in China." In: *Media, Culture & Society*, p. 01634437251380415.

Klofstad, Casey A. (2016). "Candidate voice pitch influences election outcomes: voice pitch influences voters." In: *Political Psychology* 37.5, pp. 725–738. DOI: 10.1111/pops.12280.

Klofstad, Casey A (2017). "Looks and sounds like a winner: Perceptions of competence in candidates′ faces and voices influences vote choice." In: *Journal of Experimental Political Science* 4.3, pp. 229–240.

Klofstad, Casey A., Rindy C. Anderson, and Stephen Nowicki (2015). "Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices." In: *PLoS ONE* 10.8, e0133779. DOI: 10.1371/journal.pone.0133779.

Kuran, Timur (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.

Lei, Xuchuan and Jie Lu (2017). "Revisiting political wariness in China′ s public opinion surveys: Experimental evidence on responses to politically sensitive questions." In: *Journal of Contemporary China* 26.104, pp. 213–232.

Lennes, Mietta et al. (2015). "Comparing pitch distributions using Praat and R." In: *Phonetician* 111-112, pp. 35–53.

Li, Junlin and Chu-Ren Huang (2023). "Investigating Acoustic Cues of Emotional Valence in Mandarin Speech Prosody-A Corpus Approach." In: *Workshop on Chinese Lexical Semantics*. Springer, pp. 316–330.

Lindh, Jonas (2017). "Forensic Comparison of Voices, Speech and Speakers." In: *Journal of the Royal Statistical Society. Series C, Applied Statistics* 53.1, pp. 109–122.

MacFarquhar, Roderick and Michael Schoenhals (2006). *Mao′ s last revolution*. Harvard University Press.

Mayew, William J, Christopher A Parsons, and Mohan Venkatachalam (2013). "Voice pitch and the labor market success of male chief executive officers." In: *Evolution and Human Behavior* 34.4, pp. 243–248.

Mozilla (2024). *Common Voice Corpus 19.0*. Version 17.0. URL: https://commonvoice.mozilla.org/en/datasets (visited on 02/26/2026).

Nunn, Nathan (2012). "Culture and the Historical Process." In: *Economic History of Developing Regions* 27.sup1, S108–S126.

Pan, Jennifer and Yiqing Xu (2018). "China′ s ideological spectrum." In: *The Journal of Politics* 80.1, pp. 254–273.

Phelps, Edmund S. (Sept. 1972). "The Statistical Theory of Racism and Sexism." In: *American Economic Review* 62.4, pp. 659–661.

Qin, Bei, David Strömberg, and Yanhui Wu (2017). "Why does China allow freer social media? Protests versus surveillance and propaganda." In: *Journal of Economic Perspectives* 31.1, pp. 117–140.

Reisinger, William M., Marina Zaloznaya, and Byung-Deuk Woo (2023). "Fear of punishment as a driver of survey misreporting and item non-response in Russia and its neighbours." In: *Post-Soviet Affairs* 39.1-2, pp. 49–59. DOI: 10.1080/1060586X.2022.2150490.

Rozenas, Arturas and Yuri M. Zhukov (2019). "Mass repression and political loyalty: Evidence from Stalin′ s ′terror by hunger′." In: *American Political Science Review* 113.2, pp. 569–583.

Searles, Kathleen et al. (2020). "The effects of men's and women's voices in political advertising." In: *Political Communication* 37.4, pp. 518–538. DOI: 10.1080/10584609.2019.1670903.

Shen, Xiaoxiao and Rory Truex (2021). "In search of self-censorship." In: *British Journal of Political Science* 51.4, pp. 1672–1684.

Steinert, Christoph Valentin (2024). "Elitist remedies? Complaint resources and representation in international human rights bodies." In: *International Studies Quarterly* 68.2, sqae042.

Su, Fubing (2006). "Gender inequality in Chinese politics: an empirical analysis of provincial elites." In: *Politics & Gender* 2.2, pp. 143–163.

Svolik, Milan W (2012). *The politics of authoritarian rule*. Cambridge University Press.

Teets, Jessica C. (2024). "The Fragility of Authoritarian Legitimacy under Xi Jinping: Citizen Feedback and Regime Responsiveness in China." In: *The China Quarterly* 263, pp. 597–620. DOI: 10.1017/S0305741024000096.

Truex, Rory (2017). "Consultative authoritarianism and its limits." In: *Comparative political studies* 50.3, pp. 329–361.

Wagner, Kevin M, Jason Gainous, and Jason P Abbott (2021). "Gender differences in critical digital political engagement in China: The consequences for protest attitudes." In: *Social Science Computer Review* 39.2, pp. 211–225.

Walder, Andrew G (2014). "Rebellion and repression in China, 1966–1971." In: *Social Science History* 38.3-4, pp. 513–539.

Wang, Ting, Yong-cheol Lee, and Qiuwu Ma (2018). "Within and across-language comparison of vocal emotions in Mandarin and English." In: *Applied Sciences* 8.12, p. 2629.

Wang, Yuhua (2021). "The political legacy of violence during China's Cultural Revolution." In: *British Journal of Political Science* 51.2, pp. 463–487.

Wedeen, Lisa (2022). *Ambiguities of domination: Politics, rhetoric, and symbols in contemporary Syria*. University of Chicago Press.

Wong, Stan Hok-Wui and Jiachen Liang (2023). "Attraction or distraction? Impacts of pro-regime social media comments on Chinese netizens." In: *Political behavior* 45.3, pp. 1071–1095.

Xu, Xu (2021). "To repress or to co-opt? Authoritarian control in the age of digital surveillance." In: *American Journal of Political Science* 65.2, pp. 309–325.

Yang, Byunggon et al. (2024). "The fundamental frequency (f0) distribution of American speakers in a spontaneous speech corpus." In: *Phonetics and Speech Sciences* 16.1, pp. 11–16.

Yang, Tony Zirui (2025). "Participatory censorship in authoritarian regimes." In: *Comparative political studies* 58.12, pp. 2594–2620.

Yuen, Samson and Francis LF Lee (2025). "Echoes of silence: how social influence fosters self-censorship under democratic backsliding." In: *Democratization*, pp. 1–26.

Zhang, Sheng, PC Ching, and Fanrang Kong (2006). "Acoustic analysis of emotional speech in Mandarin Chinese." In: *International symposium on chinese spoken language processing*, pp. 57–66.

Appendix

List of Tables

1 18
2 20
3 26
4 27
A.1 Descriptive statistics, individual level A.5
A.2 Accents coding A.6
B.1 Robustness: Dependent variable definition B.14
B.2 Speaking fast or clicking fast? B.15
B.3 Political Sentences in Main Specification B.16
B.4 Covariate balance for sentence characteristics B.17
B.5 Robustness: Treatment definition B.18
B.6 Robustness: by sub-treatment B.19
B.7 Robustness: specifications B.20
B.8 Voice quality: Jitter and shimmer B.21
B.9 Randomization of treatment B.22

List of Figures

1 Speech Acoustics by Country and Gender 10
2 13
3 14
4 22
5 25
A.1 Voice clips recording A.7
A.2 Pitch and gender A.7
B.1 Attrition by gender and area B.13

A Data construction and Descriptive Statistics

A.1 Dataset collection and cleaning

We use data from the Mozilla Common Voice corpus (CVC). Launched in 2017, the project aims to build the largest open-source, multilingual dataset of voice recordings for machine-learning applications in speech technology (e.g., automatic speech recognition). We use the latest available release as of November 2024 (CVC 19.0), downloaded under the CC0 1.0. While the CVC is generally a language rather than country-based dataset, it includes separate datasets for Mainland Chinese and Taiwanese speakers, the former relying on simplified written Mandarin and the latter on Traditional Mandarin.

