# Speech and Voice in Politics:

## Multimodal Analysis of Presidential Addresses to the Korean National Assembly[*]

*Kyusik Yang*[**]

This study examines presidential addresses to the Korean National Assembly from two perspectives: text-as-data and audio-as-data. To overcome the limitations of single-source approaches, it jointly analyzes what presidents say and how they say it. Twenty speeches delivered by Presidents Roh Moo-hyun, Lee Myung-bak, Park Geun-hye, and Moon Jae-in are examined within a multimodal framework. BERTopic identifies shifts in policy agendas across presidents and time periods, while audio embeddings from Emotion2Vec capture vocal patterns, including emotional expressions. The results reveal clear differences in speaking styles and notable variation over time. Cross-validation of textual sentiment classification using KoELECTRA with audio-based analysis reveals frequent discrepancies, underscoring the importance of integrating text and audio in political speech research. This study shows how multimodal computational approaches can advance the analysis of presidential speech and contribute to broader debates in political communication and executive politics.

**Keywords:** presidential archives, presidential speech, text-as-data, audio-as-data, multimodal analysis

## I. Introduction: Beyond the Written Text

On October 13, 2003, amid the political turmoil surrounding the special investigation into the North Korea remittance scandal, President Roh Moo-hyun took the podium of the National Assembly. The content

[*] I am grateful to two anonymous reviewers for their insightful feedback and suggestions. This article received the Outstanding Research Paper by the Presidential Archives of Korea in 2025.
[**] Ph.D. Candidate, Wilf Family Department of Politics, New York University; E-mail: kyusik.yang@nyu.edu

of his speech served as a political statement, an unprecedented proposal to place his political future in the hands of the legislature. Yet his vocal delivery expressed both firm determination and great distress. Thirteen years later, on October 24, 2016, President Park Geun-hye, amid a growing scandal involving undue influence over state affairs, attempted to shift the political landscape by unexpectedly proposing a constitutional revision in her budget speech to the National Assembly. While the text calmly outlined the institutional rationale for reform, her delivery was marked by a rigid, defensive intonation. Subtle vocal tremors indicated not confidence but anxiety and isolation under intense political pressure.

Presidential systems often create sharp tensions between the executive and the legislature (Linz 1990; Shugart and Carey 1992; Samuels and Shugart 2010; Jin and Yun 2024). These tensions are extreme during periods of divided government, when the president's party lacks a majority in the National Assembly. In these situations, pushing forward the presidential policy agenda requires overcoming significant institutional barriers to legislative cooperation. In this context, a president's appearance at the National Assembly podium is more than just a ceremonial routine (Campbell and Jamieson 1990; Cohen 1995; Kwon and Choi 2013; Kim 2014). Presidential addresses to the Assembly usually involve budget or policy speeches, but they also include inaugural addresses to a new legislature and special statements. On this political stage, presidents aim to persuade both legislators and the public. Sometimes, they try to confront and overcome institutional conflicts directly.

Existing research has primarily focused on presidential speeches as written texts, systematically analyzing policy agendas and ideological orientations through content analysis (Campbell and Jamieson 1990; Hur and Eom 2012; Moon and Lim 2020; Park and Yoo 2020; Rhyu and Kim 2021; Park et al. 2022; Kim and Han 2024). Recently, the adoption of advanced natural language processing techniques has further advanced the study of presidential discourse (Grimmer and Stewart 2013; Lucas et al. 2015). Despite these contributions, however, prior research has struggled to capture the inherently multimodal nature of political speech. Although textual analysis can reveal what presidents demand, it cannot fully explain how those demands are carried out (Dietrich et al. 2019; Knox and Lucas 2021). It is difficult to discern from text alone whether a president is pressuring the legislature with a strong tone, seeking compromise with a

respectful voice, or claiming legitimacy through neutral policy explanation. Moreover, presidential speaking styles vary across political contexts, and such contextual variation cannot be adequately captured using text-as-data alone. To fully understand presidential addresses as complex political actions, therefore, it is essential to complement textual analysis with audio-as-data through a multimodal approach.

This study departs from these concerns by undertaking an integrated analysis of presidential speeches delivered before the National Assembly and their corresponding audio recordings. To address the limitations of text-centered approaches, it employs BERTopic (Grootendorst 2022), a transformer-based topic modeling technique, to trace changes in key policy agendas over time. For the analysis of audio data, the study uses Emotion2Vec (Ma et al. 2024) to quantify emotional signals embedded in prosody, such as intonation, speech rate, and emphasis. In addition, by cross-validating sentiment analyses derived from a text-based KoELECTRA model and the audio-based Emotion2Vec, the study explores the strategic duality of presidential speech (Park and Yoo 2020; Knox and Lucas 2021).

The scholarly contributions of this study can be summarized in two respects. First, by adopting a multimodal approach that integrates text and audio, the study quantitatively examines dimensions of political communication beyond the written text that have been largely overlooked in existing research. Second, it reinterprets the classic executive–legislative relations using new data and methodological tools. This sheds light on strategic interactions that are often unseen when analysis is limited to textual data.

This article is organized as follows. It first reviews the institutional significance of presidential speeches before the National Assembly, along with the contributions and limitations of existing research. It then presents the research design and analytical methods based on textual and audio data, followed by a discussion of results. Finally, the article concludes by outlining the implications of this study for research on Korean politics and suggesting directions for future research.

## II. Theoretical Background

### *A. Presidential Addresses to the National Assembly and Executive-Legislative Relations*

Presidential addresses to the National Assembly are among the most symbolic institutional acts in which the head of the executive stands before the legislature. In this setting, the president presents a vision for state governance and requests cooperation, while simultaneously constituting and exercising power in relation to the National Assembly. Therefore, these speeches are not just explanations of policies; they can also be seen as strategic actions that produce political outcomes within an institutional context.

Just as the State of the Union in the United States has functioned as a key device for revealing the power relationship between the president and Congress (Proksch & Slapin 2012; 2015), Korea's budget and policy speeches to the National Assembly likewise represent moments in which negotiations and conflicts over the budget and the direction of governance are condensed. Presidential addresses to the National Assembly thus reflect the power configuration and political environment of a particular period (Kim 2014; Moon and Lim 2020). On this stage, informal interactions that are difficult to capture through legislative outputs, such as lawmaking or roll-call votes, become visible. As these scenes are transmitted through the media and made public, they serve as essential cues in shaping the substantive dynamics of executive-legislative relations.

### *B. Achievements and Limitations of Text-Based Analysis*

Presidential speeches have long been treated as a core object of analysis in political science. Speech texts have been widely used as key data for tracing presidents' policy priorities, patterns of interparty conflict, and shifting political agendas across historical periods (Hur and Eom 2012; Kim 2014; Park and Yoo 2020; Park et al. 2022). In particular, the introduction of computer-based text analysis methods since the 2000s has substantially expanded both the scope and precision of research on presidential speeches.

Lucas et al. (2015) established a methodological foundation for quantitative political text analysis by specifying standards for research design, validation, and interpretation. Roberts et al. (2014) proposed the Structural Topic Model (STM), which introduced a new analytical approach by incorporating metadata, such as speaker characteristics, time, and party affiliation, into text analysis. This enables the systematic tracking of temporal and administration-level shifts in presidential agendas. Together, these studies demonstrate that presidential speeches are political actions shaped by an institutional context, not just individual expressions (Proksch and Slapin 2012; 2015). Therefore, they provide the direct theoretical and methodological precedents for this study's analysis of presidential speeches.

Presidential speech analysis has also accumulated steadily in South Korea. Kang and Kim (2004) analyzed the rhetorical characteristics found in the speeches of successive presidents, while Kwon and Choi (2013) explored the political functions of presidential discourse by comparing language-based symbolic strategies across administrations. Kim (2014) examined patterns of presidential leadership and policy agenda-setting through content analysis, and Park and Yoo (2020) explored trends in regulatory reform policy reflected in presidential speeches using text mining and topic modeling techniques. Rhyu and Kim (2021) analyzed Korean nationalist discourse through text mining of presidential speeches, while Park et al. (2022) traced patterns of topic change in unification-related presidential addresses. Recently, Kim and Han (2024) further advanced the analysis of presidential discourse by applying large-scale text analysis to presidents' speeches. These studies have provided a comprehensive understanding of presidential rhetorical strategies and the evolution of political discourse over time. They have also demonstrated differences across administrations.

Additionally, text-centric analyses have evolved from basic topic classification to include new methods such as sentiment analysis and semantic network analysis. Park and Lee (2024) empirically examined patterns of aggressive language use by the two major parties through an analysis of party commentaries, while Cho (2024) linked the negativity of legislators' speech to political factors such as ideological extremity. These efforts demonstrate that research on presidential speeches has progressed

beyond the identification of policy agendas toward a more comprehensive analytical scope that includes political emotions, discursive strategies, and the structure of semantic networks.

Taken together, domestic and international studies have made significant progress in quantitatively analyzing presidential speeches to identify changes in policy agendas, rhetorical strategies, and emotional content in South Korea. However, these studies have primarily focused on *what* presidents said, offering limited insight into *how* they delivered those messages. In other words, these studies have not examined the attitudes presidents adopted toward the legislature or the emotional signals they conveyed when addressing it. As a result, existing research faces structural limitations in explaining the interactive dynamics of tension, cooperation, and conflict between the executive and the legislature. To overcome these limitations, this study seeks a new approach that moves beyond text-based analysis.

### *C. The Need for a Multimodal Approach*

Presidential speeches are political actions with both textual and auditory dimensions. Elements such as intonation, speaking rate, emphasis, and emotional expression are key factors that shape audience reception and political effects, beyond the content of words themselves (Grebelsky-Lichtman 2016; Dietrich et al. 2019). Nevertheless, research that systematically and quantitatively uses political leaders' vocal data remains limited worldwide. Attempts to analyze presidential speech leveraging audio-as-data are exceedingly rare in the study of Korean politics. While some studies have incorporated presidential vocal data, they have focused mainly on simple comparisons of pitch or speaking rate (Chung and Lee 2016; Chung and Nam 2016; Choi et al. 2016). These approaches alone are insufficient to explain the political nature of presidential speeches comprehensively.

Recent advances in speech analysis techniques based on artificial intelligence have opened up possibilities for overcoming these limitations. In particular, rapid developments in computing technology and increasingly active interdisciplinary exchange have expanded the window of opportunity for examining presidential speeches in a multidimensional manner. As a variety of methodologies have been developed,

including acoustic feature filtering, multimodal fusion, and self-supervised embedding approaches (Ververidis and Kotropoulos 2006; Hashem et al. 2023; Ma et al. 2024; Kim et al. 2021; Kim and Lee 2021; Shin and Hong 2023; Kim et al. 2024; Lim et al. 2024; Lee 2025), a foundation has been established for quantitatively exploring the emotional and political dimensions of speech.

Accordingly, this study analyzes presidential addresses to the National Assembly simultaneously at two levels: speech as text (*text-as-data*) and speech as voice (*audio-as-data*). Through this approach, it seeks to identify how signals of tension, cooperation, and conflict are conveyed alongside the presentation of policy agendas. More broadly, a multimodal approach that combines textual and audio data offers a new analytical framework for the study of presidential discourse by capturing both the interaction and the disjunction between these two data. For example, if text analysis reveals conflict-oriented vocabulary but audio data indicates cooperative intonation, then the presidential speech could be interpreted as performing contradictory strategies on the surface and underlying levels simultaneously. Analysis of such mismatches can be further extended into a multilayered examination of political effects when combined with additional data sources, such as legislative responses, media coverage, and public opinion surveys.

Such an approach carries scholarly significance in the institutional context of South Korea's presidential system. The inherent tension of presidentialism, combined with divided government, makes the political messages conveyed through the delivery of speech especially important. Accordingly, a multimodal approach not only illuminates the distinctive features of Korean politics but also holds potential for extension to cross-national comparative research. Although this type of analysis cannot directly resolve political conflicts, it can provide valuable insights into the patterns of interaction and the origins of conflict that serve as its starting point. The following sections first detail the data and methodology used to analyze presidential addresses to the National Assembly and then present the results.

## III. Data and Methodology

This study takes a multimodal approach, encompassing the textual and audio dimensions of presidential addresses to the National Assembly. The research design consists of three stages. First, a sentence-level parallel corpus is constructed so that individual presidential speeches to the National Assembly can be analyzed simultaneously in both textual and audio forms. Second, independent analyses are conducted on text and audio data. The text data are used to extract policy agendas and rhetorical structures. The audio data are used to derive speaking styles, as well as vocal and emotional characteristics. Third, the two analytical approaches are integrated to measure the interaction and disjunction between "speech as text" and "speech as voice," followed by an attempt to provide political interpretations of these patterns.

### *A. Text-as-Data and Audio-as-Data in Presidential Addresses to the National Assembly*

This study analyzes twenty presidential addresses delivered by Presidents Roh Moo-hyun, Lee Myung-bak, Park Geun-hye, and Moon Jae-in on the floor of the National Assembly between 2003 and 2021. Full-text transcripts of the speeches were obtained from the official archive of the Presidential Archives of Korea (Records Collection > Speech Records > Speech Texts). Corresponding audio data were collected as follows. For two speeches by President Roh Moo-hyun (June 7, 2004; February 25, 2005), audio was extracted from the Presidential Archives' video archive (Records Collection > Speech Records > Speech Videos). For the remaining eighteen speeches, audio data were processed from video recordings available through the National Assembly of the Republic of Korea's VOD Service for Conferences.[1] Data that allow for the simultaneous acquisition and direct comparison of text and audio from the same

[1] Detailed information on the timing of each speech, the associated political issues is provided in Table 2.

presidential speech are rare in Korean political research. As such, this dataset provides an important foundation for multimodal analysis.