The CVC is crowdsourced: contributors (i) record sentences in voice clips, (ii) submit additional sentences, and (iii) validate (i) and (ii). Most volunteers only record voice clips and most sentences are scraped from mainstream internet sources, especially Wikipedia, and added to the sentence collector according to a specific protocol discussed in a dedicated github page here. Sentences from a particular collector can be assigned to both the Mainland Chinese and Taiwanese speakers. After agreeing to donate their voice at commonvoice.mozilla.org, volunteers get random sentences on their screen and read them out aloud (see e.g. A.1). Each of these voice clips is recorded as separate MP3. Each MP3 has an ID that matches to the ID of each sentence/speaker couple. Speakers can stop recording voice clips at any point in time. When they sign up to donate their voice, they are asked to report their vocal gender (male-masculine, female-feminine, nonbinary or transgender voice), their age (binned in intervals of ten years) as well as their accent (individuals can either click a regional option or write a short summary of their vocal influences).

The CVC is structured around standard subsets (train, unvalidated, test, dev, other, and validated). The validated subset consists of voice clips and sentences that have been reviewed and confirmed by at least two additional volunteers beyond the speakeras accurate transcriptions—meaning that the voice matches the sentences, avoiding people playing with the dataset and recording something else. We rely on this large subset. After performing some id, sentence and MP3 level data cleaning, we have an effective sample of 239,576 voice clips for 6,114 individuals recording 64,157 sentences.

The raw dataset covers 186,502 and 83,118 voice clips recording respectively 50,676 and 17,204 sentences by respectively 5,582 and 1,732 Mainland Chinese and Taiwanese speakers. First, we can calculate pitch, tempo and volume for respectively 99.98%, 99.99% and 100% of MP3. For 537 entries, either pitch or volume is not recorded. For 83 entries, both volume and pitch metric are missing. In those cases, the file is basically silent. It still has a tempo, because it has a beginning and an end (i.e. the MP3 is not "corrupted"). Utterances

that could be ascribed a volume but not a pitch are typically covered in noise (listen e.g. zh-CN_33489767) while in the opposite case the speaker is either whispering or recording from too far (listen e.g zh-CN_32981377). We remove rows with missing volume and/or pitch, typically of low quality, and our sample slightly decreases to 7,256 unique IDs and 269,083 utterances. Second, we identify and drop 53 IDs for a total of 10,302 entries that appear in both the Mainland Chinese and Taiwanese corpora, since individuals could in principle record their voice in multiple datasets, and a very small number of them did. This could reflect e.g. the fact that the Chinese corpus was first published in 6/11/2019, while the Taiwanese dataset began earlier in 2/25/2019, and hence some Mainland China speakers wanting to donate their voice had nowhere to go for a few months, or simply wanting to engage with multiple variations of written Mandarin. After this second cut, we have 7,203 unique individuals and 258,781 recorded MP3. Third we focus on sentences, not one word utterances which, as we later show, are never "treatment sentences". Cutting sentences with one word drops 19,205 entries, leaving us with 239,576 entries for 6,114 individuals reading 64,157 unique sentences, with 2,407 of them assigned at least once to Mainland Chinese and Taiwanese speakers. One last cleaning reveals that some "unique" sentences only differ to others because of a space before or after the sentence, a comma, or a slightly different translation from simplified and traditional Mandarin to English. Examples include "I am good" vs "am good"; "The two *parties* are not in a good mood" vs "The two *sides* are not in a good mood" (emphasis added). After taking this into account we end up with 61,468 unique sentences, with 2,403 shared across the Mainland Chinese and Taiwanese "sentence collectors".

Table A.1 provides the descriptive statistics for each of the key dependent and independent variables discussed in the paper and here below.

Table A.2 instead shows how we created geographic variables based on people self-reported accent, which often names the place of birth.

A.2 Imputation

Pitch is highly gendered as figure A.2 shows. Our pitch descriptives are in line with extant work showing with very moderate overlap and small cross-country and cross-age differences in gender pitch (Yang et al. 2024, Fouquet et al. 2016, Lennes et al. 2015, Lindh 2017). That pitch is highly gendered allows us to confidently impute missing gender information. The simplest criterion would impute as male every speaker whose average pitch across all spoken sentences is lower than the average pitch of the lowest female pitch decile and as female every speaker whose average pitch across all spoken sentences is higher than the

average pitch of the highest male pitch decile—moving the ceiling would allow trading off imputation coverage vs precision. A more sophisticated criterion builds on a version of the Wav2Vec 2.0 deep learning model developed by Meta AI (Baevski et al. 2020), pre-trained on large cross-lingual datasets and fine-tuned for gender recognition by a voice gender recognition scholar (available at https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech). The graph in the right panel of figure A.2 shows how Wav2Vec 2.0 performs. Recall that each speaker reads on average 35 sentences, recorded in 35 separate MP3. The deep learning model assigns vocal gender—0 for male and 1 for females—to each of these MP3 files separately. First, we take the average vocal gender by speaker id: 0 (1) means that each MP3 was assigned "male" ("female") while 0.1 means that 90% MP3s were assigned "male" and 10% "female" by the deep learning model. Figure A.2 shows that, in respectively 86% and 91% of the cases, 100% of the on average 35 MP3s of actual respectively female and male speakers are indeed imputed as female. The quick decay of the bars means that it is only in exceptional cases that someone who reports one gender is often assigned a different gender by the deep learning model. Impute as "male" and "female" those speakers for whom at least 2/3 of the MP3s are assigned respectively "male" and "female" gender results in an accuracy of 94.15% with respect to the part of the sample for which we have actual gender data. Such accuracy is of course lower when we consider teenagers, whose voice may have not fully developed, meaning that boys can be misgendered.

Table A.1: Descriptive statistics, individual level

	China male/masculine voice			China female/feminine voice			Taiwan male/masculine voice			Taiwan female/feminine voice			All	
	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs	Min	Max
Main Acoustic Metrics														
Tempo (characters/second)	2.67	1.19	58262	2.67	1.29	123852	2.60	0.75	45592	2.61	0.71	30947	0.10	62.50
Volume (dB, mean)	59.78	5.96	58262	59.52	5.55	123852	61.41	5.76	45592	60.21	5.10	30947	30.03	89.05
Pitch (Hz)	130.79	26.53	58262	226.08	30.71	123852	136.48	32.07	45592	224.23	35.70	30947	40.21	399.12
Alternative Acoustic Metrics														
Tempo (standardized)	-1.10	0.70	52614	-0.85	0.80	113440	-1.75	0.84	45555	-1.16	0.87	30927	-5.06	44.91
Volume (dB, median)	61.89	7.17	58262	61.73	6.60	123852	63.76	6.74	45592	62.62	5.69	30947	30.03	88.69
Volume (standardized)	-0.46	0.96	52614	-0.41	0.95	113440	-0.45	1.00	45555	-0.55	1.00	30927	-6.27	5.18
Pitch (Hz, median)	128.95	27.14	58262	225.37	31.72	123852	134.17	32.57	45592	221.86	37.42	30947	40.21	399.12
Pitch (standardized)	-0.24	0.94	52614	0.59	0.98	113440	0.38	1.02	45555	-0.33	1.03	30927	-6.05	10.56
Treatment Variables														
Treatment: main	0.09	0.29	58041	0.09	0.29	123401	0.06	0.24	45591	0.05	0.23	30947	0.00	1.00
Treatment: stricter	0.06	0.24	58041	0.06	0.24	123401	0.04	0.20	45591	0.04	0.19	30947	0.00	1.00
Treatment: vague	0.07	0.26	58041	0.07	0.25	123401	0.04	0.21	45587	0.04	0.19	30945	0.00	1.00
Treatment: high temperature	0.10	0.30	58041	0.10	0.30	123401	0.06	0.25	45590	0.06	0.23	30946	0.00	1.00
Treatment: continuous	0.06	0.17	58041	0.06	0.17	123401	0.08	0.20	45591	0.07	0.19	30947	0.00	1.00
political-MG	0.00	0.03	58262	0.00	0.03	123852	0.00	0.07	45592	0.01	0.08	30947	0.00	1.00
Sociodemographics														
Female/feminine speaker	0.00	0.00	58262	1.00	0.00	123852	0.00	0.00	45592	1.00	0.00	30947	0.00	1.00
Age: 1: teen;...7: seventy+	2.13	0.70	26839	2.38	0.66	5289	2.58	0.86	37358	2.53	0.91	19909	1.00	7.00
Born in special economic zone	0.08	0.27	25326	0.17	0.38	5372	0.00	0.00	21955	0.00	0.00	9501	0.00	1.00
Born in open coastal city	0.47	0.50	25326	0.55	0.50	5372	0.00	0.00	21955	0.00	0.00	9501	0.00	1.00
Born in inland China	0.53	0.50	25228	0.44	0.50	5293	.	.	0	.	.	0	0.00	1.00
Report Chinese accent	1.00	0.05	25291	0.99	0.12	5370	.	.	0	.	.	0	0.00	1.00
Report Taiwanese accent	.	.	0	.	.	0	1.00	0.07	21938	0.98	0.15	9270	0.00	1.00