Each president's speeches were selected to cover major political milestones and distinct institutional contexts: early-term policy addresses, budget speeches during divided government, and politically critical moments, such as presidential crises. This selection strategy prioritizes variation in political context over uniform temporal spacing, ensuring that the corpus captures the range of rhetorical situations presidents face at the National Assembly. As summarized in Table 2, the selected speeches span a range of political conditions, from legislative cooperation to constitutional crisis.

Each speech video was first extracted into an audio file and then processed with an automatic speech recognition (ASR) system using the CLOVA Speech service on the Naver Cloud Platform. CLOVA Speech is a service based on CLOVA's NEST (Neural End-to-End Speech Transcriber) technology that converts audio files into sentence-level text. In this study, the service was used to obtain foundational data that allows for precise alignment between the audio and textual components of presidential speeches.[2]

The officially released presidential speech transcripts were then compared with the outputs generated by CLOVA Speech to verify the start and end points of each sentence, and precise timestamps were assigned at the millisecond (ms) level. In particular, the syllable-level segmentation results from the JSON files generated by CLOVA Speech were used to improve sentence-level alignment accuracy.[3] The final text-audio parallel corpus was constructed such that each unit consisted of the following elements: president, date of speech, sentence number, sentence text, the corresponding sentence-level audio file, and start and end timestamps. This structure establishes a basis for directly comparing and validating the same sentence in textual and audio formats. The complete corpus comprises 3,900 individual sentence-level observations.

[2] For an overview of the CLOVA Speech model and details, see the user guide provided by the Naver Cloud Platform. (https://guide.ncloud-docs.com/docs/en/clovaspeech-overview)
[3] In this study, audio data were loaded using Python's *librosa* library and preprocessed into a form suitable for analysis through a normalization procedure (McFee et al. 2025).

The variation in the number of speeches per president (3-7) reflects differences in tenure length and in the frequency with which each president addressed the National Assembly: Moon Jae-in's seven speeches span six budget cycles plus the opening of the 21st National Assembly, while Lee Myung-bak's three speeches are concentrated in the first half of his term. To assess whether this imbalance affects comparability, I checked that the per-speech sentence counts are broadly similar across presidents, suggesting that differences in corpus size reflect differences in the number of available speeches rather than in speech length. Results are reported as proportional distributions (e.g., topic shares, emotional profiles) rather than raw counts, which mitigates but does not eliminate the comparability concern raised by unequal N. Nonetheless, the primary analytical focus is on within-presidency variation and proportional comparisons across presidents, neither of which requires equal sample sizes.

***Table 1.*** *Descriptive Statistics of Presidential Addresses to the National Assembly*

| President | N (Speeches) | N (Sentences) | Sentence range (per speech) | Period |
|---|---|---|---|---|
| Roh Moo-hyun | 4 | 1,127 | 204-359 | 2003-2005 |
| Lee Myung-bak | 3 | 530 | 135-217 | 2008-2012 |
| Park Geun-hye | 6 | 798 | 100-172 | 2013-2016 |
| Moon Jae-in | 7 | 1,445 | 166-241 | 2017-2021 |
| **Total** | 20 | 3,900 | 100-359 | 2003-2021 |

***Table 2.*** *Presidential Addresses to the National Assembly and Major Political Issues*

| President | Date | Type of Address | Major Political Issues |
|---|---|---|---|
| Roh Moo-hyun | 2003-04-02 | Address on State Affairs | Controversy over troop deployment to the Iraq War |
| Roh Moo-hyun | 2003-10-13 | 2004 Budget Address | Special prosecutor probe into North Korea remittance; economic uncertainty |
| Roh Moo-hyun | 2004-06-07 | Opening of the 17th National Assembly | Immediately following the impeachment crisis |
| Roh Moo-hyun | 2005-02-25 | Address on State Affairs | North Korea's declaration of nuclear weapons possession |
| Lee Myung-bak | 2008-07-11 | Opening of the 18th National Assembly | Candlelight protests over U.S. beef import; high oil prices |

| | | | |
|---|---|---|---|
| Lee Myung-bak | 2008-10-27 | 2009 Budget Address | Global financial crisis |
| Lee Myung-bak | 2012-07-02 | Opening of the 19th National Assembly | End-of-term lame-duck; demands for economic democratization |
| Park Geun-hye | 2013-11-18 | 2014 Budget Address | Controversy over retreat from the basic pension pledge |
| Park Geun-hye | 2014-10-29 | 2015 Budget Address | Follow-up measures after the Sewol ferry disaster |
| Park Geun-hye | 2015-10-27 | 2016 Budget Address | Controversy over state-authored history textbooks |
| Park Geun-hye | 2016-02-16 | Address on State Affairs | North Korea's 4th nuclear test; missile launch |
| Park Geun-hye | 2016-06-13 | Opening of the 20th National Assembly | Divided government (ruling party in minority) |
| Park Geun-hye | 2016-10-24 | 2017 Budget Address | Escalation of allegations in the Choi Soon-sil scandal |
| Moon Jae-in | 2017-06-12 | 2017 Supplementary Budget Address | Launch of a new administration; employment issues |
| Moon Jae-in | 2017-11-01 | 2018 Budget Address | Debate over income-led growth |
| Moon Jae-in | 2018-11-01 | 2019 Budget Address | Korean Peninsula peace process |
| Moon Jae-in | 2019-10-22 | 2020 Budget Address | Cho Kuk controversy; Japan's export restrictions |
| Moon Jae-in | 2020-07-16 | Opening of the 21st National Assembly | COVID-19 pandemic |
| Moon Jae-in | 2020-10-28 | 2021 Budget Address | Prolonged COVID-19 pandemic |
| Moon Jae-in | 2021-10-25 | 2022 Budget Address | Gradual return to normalcy ("With COVID") |

### *B. Text Analysis Using BERTopic*

To identify the policy agendas and rhetorical structures embedded in presidential addresses to the National Assembly, this study employs BERTopic, a recently adopted topic modeling technique leveraging Bidirectional Encoder Representations from Transformers (Grootendorst 2022). Frequency-based analyses and traditional topic modeling approaches, most notably Latent Dirichlet Allocation (LDA), have been widely used in political science research. These approaches have contributed to the quantitative analysis of presidential discourse; however, they have limitations in adequately capturing contextual meaning. Because LDA is based on the bag-of-words assumption and does not consider word order or syntactic structure, it is ideal for examining the overall theme of longer documents. However, it is not suitable for capturing the nuances of context in individual sentences.

In contrast, BERTopic employs Sentence-BERT-based embeddings to transform speech texts at the sentence level and to group semantically similar sentences into topics.[4] This approach enables visualization of changes in the relative prevalence of major topics over time and the extraction of representative sentences for each topic, thereby allowing us to identify which policy agendas and messages were emphasized by different presidents and across different periods.

The number of topics was determined through an iterative reduction procedure. BERTopic initially identified 160 topics from the 3,900-sentence corpus; these were reduced to 30 through hierarchical topic merging, balancing granularity with interpretability. To assess the coherence of the resulting topics, I computed coherence scores, yielding a mean coherence of 0.39 across the 29 non-outlier topics (Röder et al. 2015).[5] The highest-coherence topics ($C_v$ > 0.50), including formal address conventions, inter-Korean affairs, and welfare support, are substantively interpretable and consistent with the thematic labels in Table 3. A random baseline (shuffled topic assignments, 100 iterations) yielded a mean $C_v$ of 0.32, significantly lower than the observed 0.39 ($p < 0.001$), confirming that the topic assignments capture meaningful lexical co-occurrence patterns above chance. The proportion of sentences classified as outliers was 31.6% (N = 1,231), consistent with BERTopic behavior on short-text corpora (see Appendix B for per-topic coherence scores and validation details).

The analysis confirms that the core issues addressed in presidential speeches to the National Assembly varied in accordance with the political and economic conditions of specific periods and the governing priorities of different administrations.

Nevertheless, these analyses are limited to presidential speeches in written form. In other words, while they reveal *what* presidents conveyed to the legislature, they do not explain *how* those messages were

[4] This study uses the *ko-sroberta-multitask* model, which is specialized for Korean sentence representation learning (Ham et al. 2020; Reimers and Gurevych 2019; 2020). This model is based on the Sentence-BERT architecture and is trained using a multitask learning framework, and it is known to exhibit strong performance in Korean natural language understanding.

[5] $C_v$ is a topic coherence metric that measures how frequently a topic's top keywords co-occur within a sliding text window, using normalized pointwise mutual information and cosine similarity. Among existing coherence measures, $C_v$ achieves the highest correlation with human judgments of topic quality (Röder et al. 2015).

delivered. Still, text-based analysis remains important because it has quantitatively expanded upon previous research and established a systematic approach to the policies and discourses of presidential speeches. To achieve a more comprehensive understanding of executive-legislative relations, however, analysis must move beyond text and incorporate a new dimension of data: the president's voice and modes of speech delivery.

### *C. Audio Embedding Analysis*

To address the limitations of text-based analysis, this study uses Emotion2Vec, a cutting-edge audio embedding method (Ma et al. 2024). Emotion2Vec is a model that focuses on the acoustic and emotional characteristics of speech rather than linguistic meaning. It has demonstrated strong performance in speech emotion classification and vocal expression analysis across diverse languages and environments. However, publicly available pretrained models lack data from Korean speakers, which limits their direct applicability to the analysis of Korean political speeches. To address this limitation, this study further fine-tunes Emotion2Vec using the "Conversational Speech Dataset for Emotion Classification" released on AI Hub by the National Information Society Agency (NIA).[6] Training was conducted in a Google Colab environment equipped with an A100 GPU, and the fine-tuned model was applied to sentence-level audio files from presidential speeches. This enabled the quantification and visualization of speaking styles, intonation, and emotional patterns within an embedding space.

Traditional acoustic approaches, such as waveform analysis, Mel-spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs), provide intuitive representations of speech intensity, rate, and intonation. Nevertheless, they are limited in their ability to explain speech within a political context quantitatively. Existing audio embedding models such as HuBERT (Hsu et al. 2021) and Wav2Vec 2.0 (Baevski et al. 2020; Kim and Kang 2021) are primarily optimized for speech recognition tasks. In contrast,

[6] This dataset consists of Korean conversational speech and includes a total of 43,975 utterances.

Emotion2Vec is designed to efficiently learn nonverbal prosodic features, such as intonation, rhythm, and emphasis, as well as emotional signals. This makes it particularly well-suited for analyzing political speeches, such as presidential addresses.

This study extracted 768-dimensional embedding vectors from sentence-segmented presidential speech audio using Emotion2Vec. These embeddings were reduced to two dimensions and visualized using UMAP (Unified Manifold Approximation and Projection). Then, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) was applied to derive natural clusters from the underlying data distribution.

### *D. Multimodal Emotion Analysis: Mismatches Between Text and Audio*

This study conducts an additional multimodal analysis by simultaneously examining presidential addresses to the National Assembly through two channels, text and audio, and cross-validating the results derived from each. For text analysis, KoELECTRA (Park 2020), which has been validated on Korean natural language understanding tasks, was employed to classify each sentence into three categories: positive, neutral, and negative. For audio analysis, the aforementioned embedding model, Emotion2Vec (Ma et al. 2024), was applied. Both models were fine-tuned on the NIA AI Hub Korean Conversational Speech corpus (N = 43,975 utterances; 7-class emotion labels) and evaluated on held-out stratified test sets. KoELECTRA achieved a 7-class macro-F1 of 0.92; Emotion2Vec achieved a 3-class macro-F1 of 0.69, with F1 = 0.90 for the negative class, which is the dominant valence category in this analysis. The performance gap reflects the inherent difficulty of inferring emotion from acoustic properties alone compared to text-based classification (see Appendix A for full model specifications, per-class performance, and discussion of the cross-modal performance asymmetry).

Subsequently, sentiment classification results from text and audio were compared for the same sentence-level utterances to analyze patterns of agreement and mismatch. Rather than limiting the analysis to a simple accuracy comparison between the two approaches, this study systematically examined transition

patterns across specific emotional categories using a confusion matrix. The results indicate that both models tended to classify presidential speeches as negative utterances, suggesting the need for supplementary training that better reflects the distinctive characteristics of political speech. At the same time, cases in which the text and audio diverge can be interpreted as more than just technical errors. They can also be seen as potential clues to the strategic rhetorical choices presidents make. These mismatches thus offer important implications for subsequent research.

## IV. Results

### *A. Text-Based Analysis Results*

Figure 1 visualizes year-by-year changes in the relative proportions of the top ten topics, while Table 3 presents the thematic labels and representative sentences for each topic. During the Roh Moo-hyun administration, topic distributions appear relatively balanced overall; however, a notably high share of Topic 7 (Politics & Conflict) distinguishes this period from other presidencies. This pattern reflects a substantive tendency to address political conflict directly and attempt to break through it rhetorically. In the case of President Lee Myung-bak, the 2008 speeches show a relatively even distribution across Foreign Affairs & Security, Economy & Industry, and Politics & Communication. In contrast, toward the end of his term in 2012, the relative weight of Foreign Affairs & Security increases. President Park Geun-hye's speeches primarily emphasize Economy & Industry, alongside a consistent treatment of Society & Welfare. During the 2016 political crisis, however, Foreign Affairs & Security became particularly salient. President Moon Jae-in places greater emphasis on Society & Welfare early in his term.