Notes. Authors' calculations based on mp3 files and metadata from Mozilla's Common Voice corpus (commonvoice.mozilla.org). Each observation is an individual–sentence pair. Tempo is defined as characters per second over the full mp3 duration. Volume is the mean RMS amplitude in decibels across voiced frames (threshold >30 dB). Pitch is the mean fundamental frequency in Hertz across voiced frames (range 40–400 Hz). Standardized variables subtract the country–gender-specific mean and divide by the corresponding standard deviation. Treatment variables indicate whether a sentence is classified as political under various prompts and thresholds; see Section 3 for details. Sociodemographic variables are self-reported by speakers on the Common Voice platform.

Table A.2: Accents coding

Self-reported accent	Deduced area	Special zones	Coastal province	Close to Beijing	H repression 1966-71	H repression 2006-18
Malaysia Chinese	Malaysia	0	0	0	0	0
Malaysian Chinese	Malaysia	0	0	0	0	0
出生地：11 北京市	Beijing	0	1	0	1	0
出生地：11 北京市, 標準漢語普通話	Beijing	0	1	0	1	0
出生地：12 天津市	Tianjin	1	1	1	0	1
出生地：13 河北省	Hebei	0	1	1	1	1
出生地：14 山西省	Shanxi	0	0	1	1	1
出生地：15 内蒙古自治区	Inner Mongolia	0	0	0	1	1
出生地：21 辽宁省	Liaoning	1	1	0	1	1
出生地：21 辽宁省, 辽宁西部, 非正经东北话	Liaoning	1	1	0	1	1
出生地：22 吉林省	Jilin	0	0	0	1	1
出生地：23 黑龙江省	Heilongjiang	0	0	0	0	1
出生地：31 上海市	Shanghai	0	1	0	1	0
出生地：32 江苏省	Jiangsu	0	1	0	0	0
出生地：33 浙江省	Zhejiang	1	1	0	1	0
出生地：33 浙江省, 普通话等级水平三甲, 较为标准, 普通话	Zhejiang	1	1	0	1	0
出生地：34 安徽省	Anhui	0	0	0	0	0
出生地：35 福建省	Fujian	1	1	0	0	0
出生地：35 福建省, 轻微的方言口音	Fujian	1	1	0	0	0
出生地：36 江西省	Jiangxi	0	0	0	1	0
出生地：36 江西省, 出生地：11 北京市	Jiangxi / Beijing	0	1	0	1	0
出生地：37 山东省	Shandong	0	1	0	0	0
出生地：41 河南省	Henan	0	0	0	0	
出生地：41 河南省, 标准, 普通话二级甲等	Henan	0	0	0	0	1
出生地：42 湖北省	Hubei	0	0	0	0	0
出生地：43 湖南省	Hunan	0	0	0	1	1
出生地：43 湖南省, 株洲普通话	Hunan	0	0	0	1	1
出生地：44 广东省	Guangdong	1	1	0	1	1
出生地：44 广东省, 第二語言 · 自認有母語程度	Guangdong	1	1	0	1	1
出生地：45 广西壮族自治区	Guangxi	0	1	0	1	1
出生地：45 广西壮族自治区, 出生地：44 广东省	Guangxi/Guangdong	1	1	0	1	1
出生地：46 海南省	Hainan	1	1	0	0	1
出生地：50 重庆市	Chongqing	0	0	0	0	1
出生地：50 重庆市, 出生地：13 河北省	Chongqing/Hebei	0	1	0	1	1
出生地：51 四川省	Sichuan	0	0	0	0	1
出生地：52 贵州省	Guizhou	0	0	0	0	1
出生地：52 贵州省, 出生地：33 浙江省, 出生地：21 辽宁省	Guizhou	0	0	0	0	1
出生地：53 云南省	Yunnan	0	0	0	1	1
出生地：61 陕西省	Shaanxi	0	0	0	1	1
出生地：62 甘肃省	Gansu	0	0	0	1	0
出生地：64 宁夏回族自治区	Ningxia Hui	0	0	0	1	0
出生地：65 新疆维吾尔自治区	Xinjiang Uyghur	0	0	0	1	1
出生地：81 香港特别行政区	Hong Kong	0	1	0	0	1
广州	Guangdong	1	1	0	1	1
我在廣東出生。	Guangdong	1	1	0	1	1
昆明市	Yunnan	0	0	0	1	1
普通话	Unspecified	.	.	.	.	.
普通话, Unspecified	Unspecified	.	.	.	.	.
标准普通话	Unspecified	.	.	.	.	.
標準音	Unspecified	.	.	.	.	.
海外华裔	Unspecified	.	.	.	.	.
深圳市, 母语, 潮州人	Guangdong	1	1	0	1	1
马来西亚	Malaysia	0	0	0	0	0
马来西亚华文	Malaysia	0	0	0	0	0
外国人 · 北方口音	Foreigner	0	0	0	0	0
Mixed origin	Mixed	.	.	.	.	.
Unspecified origin	Unspecified	.	.	.	.	.

"Costal" = 1 for Beijing, Fujian, Guangdong, Guangxi, Hainan, Hebei, Jiangsu, Liaoning, Shandong, Shanghai, Tianjin, and Zhejiang; 0 otherwise. "High death ratio during CR" = 1 for Beijing, Gansu, Guangdong, Guangxi, Hebei, Hunan, Inner Mongolia, Jiangxi, Jilin, Liaoning, Ningxia, Shaanxi, Shanghai, Shanxi, Xinjiang, Yunnan, and Zhejiang; 0 otherwise.