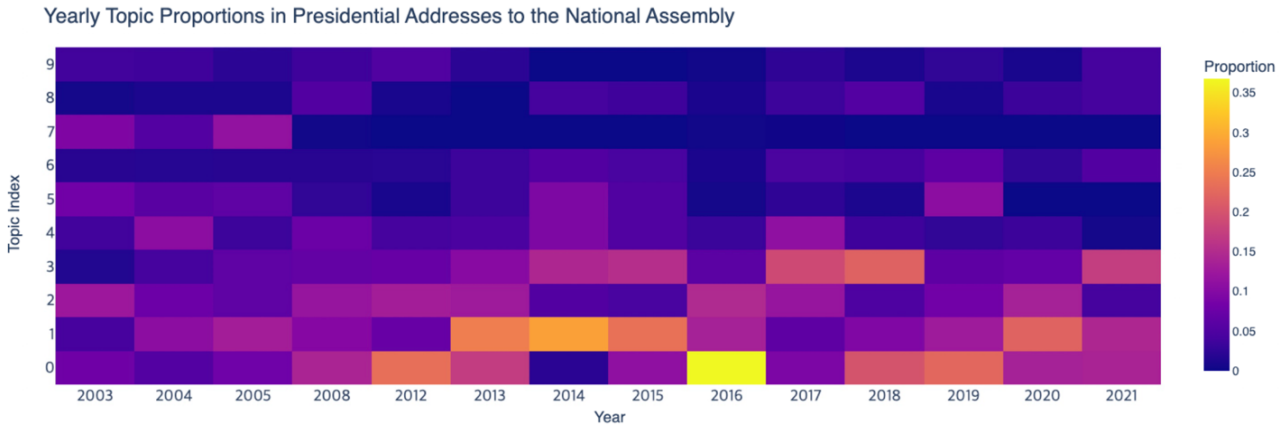


***Figure 1.*** *Results from the BERTopic Analysis*

At the same time, from 2018 onward, the relative prominence of topics in Foreign Affairs & Security and Economy & Industry expands. This shift reflects the diplomatic context of the period, including inter-Korean summits. These patterns suggest that presidential addresses to the National Assembly reflect core policy agendas that are important at specific times.

***Table 3.*** *Major Topics and Representative Sentences*

| Topic | Theme | Representative Sentence 1 | Representative Sentence 2 |
|---|---|---|---|
| **0** | **Foreign Affairs & Security** | Going forward, the government will promote the Korean Peninsula trust process based on firm principles and patience.<br>(Park Geun-hye, 2013-11-18) | Through dialogue and diplomacy, we will continue to strive until a new order that brings peace and prosperity to the Korean Peninsula is established.<br>(Moon Jae-in, 2021-10-25) |
| **1** | **Economy & Industry** | Technological development aimed at improving energy efficiency will create new markets and jobs, becoming a new driving force for economic growth.<br>(Lee Myung-bak, 2008-07-11) | The government will support for key industries, including K-semiconductors, K-batteries, K-biotechnology, K-hydrogen, and K-shipbuilding, through sector-specific strategic assistance.<br>(Moon Jae-in, 2021-10-25) |
| **2** | **Politics & Communication** | I believed that the nation could develop and its people can become happier only when our politics changes.<br>(Roh Moo-hyun, 2003-10-13) | I sincerely ask for your cooperation so that these bills can be passed during the current session of the National Assembly.<br>(Park Geun-hye, 2013-11-18) |

| | | | |
|---|---|---|---|
| **3** | **Society & Welfare** | We will ensure that low-income groups can secure a minimum standard of living and, through this, be supported toward self-reliance. (Park Geun-hye, 2015-10-27) | By increasing housing and education benefits, we will make the basic livelihood security benefits a reality. (Moon Jae-in, 2017-11-01) |
| **7** | **Politics & Conflict** | Even when we oppose one another and engage in confrontation, we must compete fairly and honorably in accordance with the norms and principles of democracy. (Roh Moo-hyun, 2003-04-02) | Our politics has become trapped in a vicious cycle in which, from the day after a presidential election, extreme confrontation and antagonism dominate daily politics. (Park Geun-hye, 2016-10-24) |

This study used BERTopic to extract the policy agendas and rhetorical structures of presidential speeches with precision at the sentence level. This made it possible to quantitatively identify differences in topic distribution across years and presidents and to systematize the political messages reflected in "written words" by deriving representative sentences for each topic. However, this approach is limited to the textual dimension. That is, while it reveals what presidents said to the National Assembly, it does not capture additional forms of political interaction embedded in vocal intonation and emotional signals. The following section presents an analysis of audio embedding to quantitatively explore the vocal characteristics of presidential speeches.

### *B. Audio-Based Analysis Results*

Before conducting the complete embedding-based analysis, I visualized selected segments of presidential speeches using Mel spectrograms (Figure 2).[7] Figure 2 presents a comparison of Mel

[7] A Mel spectrogram shows the distribution of energy across frequency bands over time, enabling comparison of basic acoustic characteristics such as intensity, intonation, and speech rhythm. Mel spectrograms are generated by decomposing audio signals into frequency components through a standard Fast Fourier Transform (FFT) and then remapping these components onto the Mel scale. Through this transformation, the frequency axis of the spectrogram is expressed in units that are closer to human auditory perception than raw physical frequencies (Rabiner and Schafer 2011).

spectrograms for the opening sentences of presidential addresses delivered by President Park Geun-hye on November 18, 2013, and President Moon Jae-in on June 12, 2017. In both cases, the excerpts correspond to the presidents' initial greetings to the National Assembly, beginning with a formal address to citizens, the Speaker, and members of the legislature. President Park opened her address by stating, *"Distinguished citizens, Speaker Kang Chang-hee, and honorable members of the National Assembly, I am pleased to deliver a budget address here in this hall of the people…"*, while President Moon began with, *"Distinguished citizens, Speaker Chung Sye-kyun, and honorable members of the National Assembly, it was right here that I delivered a party leader's address during the 19th National Assembly."*