Figure A.1: Voice clips recording

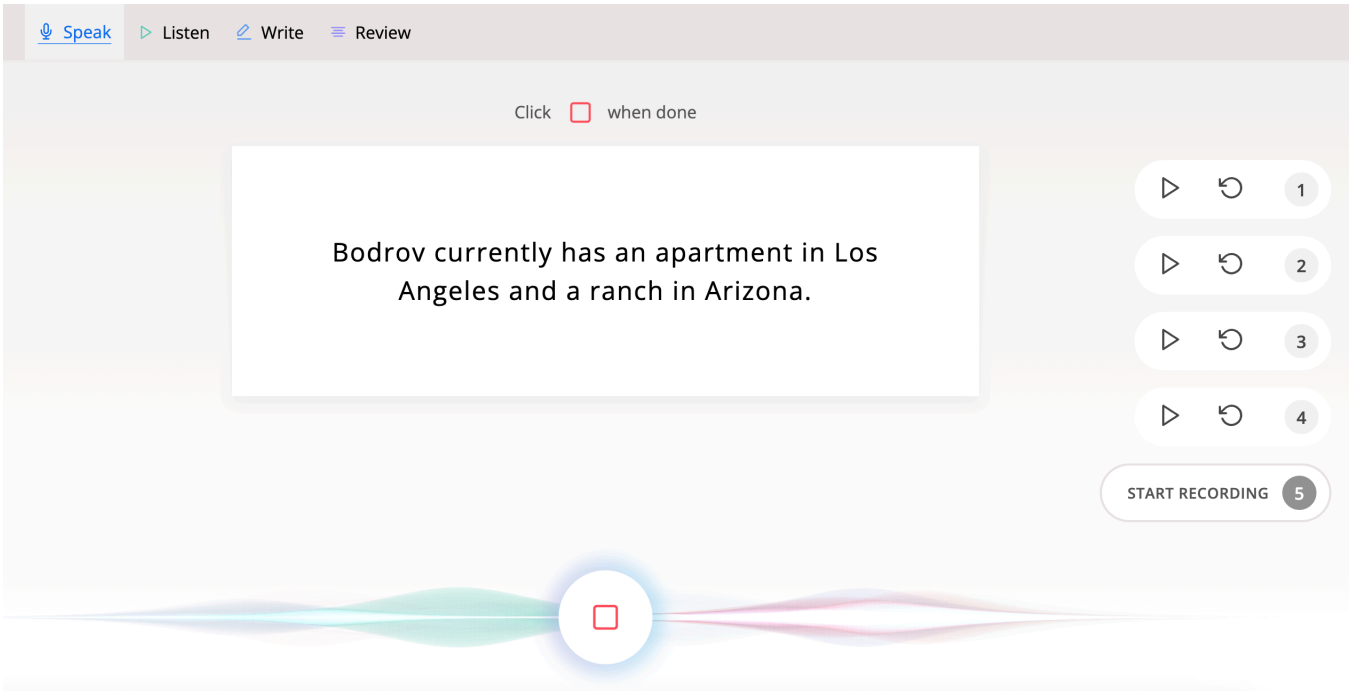
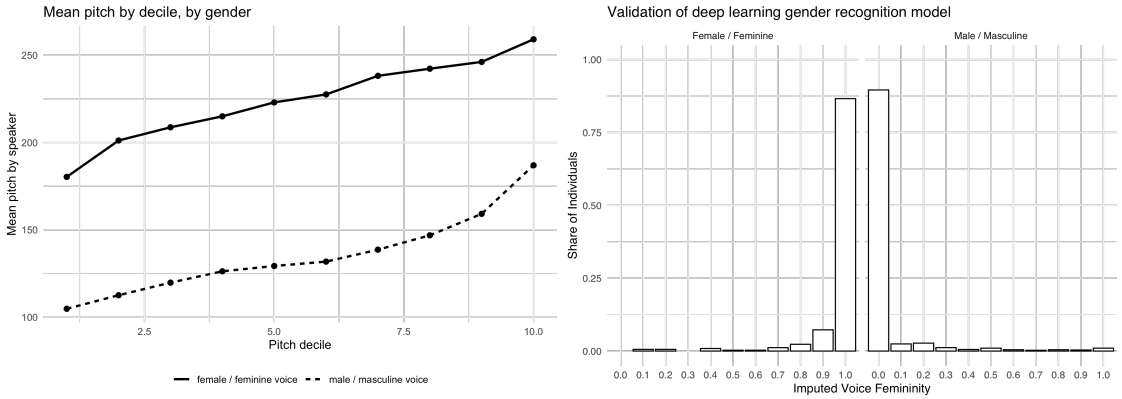


Figure A.2: Pitch and gender



B Robustness checks

B.1 Alternative construction of the dependent variables

Table B.1 checks the robustness of speaking fast and low to changes in how the dependent variables are constructed. The first column simply reports the coefficients from the main specification. In the second column 'Log' we take the natural logarithm of each dependent variables, the resulting magnitudes can hence be interpreted as percent changes, since they are somewhat limited, and we used those ones in the text to discuss the magnitudes. In the third column we standardize each DV and again use the resulting magnitudes in the main body of the paper. The 'AI-Std' column shows DiD estimates where pitch, tempo and volume are standardized deviation from the pitch, tempo and volume produced by synthetic AI-bots reading the exact same sentence as explained in the previous section. Here we see that pitch loses significance. Moreover, we see that for both tempo and volume, the key interaction term has a similar magnitude to that obtained when implementing a simple standardization, suggesting that AI bots are indeed not able to mimic accurately those human contextual emotions behind speaking fast and low. When we use the instead of mean for both pitch and volume, in column five, we get again an insignificant DiD estimate for pitch and a $\approx 40\%$ stronger estimate for volume. This is particularly reassuring and meaningful. It suggests that our results on volume are not driven by sudden changes in volume that produce outlier dB in some milliseconds. In the fifth and sixth columns we change how we measure pitch and volume, leaving tempo blank. What we see in the last two columns of table B.1 is that once again pitch outcomes seem particularly sensitive to their measurement, while volume outcomes are very stable.

Table B.2 analyzes the subcomponents of tempo. In the main analysis, we derived ″tempo″ , the typical metric used in the phonetics literature, as number of characters divided by the length of the MP3. However, the total length of the MP3 includes periods of silence before and after speech. It is therefore interesting to examine whether the main result—that Mainland Chinese men speak ″faster″ —reflects actual speech tempo or instead differences in recording behavior, whereby silences are shorter because people start or stop recordings more quickly. If Mainland Chinese men feel urged to complete the recording of political sentences due to some embodied stress, this should manifest not only in how fast they speak but also in how quickly they handle the recording itself. To test this possibility, we examine a few categories. Identifying precisely when a sentence begins and ends is not straightforward, as background noise can be difficult to distinguish from speech. We are therefore aware that some measurement error is unavoidable. Using the same criteria adopted for pitch and volume, we consider admissible pitch values between 40 and 400 Hz and admissible volume

levels above 30 dB. We then record the moments when the first and last pitch or volume within these ranges occur. To clarify: every sound has volume, but not every sound has pitch. For example, the *s* (or *f*) sound in English has no pitch—it cannot be sung on a musical scale—while *z* or *n* do. Using the pitch criterion is therefore conservative: noise such as a tap on the table may exceed the volume threshold but has no pitch, and will thus be excluded. However, if a word begins or ends with an unvoiced consonant, that portion will also be omitted. The volume criterion has the opposite limitation. With these caveats in mind, table B.2 reports DiD estimates and shows that "speaking fast" is both a matter of attitude and one of vocal delivery: the reduction in .2 seconds documented in column 1 and discussed in the paper seems similarly composed by the various sub-intervals of time in the MP3. In columns 2-3, we report DiD estimates on the *onset*: the first sign of voice, either with a pitch or volume criterion. We see that the onset is $\approx .1$ second quicker among Mainland Chinese men when prompted with political sentences, significant when using the pitch and non significant when using the volume criterion. Furthermore, when analyzing DiD estimates on the *offset*: the duration from the start to the last sign of voice, the results using pitch are basically the same as when using the whole duration as in the baseline approach, while they amount to a $\approx .1$ again when using the volume criterion. Instead, when analyzing the silence at the end of the sentence, we find small magnitudes and no significance, though the sign is still negative.

B.2 Alternative construction of the treatment

Table B.5 checks the robustness of speaking fast and low to changes in how the treatment is constructed. The details are provided in the table note. Throughout seven different operationalizations of the treatment, involving varying prompt, model settings, model and coding agents, we find remarkably stable results for tempo and volume but not for pitch, once again. Differences in tempo outcomes are negligible, while differences in volume outcomes are a bit more marked. The reader should have in mind that the subset of sentences manually labelled by two RAs is different from the pool of common sentences, hence the different outcomes we get in such a case—with considerably lower standard deviation and DiD estimates alike—are not directly comparable to the other ones reported in table B.5. The way in which all sentences are coded by these different approaches in reported in table B.3, while table B.4 reports the details of the balance test in the main body (figure 3).