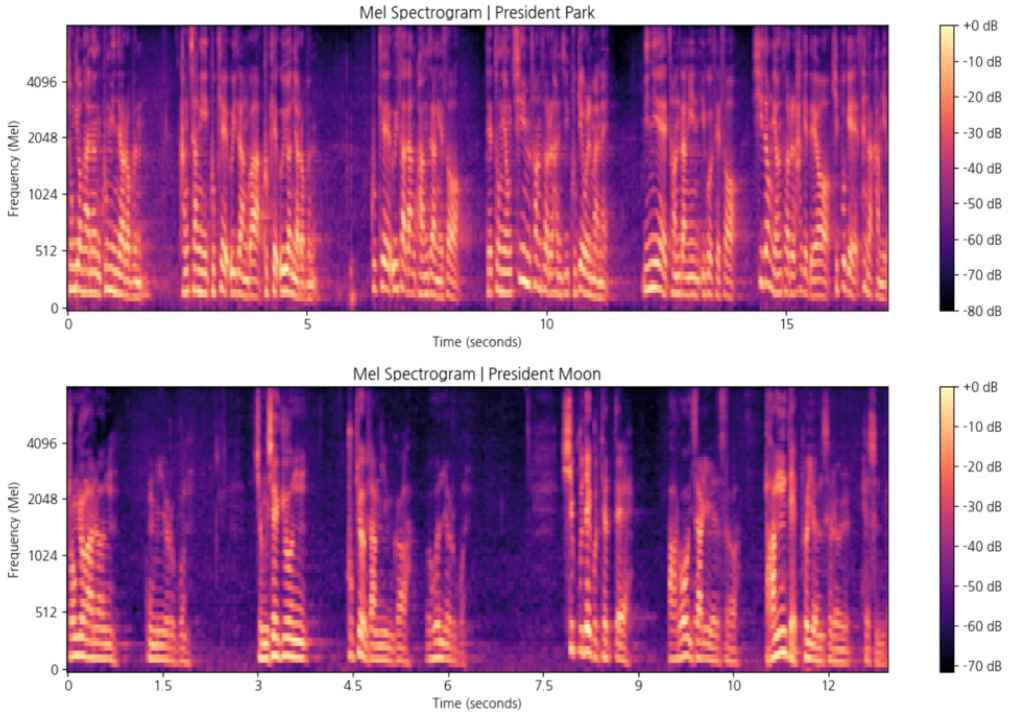


***Figure 2.*** *Comparison of Mel Spectrograms*

This comparison highlights differences in vocal delivery rather than in content, holding the institutional context and the rhetorical function of the utterance constant, namely, the first sentence of a presidential address to the National Assembly. The Mel spectrograms indicate relatively recurrent frequency patterns in President Park's speech, whereas President Moon's speech shows greater dispersion

and temporal variation in energy distribution across frequency bands. These visual differences underscore how presidential speech can convey distinct vocal styles even within standardized institutional settings.

Mel spectrograms have thus been widely used in prior research to intuitively examine differences in vocal characteristics. However, visualizing physical acoustic signals alone makes it challenging to interpret a speaker's strategic intent or emotional signals. In other words, the fact that certain patterns appear repetitive or stable in specific segments cannot, by itself, be interpreted as evidence of political communicative effects. While Mel spectrograms provide a useful starting point for comparing the acoustic properties of presidential speeches, analyzing speech strategies and their political implications requires a quantitative, embedding-based approach.

In this study, Emotion2Vec was used to extract 768-dimensional audio embeddings from individual sentence-level utterances in presidential addresses to the National Assembly. These embeddings were subsequently projected into a two-dimensional space using UMAP for visualization and clustered using HDBSCAN. The results reveal clear sentence-level variation in intonation and emotional tone across presidential speeches (Figure 3). However, clustering at the aggregate level alone has limitations in capturing how individual presidents strategically utilized their addresses to the National Assembly in political terms. Accordingly, this study visualizes each president's speeches separately at the sentence level (Figures 4 and 5). This approach enables us to examine how distinctive vocal characteristics vary across presidents. It provides a basis for inferring how presidential addresses to the National Assembly were used as political instruments in specific contexts.

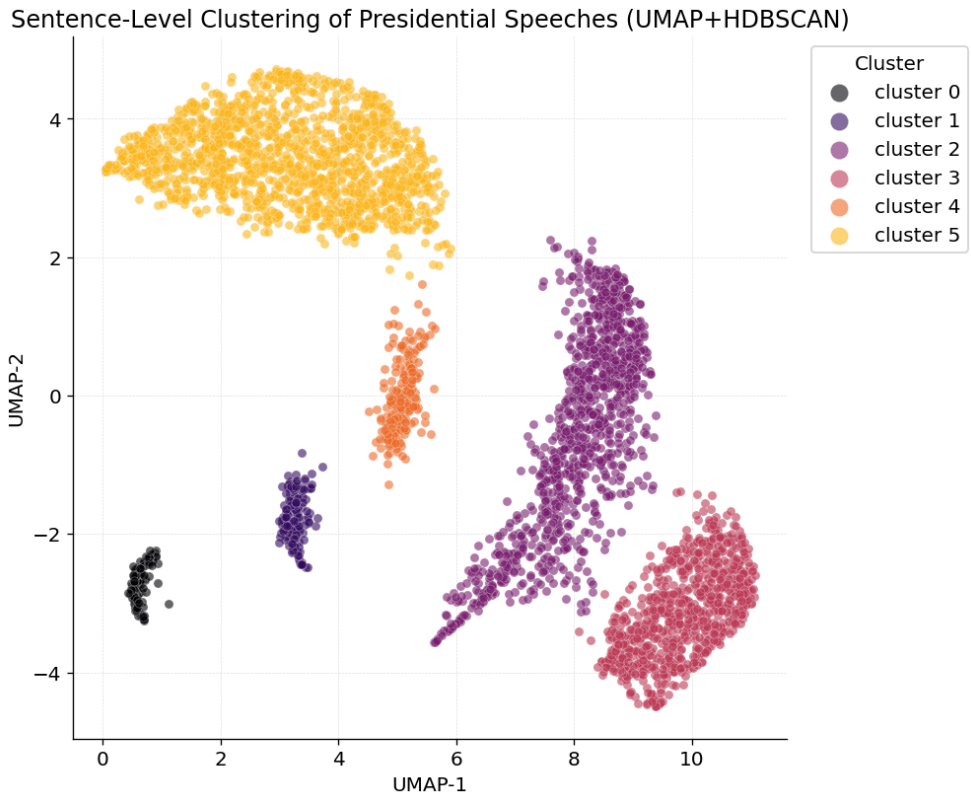


***Figure 3.*** *Audio Embedding Clustering*

At the same time, because the horizontal and vertical axes produced through UMAP dimensionality reduction represent simplified projections of high-dimensional embedding vectors, it is difficult to specify precisely which acoustic or emotional features each axis reflects. In general, the horizontal axis appears to capture speaker-specific vocal characteristics, such as habitual speaking patterns and fundamental frequency range, while the vertical axis appears to reflect emotional dimensions related to intonation and emphasis. These interpretations are tentative, as UMAP dimensions do not map directly onto acoustic features. Nonetheless, they are consistent with the design of Emotion2Vec, whose self-supervised pre-training objective organizes speech representations along dimensions of speaker identity and emotional prosody (Ma et al. 2024).[8]

[8] A full acoustic validation, correlating UMAP dimensions with extracted features such as pitch (F0), intensity (dB), and speaking rate, would provide stronger grounding for these interpretations. I treat this as a direction for future work, noting that the aggregate clustering patterns across presidents are consistent with the interpretive logic described above.