Table B.6 tentatively looks at sub-treatment effects, sharing treatment sentences by whether they are opinions or factual statements, about foreign or domestic policies, with a negative or neutral tone. Admittedly, the analysis is imperfect in two ways. First, the

number of sentences becomes quite small when we seek to assess thee sub-treatments, hence possibly unrepresentative. Second, the sentences in various sub-treatment groups may be different in other ways than those captured by their label. It should be taken with caution as a supplement of column 6 in table B.5: it helps us qualifying what political domain might be particularly sensitive. Unsurprisingly, speaking fast and low is more marked when sentences are about opinions rather than facts and when they present a negative than neutral tone. It is also more marked on foreign than domestic policy.

B.3 Alternative specifications

Table B.7 checks the robustness of our main results to alternative sample restrictions, clustering choices, and additional controls. The first column simply reports the coefficients from the baseline specification for ease of comparison. In column two ('Sample: Has $T=1$'), we restrict the analysis to individuals who were at least once assigned a political sentence, ruling out the possibility that some respondents may have quit the recording process upon encountering politics. The results remain significant and comparable in magnitude. In column three ('Sample: Has PoB'), we limit the sample to respondents reporting to have been born in China or Taiwan. The sharp drop in sample size reflects missing sociodemographic information rather than a large share of speakers born elsewhere. Despite the reduced sample, the tempo and volume coefficients remain strong, while the pitch effect loses significance. A similar picture emerges in column four ('Sample: Has gender'), where restricting to individuals reporting their gender again shrinks the sample but leaves the main results on tempo and volume robust. Column five ('Cluster: id') and column six ('Cluster: id+sentence') vary the clustering level of standard errors; point estimates remain stable, and statistical significance is preserved, indicating that the results are not an artifact of clustering choices. In column seven ('Control: linguistic'), we add controls for Mandarin tone distribution and sentence readability to capture additional text-level variation. The main effects on tempo and volume remain intact, suggesting that unobserved linguistic differences are not driving our findings. Finally, in column eight ('Control: order T'), we account for the random order in which treatment sentences were assigned and for the total number of political sentences seen by each individual; results again hold. Overall, table B.7 demonstrates that our findings—speaking faster and more quietly about politics in Mainland China—are robust across a wide range of alternative specifications. Pitch results, by contrast, are consistently fragile, underlining that tempo and volume are the most reliable prosodic markers of political stress.

B.4 Additional speech acoustics: Jitter and shimmer

In table B.8, we estimate null DiD effects of speaking about politics on voice quality, captured by two further metrics. Jitter and shimmer are classical measures of voice quality that capture microperturbations in the speech signal, often associated with phonatory stability and vocal health.Jitter and shimmer capture the stability of the voice signal at the micro level. Jitter measures how much the pitch wavers from one vocal cycle to the next, while shimmer measures the equivalent variation in loudness. Both are standard in clinical voice research, where high values typically indicate irregular vocal fold vibration associated with voice disorders or strain. In non-clinical contexts, they reflect subtler variation in how controlled or effortful a speaker sounds—breathier or more tense phonation, for instance. Including them here allows us to test whether speaking about politics affects not just how fast, loud, or high someone speaks, but also how steady and controlled their voice is at a finer level of detail.

B.5 Alternative mechanisms

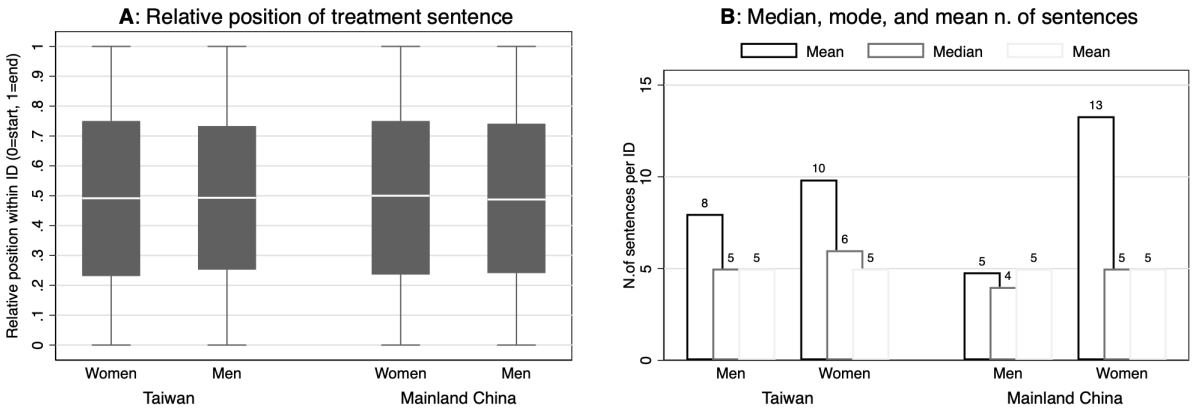
Gender socialization. A recent paper by Fan, Pan, and Zhang (2025) highlights that the uncertainty generated by the lack of specific guidelines about what is or is not allowed in contemporary China, possibly kept in place by the authoritarian regime to stimulate costless compliance, generates gender-diverging compliance strategies. Due to different early socialization patterns shaping men to be assertive and women to be agreeable, Fan, Pan, and Zhang (2025, p 1892) document that when tasked with defending the regime′ s image but given no strict instructions, male instructors tended to aggressively assert the Party line and censor sensitive discussion, while female instructors fostered open dialogue and tried to gently persuade others toward pro-China views. Such an attitude may also take place inside our natural experiment: men self-censor, while women speak normally. The plausibility of such ″gendered socialization″ mechanism would increase if the acoustic markers of stress are stronger in provinces with higher gender inequality, where, presumably, the extent to which gender matters for communicative socialization is more marked. We, however, do not find evidence for this possibility. Data on gender equality in China are scant, however, we found data on the the ratio of the female to the male Subnational Human Development Index (SHDI) by province, allowing to replicate the same within China analysis in table 3. Each SHDI is the geometric mean of three dimension indices—education, health, and standard of living—constructed identically but with gender-specific indicators. Values < 1 indicate disadvantage for women, > 1 disadvantage for men, and $= 1$ parity — higher values mean higher gender equality. Data come from national statistical offices (incl. Eurostat), the

Global Data Lab's Education, and UNDP's HDI site. For 1990–2017, missing subnational years were filled by interpolation/extrapolation, anchored to complete national UNDP series. We create a dummy that takes the value of 1 if the province has higher than median level of (SHDI). We find that the DiD estimate for Political×China is negative on tempo and positive on volume, but the effect is very mild ($\hat{\beta} = -.11$,$SE = .09$ for tempo and $\hat{\beta} = .37$,$SE = .75$ for volume). It is hence unlikely that men and women are equally afraid to talk about ordinary politics and only differ in how they cope with the same feelings.

Sampling. Because speakers are volunteers, the results may partly reflect differential selection or attrition between participants from Mainland China and Taiwan and /or between men and women. While individual FE net out time-invariant characteristics such as e.g. different recording technologies that may be more or less popular in Mainland China and Taiwan and e.g. display different sensitivities to e.g. pitch variation, it could be that the self-selection of Mainland Chinese and Taiwanese speakers dramatically differ in a way that would not generalize to a more representative sample. Three observations here: (i) if anything, we would expect the Mainland Chinese sample to exclude individuals that are particularly afraid of their digital footprint, hence if a sampling bias exists then it runs against our findings; (ii) accessing Modzilla common voice corpus to donate one's voice did not require a vpn in China; (iii) the distribution of aggregate pitch, tempo and volume reproduce stylized facts: Taiwanese are louder and more melodic (see 1). (iv) The share of women is 39.6% in China and 34.7% in Taiwan while the share of speakers aged less than 30 is 95.4% in China and 85.3% in Taiwan: while some differences in the few demographics we know exist, they are somewhat limited in size.