Note that the distribution along the horizontal axis shows that President Park Geun-hye's utterances are positioned farther from the center than those of other presidents. This pattern suggests that differences in frequency range, which vary systematically between male and female speakers, may be reflected in the embedding space. Emotion2Vec, a self-supervised model trained on speech audio, encodes acoustic properties, including pitch and timbre, that carry speaker-identity and gender-related information alongside emotional content (Ma et al. 2024). Whether the observed separation reflects gender-based acoustic differences, President Park's distinctive, restrained speaking style, or a combination of both cannot be determined from the current analysis alone. Importantly, this separation does not invalidate the cross-presidential comparisons: because the analysis focuses on variation within each president's speeches and relative positional patterns across the embedding space, not absolute distances, the gender-related displacement does not introduce systematic bias into the political stylistic interpretations.

In the case of President Roh Moo-hyun, differences across speeches are relatively limited compared to those of other presidents; however, dispersion along the vertical axis within individual speeches is comparatively pronounced (Figure 4). This pattern suggests that the president actively used addresses to the National Assembly as a political action, adjusting vocal tone and intonation in response to situational demands to maximize rhetorical effect. The political context, including periods of heightened tension and conflict between the executive and legislative branches, may influence such variation in speaking style.

By contrast, President Lee Myung-bak's utterances form relatively cohesive, small-scale clusters within individual speeches, while exhibiting clear separation across different speech occasions (Figure 4). This pattern shows that President Lee's speech remained consistent in tone while producing different vocal patterns in response to changing political situations. In other words, variation in his speech seems to have occurred primarily between speeches rather than within them. This indicates a strategy where vocal delivery was adapted to larger political moments rather than through detailed adjustments within a single speech.