Additionally, male and female speakers across Taiwan and Mainland China do not seem to substantially differ in terms of attrition. The panel A of figure B.1 shows that the probability distribution of treatment sentences. If such probability distribution has some positive skewness, we mat suspect that, as soon as they see a sentence with a political connotation, Chinese men stop recording. Instead, the probability distribution is a uniform within the recording session for each subgroup. This likely reflects, in part, the fact that even political sentences can be sensitive enough from the perspective of some speakers to trigger a specific speech acoustics pattern but not so strong as to lead speakers to stop recording—the reader may remember that they can exit any time). Moreover, while some differences in the mean or median number of recorded sentences can be observed in panel B of figure B.1, these differences are overall limited, with the mean being sensitive to outliers who recorded an extremely high number of sentences. The fact that the mode is 5 for every subgroup is not surprising: recording sessions come in batches of 5, after which speakers can decide to

Figure B.1: Attrition by gender and area



Notes. Panel **A** shows the relative position of the treatment sentence within each speaker′ s set of recorded sentences. The variable on the y-axis, relative position within ID, ranges from 0 (first sentence) to 1 (last sentence) within each individual′ s ordered sequence. The plot focuses on treatment sentences. Panel **B** summarizes the number of recorded sentences per individual (ID) by gender and country. For each subgroup, it reports three measures of central tendency: the median, the mode, and the mean number of sentences. The mean is calculated after trimming the top and bottom 20 percent of observations within each subgroup to limit the influence of extreme values. Bars are white with distinct outline shades to differentiate the three statistics, and values are displayed separately for men and women in Taiwan and Mainland China.

either record again or stop).

Biological differences. One possibility could be that gender differences observed in Mainland China but not in Taiwan have to do with biological differences between genders that are to be expected in China but not in Taiwan. The gender difference would be explained by the fact that men have more flexibility of their speech acoustics due to biological differences, but such gender difference is specific to Mainland Chinese. This possibility is extremely unlikely. First, women have greater, not smaller, vocal flexibility than men. Moreover, mainland Chinese and Taiwanese are by far and large ethnically similar. Another related possibility is that men being more interested in politics, they are quicker at spotting that a sentence is political and at adapting. Again, it seems very implausible because our sentences are short and simple.

Table B.1: Robustness: Dependent variable definition

	Baseline	Log	Std	AI-Std	Median	L-bounds up	U-bounds down
A: Tempo (characters/second)							
Political × China	.201***	.080***	.219***	.196**			
	(.072)	(.030)	(.085)	(.087)			
Output mean	2.756	.977	.147	-1.525	.	.	.
F-stat	7.765	7.374	6.658	5.065	.	.	.
N. obs	14,967	14,967	14,967	14,909	.	.	.
B: Volume (dB)							
Political × China	-2.270**	-.040**	-.383**	-.440***	-3.127**	-2.284***	-1.612***
	(.892)	(.016)	(.153)	(.157)	(1.232)	(.851)	(.611)
Output mean	61.737	4.119	.109	-.430	64.114	61.149	64.877
F-stat	6.473	6.236	6.312	7.851	6.447	7.194	6.958
N. obs	14,967	14,967	14,967	14,909	14,967	14,948	14,946
C: Pitch (Hz)							
Political × China	4.733**	.035**	.162*	.089	4.416	2.584	.648
	(2.280)	(.016)	(.087)	(.082)	(3.161)	(2.387)	(2.580)
Output mean	136.550	4.892	.037	.286	134.527	139.699	140.426
F-stat	4.312	4.550	3.421	1.165	1.952	1.171	.063
N. obs	14,967	14,967	14,967	14,909	14,967	14,946	14,946
Common modelling assumptions							
Sentence FE	yes	yes	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The first column 'Baseline' reports the coefficients from the main specification for comparison. In the second column 'Log' we take the natural lograrithm of each dependent variables, magnitudes can hence b e interpreted as semi-elasticities. In the third and fourth columns, 'Std' and 'AI-Std' we replace the original DV with a standardized version of them. For the 'Std' column, we apply the typical standardization by subtracting to each observation the mean and dividing by the standard deviation, both country-gender specific. The 'AI-Std' column operates a more complex and original standardization. Specifically, the main dependent variables are defined as $z_i, s = \frac{y_i, s - \hat{y}_s}{\sigma_\hat{y}}$, where $\sigma_\hat{y}$ is the standard deviation across all AI sentence-level values for each output, hence z_i, s captures the standardized deviation of each spoken sentence with respect to the AI benchmark. In the fourth column 'DV median' we use the median instead of mean for both pitch and volume – of course e do not have an equivalent for 'tempo', hence we leave it blank. In the fifth and sixth columns we change how we measure pitch and volume, leaving tempo blank. 'Lowerbound up' lifts lowerbound Hz for recording pitch from 40 to 60 and lowerbound db for volume from 30 to 45, leaving upperbounds unchanged to 400Hz and 200db resp. 'Upperbound down' we set maximum pitch to 350hz and maximum volume to 90db.

Table B.2: Speaking fast or clicking fast?

	Overall length	Start mp3 first Hz	Start mp3 first dB	Start mp3 last Hz	Start mp3 last dB	Last Hz end mp3	Last dB end mp3
	Length of speech (seconds)						
Political × China	-.211**	-.131**	-.109	-.218***	.110	-.021	-.024
	(.104)	(.052)	(.075)	(.082)	(.093)	(.033)	(.032)
Output mean	2.972	.719	.481	2.365	2.552	.605	.418
F-stat	4.096	6.354	2.125	7.018	1.395	.393	.599
N. obs	14,967	14,946	14,948	14,946	33,118	33,114	33,118
	Common modelling assumptions						
Sentence FE	yes	yes	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. We focus on men and use the baseline specificaiton in equation 1 in the main body of the paper. In the first column, we report the DiD estimate on the overall duration of the mp3. In the second and third columns, we focus on the distance between the start of the mp3 and a candidate for the start of the spoken part. Such candidates are the first pitch bounded between 40 and 400 Hz in column 2 and the first intensity above 45dB in column 3—they are used in subsequent analsis as well. In the fourth and fifth columns, we focus on the distance between the start of the mp3 and a candidate for the end of the spoken part. In the sixth and last columns, we focus on the distance between the last candidate for the end of the spoken part and the end of the mp3.

Table B.3: Political Sentences in Main Specification

Sentence Mandarin	Translation English	Main	Prompt 1	Prompt 2	Model nano	Temp = 1	T $\in 0, 1$
中國有自由和民主	china has freedom and democracy	1	1	1	1	1	0.9
立法院圍牆	legislative yuan wall	1	1	1	1	1	0.9
除了選舉還是選舉	besides elections elections	1	1	1	1	1	0.8
當選三峽庄協議員	elected as a Three Gorges Village council member	1	1	1	1	1	0.8
加強城市外交	strengthen city diplomacy	1	1	1	1	1	0.8
政客的特点就是	the characteristic of politicians is that	1	1	1	1	1	0.8
農委會已修法	the council of agriculture has amended the law	1	1	1	1	1	0.8
可能會成為政治人物	may become a politician	1	1	1	1	1	0.5
当时有其特殊的政治时空背景	there was a special political time and space background at that time	1	1	1	1	1	0.5
兩國總統都沒有直接通過電話	neither president spoke directly by phone	1	1	1	0	1	0.8
津貼補助的政策奏效	the policy of allowances and subsidies works	1	1	1	0	1	0.8
部分條文修正條文	amendments to some provisions	1	1	1	0	1	0.5
在当时直接做个投票	just vote right then and there	0	1	1	1	1	0.5
净化社会治安	purify public order	1	1	1	0	1	0.5
中央目的事業主管機關查核	inspection by the central authority in charge of the target industry	1	0	0	1	1	0.9
委员遴聘条件程序	member selection criteria and procedures	1	0	0	1	1	0.8
審查會議之議決結果	the results of the review meeting	1	0	0	1	1	0.8
大學學生也只有一票	college students only have one vote	0	1	1	1	0	0.5
强化市民参与	strengthen citizen participation	0	1	1	1	0	0.5
召開一般程序審查會議	convene a general procedure review meeting	1	0	0	0	1	0.8
拆除破壞警方架設之拒馬	dismantling and destroying police-erected barricades	1	0	0	0	1	0.8
應有半數以上委員出席	more than half of the committee members should be present	1	0	0	0	1	0.8
審查委员应簽署保密协定	review committee members should sign a confidentiality agreement	1	0	0	0	1	0.8
提昇國家競爭力	improve national competitiveness	1	1	0	0	1	0.7
推动各项福利措施	promote various welfare measures	1	0	0	0	1	0.7
警察將與群眾直接肢體接觸	the police will have direct physical contact with the public	1	0	0	0	1	0.7
只是臃腫的科層架構	just a bloated bureaucratic structure	0	1	1	0	0	0.3
日俄战争爆发	the russojapanese war breaks out	1	0	0	1	1	0.2
通報警察單位處理	notify the police unit for processing	1	0	0	0	1	0.8
出席國際會議	attend international conferences	1	0	0	0	1	0.7
造成多名记者警察受伤	many journalists and police officers were injured	1	0	0	0	0	0.5
也有空襲以及戰亂	there were also air raids and wartime turmoil	1	0	0	0	1	0.5
拉近雙方的心理距離	close the psychological distance between the two parties	0	1	1	0	0	0.2
他們說出席委員都在睡覺	they said all the members present were sleeping	1	0	0	0	1	0.2
第一次世界大戰結束	end of world war 1	1	0	0	0	1	0.1
第六條第五點	article 6 point 5	0	1	0	0	0	0.0
雙方情緒都不是很好	both parties are not in a good mood	0	1	1	0	0	0.0
已经丢给社长，静候他的消息	i have already given it to the president and am waiting for news from him	0	1	1	0	0	0.0
正副社長限定	limited to president and vice president	0	1	1	0	0	0.0
本來我是想跟社長說	originally i wanted to tell the president	0	1	1	0	0	0.0
社長沒訂到甜點卻沒告知	the president didnt order dessert but didnt tell me	0	1	1	0	0	0.0
有效瓦解群眾暴力攻擊	effectively disrupt violent mass attacks	0	0	0	0	1	0.8
應有三分之二以上委員出席	more than twothirds of the members should be present	0	0	0	0	1	0.8
較能配合法令	more able to comply with laws	0	0	0	0	1	0.7
握有掌控權	have control	0	0	0	0	1	0.5
他没有权力这么做	he has no right to do this	0	0	0	0	1	0.5
健全醫療制度	improve the medical system	0	0	0	0	0	0.5
推動災區的復原重建	promote the recovery and reconstruction of disaster areas	0	0	0	0	1	0.5
保障農民參加的權益	protect farmers rights and interests in participating	0	0	0	0	1	0.5
成就現今社會局勢的歷史脈絡	the historical context that shaped todays social situation	0	0	0	0	1	0.3
任何可能侵犯講師權益之行為	any behavior that may infringe on the rights of lecturers	0	0	0	0	0	0.2
新的价值体系	new value system	0	0	0	0	0	0.2
權力使我欣喜	power makes me happy	0	0	0	0	0	0.2
建立关怀温馨的社会	build a caring and warm society	0	0	0	0	0	0.1

Sentence Mandarin here is in simplified characters, stripped of punctuations to ease merging. The same applied to the translation in English, done with Claude (Anthropic).

Table B.4: Covariate balance for sentence characteristics

DV	Mean (t = 0)	Mean (t = 1)	Political	Political × China	N. obs.
Entropy	0.446	0.526	-0.028***	0.024***	238,581
			(0.001)	(0.001)	
Readability	114.517	112.578	0.869***	-0.671***	238,581
			(0.033)	(0.041)	
Word count	6.347	7.831	-0.778***	0.608***	238,581
			(0.028)	(0.035)	
N. characters count	12.131	16.349	-0.836***	1.796***	238,581
			(0.053)	(0.064)	
Tone 1	0.231	0.213	-0.027***	0.019***	238,573
			(0.002)	(0.002)	
Tone 2	0.220	0.243	0.021***	-0.011***	238,573
			(0.002)	(0.003)	
Tone 3	0.186	0.174	0.004**	-0.010***	238,573
			(0.002)	(0.002)	
Tone 4	0.363	0.370	0.002	0.002	238,573
			(0.003)	(0.004)	

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Sentence-level entropy of the within–sentence token distribution. For a sentence s with token types $\{t_1, \ldots, t_K\}$, counts n_k, total $N = \sum_{k=1}^{K} n_k$, and empirical probabilities $p_k = n_k/N$, the entropy is $H(s) = -\sum_{k=1}^{K} p_k \log p_k$ with log base 2 (bits). Higher values indicate a more uniform token distribution and greater lexical complexity. Tokens are computed at the sentence level. Readability is proxied by the Flesch Reading Ease (FRE) index, which combines average sentence length and average syllables per word as follows: $\text{FRE} = 206.835 - 1.015 \frac{\text{words}}{\text{sentences}} - 84.6 \frac{\text{syllables}}{\text{words}}$. FRE ranges from 0 (hard) to 100 (easy). To interpret it as a *complexity* measure aligned with entropy whereby higher values indicate more complexity, we rescale the index with the transformation $C_{\text{FRE}} = 100 - \text{FRE}$. Word count and number of characters are obvious metrics. As one can imagine, these first four metrics of sentence complexity are extremely correlated. Within the sample of common sentences between China and Taiwan, for example, the weakest correlation is that between number of characters and words. Yet still is .84. For each sentence s, we segment the text into Mandarin syllables and assign each syllable a lexical tone $t \in \{1, 2, 3, 4\}$ (high level, rising, dipping, falling) from the transcription (Pinyin with diacritics or numeric tags). Let n_t be the count of syllables with tone t, and let $N = \sum_{t=1}^{4} n_t$ be the total number of tone–bearing syllables with tones 1–4. The share for tone t is $\text{share}_t(s) = n_t/N$. Neutral–tone syllables, punctuation, non-Chinese tokens, or untagged items are excluded from N (a separate neutral-tone share can be reported if needed). By construction, $\sum_{t=1}^{4} \text{share}_t(s) = 1$.

Table B.5: Robustness: Treatment definition

	Baseline	Prompt 1	Prompt 2	$T \in (0,1)$	Temp=1	Model	Raw
A: Tempo (characters/second)							
Political × China	.201***	.178**	.194***	.215**	.190***	.264**	.170***
	(.072)	(.076)	(.074)	(.090)	(.059)	(.103)	(.066)
Output mean	2.756	2.756	2.756	2.755	2.758	2.759	2.756
F-stat	7.765	5.506	6.934	5.740	10.217	6.514	6.728
N. obs	14,967	14,967	14,967	14,967	14,853	14,853	14,967
B: Volume (dB)							
Political × China	-2.270**	-1.518*	-1.610*	-1.926**	-2.214***	-2.206*	-2.011**
	(.892)	(.872)	(.846)	(.861)	(.655)	(1.276)	(.839)
Output mean	61.737	61.737	61.740	61.766	61.741	61.742	61.740
F-stat	6.473	3.033	3.623	5.008	11.410	2.990	5.744
N. obs	14,967	14,967	14,967	14,967	14,853	14,853	14,967
C: Pitch (Hz)							
Political × China	4.733**	-4.802	-2.717	8.111**	2.091	-3.512	2.249
	(2.280)	(3.272)	(3.545)	(3.984)	(2.689)	(2.774)	(2.941)
Output mean	136.550	136.562	136.566	136.567	136.563	136.580	136.543
F-stat	4.312	2.154	.587	4.145	.604	1.603	.585
N. obs	14,967	14,967	14,967	14,967	14,853	14,853	14,967
Common modelling assumptions							
Sentence FE	yes	yes	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The first column provides baseline estimates to ease comparison. The treatment is assigned to a 196 voice clips. The second column 'Prompt 1' follows the original one, but avoid giving examples in brackets about the type of sentences to be labelled as political. It reads: 'You are a text classifier. Assign a political relevance score between 0 and 1 to each sentence. A sentence if political if it mentions political institutions, political opinions, political debates or political events.' The treatment is assigned to a 164 voice clips. The third column 'Prompt 2' uses an alternative simpler prompt 'Label each sentence as 1 if it is related to politics and 0 if it is unrelated'– here the definition is left to the LLM, resulting in 177 voice clips considered treatment. The fourth column proposes a continuous treatment which is based on the main prompt: 'You are a text classifier. Assign a political relevance score between 0 and 1 to each sentence. A sentence if political if it mentions political institutions, political opinions, political debates or political events. Prioritize sentences mentioning political institutions, political opinions, political debates or political events relevant to contemporary Taiwanese or Chinese politics over historical references unrelated with contemporary Taiwanese or Chinese politics.'. The fifth column 'Temp = 1' ' checks the robustness of the baseline estimation with respect to the temperature of the labelling prompt, which is 0 (deterministic) in the baseline and 1 here, hence allowing some randomness in the labelling process and resulting in approximately 50% more treated sentences. It labels in a less conservative way.The fact that some results are weaker makes sense: sentences that are less ambiguously about politics produce stronger results. The sixth column 'Model' checks the robustness of the baseline estimation to a different, less sophisticated and expensive model (GPT 4o mini). The seventh column 'Raw' allows the treatment status to be different in China or Taiwan.

Table B.6: Robustness: by sub-treatment

	Baseline	Opinions	Facts	Foreign policy	Domestic policy	Negative tone	Neutral tone
A: Tempo (characters/second)							
Political × China	.201***	.347***	.114*	.529***	.184**	.194***	.060
	(.072)	(.128)	(.059)	(.160)	(.075)	(.074)	(.072)
Output mean	2.756	2.756	2.756	2.756	2.756	2.756	2.756
F-stat	7.765	7.371	3.748	10.938	6.088	6.907	.698
N. obs	14,967	14,948	14,967	14,967	14,967	14,967	14,967
B: Volume (dB)							
Political × China	-2.270**	-3.043*	-1.406*	-5.925***	-1.211	-2.633*	-1.389
	(.892)	(1.607)	(.840)	(1.540)	(.907)	(1.425)	(1.032)
Output mean	61.737	61.738	61.732	61.728	61.730	61.728	61.729
F-stat	6.473	3.583	2.798	14.793	1.782	3.413	1.811
N. obs	14,967	14,948	14,967	14,967	14,967	14,967	14,967
Common modelling assumptions							
Sentence FE	yes	yes	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. After reporting baseline findings, we divide sentences in 'opinions' and 'facts' in columns 2 and 3. The former category includes sentences such as 'improve competitiveness' whereas the latter includes sentences like 'the council of agriculture has amended the law'. Notice that some sentences are neither opinions neither facts, some may be ambiguous—in general, none of the dichotomic categories here are intended to cover all sentences. We then divide the sentences between foreign and domestic policy in columns 4 and 5, and finally we divide sentences depending on their tone, negative or neutral. This last division is a bit more speculative: 'many journalists and police officers were injured' has a negative tone, 'more than two thirds of the members should be present' has a neutral tone.

Table B.7: Robustness: specifications

	Baseline	Sample: Has PoB	Sample: Has gender	Cluster: only id	Cluster: id+sentence	Control: order T
A: Tempo (characters/second)						
Political × China	.201***	.200**	.212**	.201**	.201	.201***
	(.072)	(.097)	(.096)	(.088)	(.)	(.072)
Output mean	2.756	2.803	2.810	2.756	2.756	2.756
F-stat	7.765	4.221	4.883	5.284	6.333	7.765
N. obs	14,967	3,434	10,706	14,967	14,967	14,967
B: Volume (dB)						
Political × China	-2.270**	-3.607***	-3.535***	-2.270***	-2.270**	-2.270**
	(.892)	(.814)	(.810)	(.862)	(.899)	(.892)
Output mean	61.737	61.917	61.577	61.737	61.737	61.737
F-stat	6.473	19.612	19.067	6.941	6.385	6.473
N. obs	14,967	3,434	10,706	14,967	14,967	14,967
C: Pitch (Hz)						
Political × China	4.733**	1.275	1.072	4.733**	4.733***	4.733**
	(2.280)	(3.288)	(3.185)	(2.224)	(1.586)	(2.280)
Output mean	136.550	144.659	137.808	136.550	136.550	136.550
F-stat	4.312	.150	.113	4.528	8.912	4.312
N. obs	14,967	3,434	10,706	14,967	14,967	14,967
Common modelling assumptions						
Sentence FE	yes	yes	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The first column provides baseline estimates to ease comparison. The second column 'Sample: has PoB' subset on individuals who report to have been born in either Taiwan or China. Note that the drammatic drop in sample is not due to the fact that a very small minority of respondent lives somewhere else (e.g, in China only 14 respondents report to live in e.g. US or Malasya), but to the fact that most respondents do not provide their sociodemographics. Same for the third column 'has gender'. Imputing on gender is quite straightforward due to strong differences in pitch, but it could still be that women and men reporting or not reporting their vocal gender are for some reason different—if that was the case, we may learn about selection. It is hence understood mostly as a check on whether people who give their credentials are substantially different from the rest of the sample. In the fourth column 'Cluster: id' we cluster by ID and not by sentence, while in the fifth column 'Cluster: id and sentence' we cluster for both. Finally, in the column 'Control: order T' we control for the order at which the treatment sentence was exogenously assigned to respondents and for the total number of treated sentences they had.

Table B.8: Voice quality: Jitter and shimmer

	Jitter local	Jitter smooth	Shimmer local	Shimmer smooth
Voice quality				
Political × China	.000	.001	-.000	.001
	(.002)	(.001)	(.007)	(.005)
Output mean	.030	.015	.153	.058
F-stat	.002	.201	.003	.061
N. obs	14,945	14,945	14,945	14,945
Common modelling assumptions				
Sentence FE	yes	yes	yes	yes
Individual FE	yes	yes	yes	yes

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Jitter (local) is calculated as the average absolute difference between consecutive pitch periods, divided by the mean period, and expressed as a percentage; it captures short-term cycle-to-cycle irregularity in frequency. Jitter (smooth) is a smoothed version of jitter, where each period is compared to the mean of itself and its four nearest neighbors (five periods total), reducing the influence of single-period outliers and reflecting instability over a slightly longer window. Shimmer (local) is the average absolute difference in peak-to-peak amplitude between consecutive cycles, divided by the mean amplitude; it reflects cycle-to-cycle variability in vocal intensity. Shimmer (smooth) extends this measure by averaging absolute amplitude differences across a moving window of three consecutive periods, providing a smoother estimate of short-term amplitude perturbation.

Table B.9: Randomization of treatment

	Mainland China	Mainland China add missings	Taiwan	Taiwan add missings
Share of political sentences				
Age bracket	-.004	-.002	-.001	-.000
	(.007)	(.005)	(.004)	(.004)
Female	-.001	.004	.013	.010
	(.013)	(.005)	(.009)	(.006)
Total sentences recorded	.000	.000	.000	.000
	(.000)	(.000)	(.000)	(.000)
Missing gender (dummy)	-.011	-.016	-.006	-.006
	(.025)	(.017)	(.024)	(.015)
F-stat	.176	.825	.551	.829
Prob $> F$	.951	.509	.699	.507
R^2	.001	.001	.003	.002
N. obs	910.000	4436.000	682.000	1673.000

Notes. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. We collapse data at the individual level and calculate the share of treatment (political) sentences, using it as a dependent variable, regressed on the key sociodemographic covariates and three dummy variables capturing whether a datum is missing. Columns (2) and (4) keep missing values on age bracket and female giving them an arbitrary value of 99.