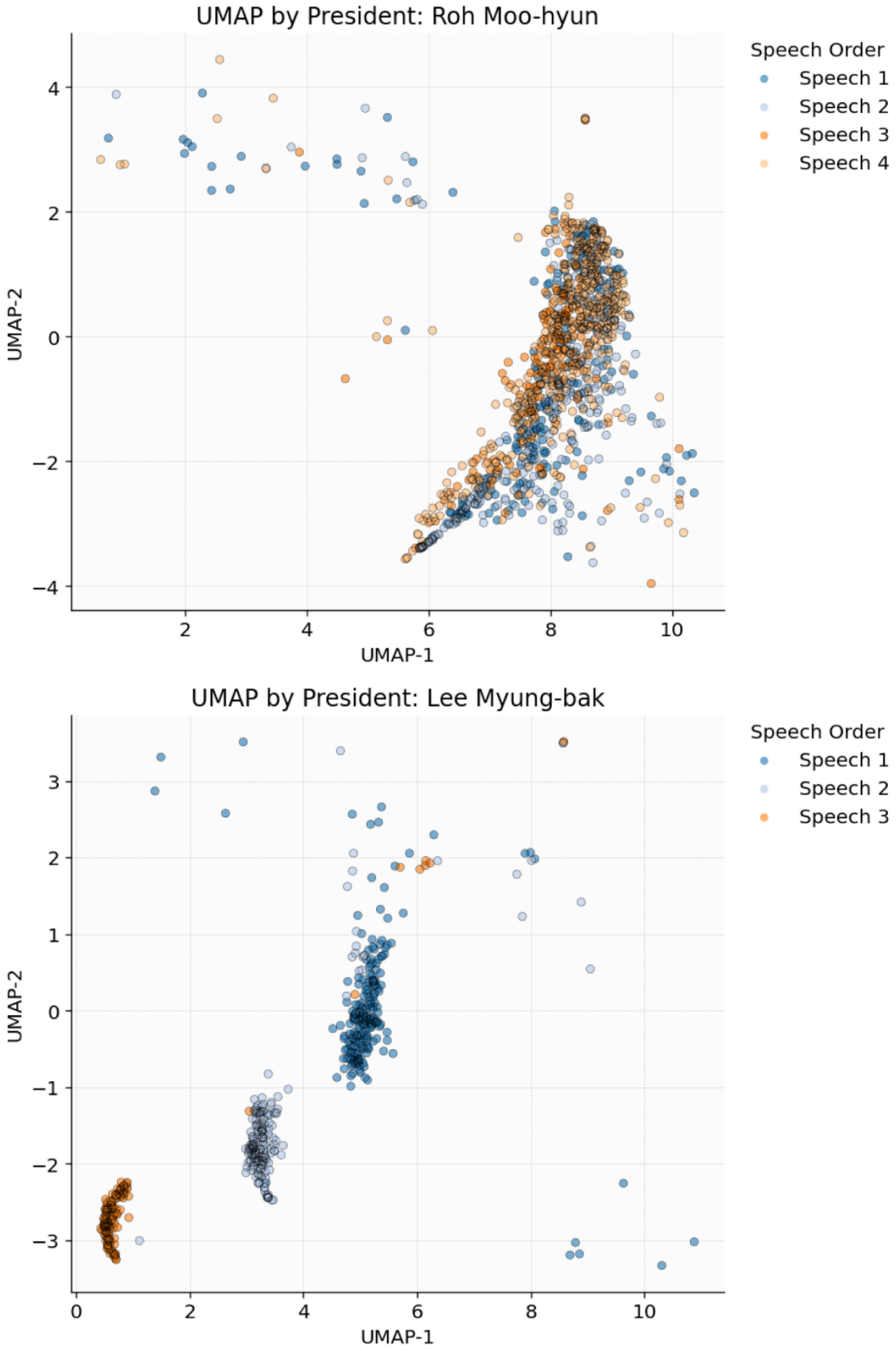


***Figure 4.*** *Clustering by President and Address (Presidents Roh Moo-hyun and Lee Myung-bak)*

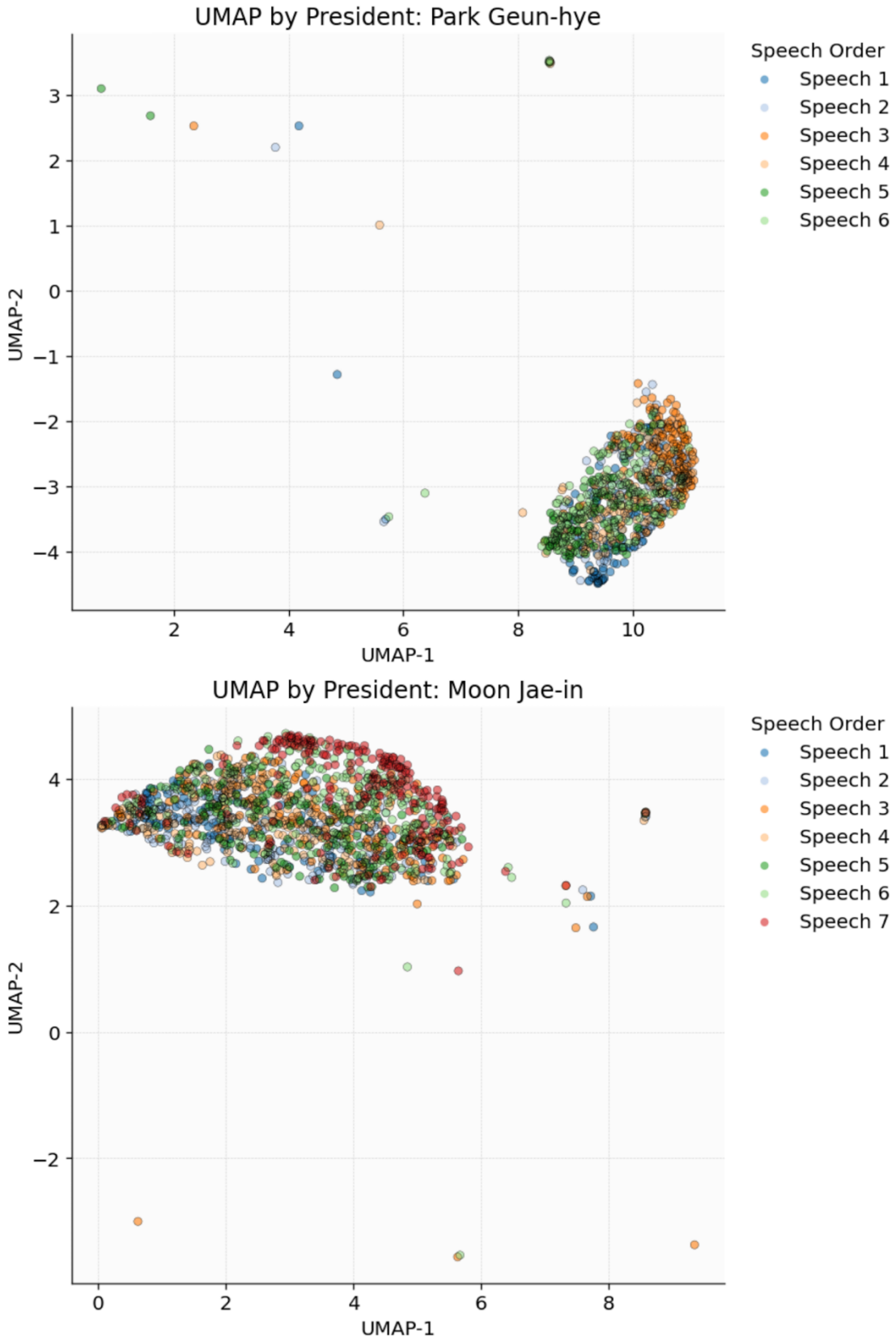


***Figure 5.*** *Clustering by President and Address (Presidents Park Geun-hye and Moon Jae-in)*

President Park Geun-hye's utterances are concentrated within one cohesive cluster with relatively little dispersion (Figure 5). This pattern suggests that her addresses to the National Assembly were delivered comparatively calmly and with minimal emotion. In other words, the limited variation in vocal tone and intonation suggests the speeches were delivered with tight control and stability. This type of clustering aligns with a communication style that conveys rhetorical emphasis through the structured presentation of policy content rather than through vocal modulation. This style is often observed when prepared remarks are delivered with limited real-time interaction with the audience.

President Moon Jae-in's statements generally maintain a consistent, policy-oriented baseline while exhibiting variation in embedded representations depending on the characteristics of the policy issues addressed in each speech (Figure 5). Notably, dispersion along the horizontal axis is relatively pronounced within individual speeches. Unlike President Roh Moo-hyun, who displayed emotional variation along the vertical axis, President Moon structured variation in his speech through the vocal dimension on the horizontal axis in a less overtly emotional manner.

This pattern suggests that President Moon's addresses to the National Assembly were fundamentally grounded in a stable framework of policy explanation. At the same time, his delivery could be seen as him strategically modulating his intonation and vocal emphasis to accompany the presentation of policy materials and explanations. President Moon achieved this by deliberately restraining his emotional expression.

### *C. Multimodal Analysis Results*

As noted earlier, presidential addresses to the National Assembly constitute political speech that simultaneously operates along two dimensions: not only *what* is said, but also *how* it is said. This study examines the extent to which results from independent text and audio analyses align or diverge, in order to assess the complementary nature of the two approaches. To this end, sentence-level emotion analysis of the same presidential addresses was conducted separately using text-based and audio-based models.

The confusion matrix comparing text-based and audio-based emotion classification results showed an overall agreement rate of 68.7% across all 3,900 sentence-audio pairs in the corpus (Figure 6). To contextualize this figure, I report two baselines. First, a baseline that predicts the most common class ('negative') for all observations would achieve an agreement rate of 80.8%, which exceeds the observed rate, confirming that the two models do not simply agree by always choosing the dominant category. Second, a chance baseline that preserves each model's observed marginal distributions (KoELECTRA: 80.8% negative; Emotion2Vec: 82.4% negative) yields an expected agreement of 68.6%, nearly identical to the observed rate. This indicates that the models' overall agreement is entirely explained by their shared tendency to classify political speech as negative, and that meaningful signal is carried in the pattern of disagreement rather than in aggregate agreement.

Two mechanisms plausibly explain this concentration of negative classifications. First, the fine-tuning corpus for Emotion2Vec (NIA AI Hub Korean Conversational Speech, N = 43,975 utterances) is itself 78.2% negative under the 3-class mapping used in this study, reflecting the emotional profile of everyday Korean conversational speech rather than formal institutional address. This training-data imbalance directly predisposes the model toward negative predictions, independent of the content of the speech being analyzed. Second, political speech genuinely differs from everyday conversation: presidential addresses frequently invoke crisis, criticize opposing forces, and make urgent demands for legislative action, rhetorical functions with a structurally negative character. Distinguishing these two mechanisms precisely would require a domain-matched validation corpus of Korean political speech with human-annotated sentiment labels, a resource that does not yet exist. For this study, the key implication is that inter-presidential comparisons remain interpretively valid on a relative basis: if both models apply the same domain-level bias uniformly, systematic differences in negative rates across presidents reflect genuine variation in rhetorical strategy rather than differential model error.

This baseline analysis reveals the core interpretive logic of the multimodal comparison. Both models apply the same domain-level bias when processing political speech: they over-predict 'negative' relative to conversational baselines, which means that aggregate agreement tells us nothing about genuine

convergence on emotional content. What carries interpretive value is the pattern of disagreement: the 1,222 mismatch cases (31.3% of pairs) where text and audio models diverge beyond what shared negative bias would produce. These are cases where the text channel signals one emotional register while the vocal delivery signals another. The per-president mismatch patterns (Roh 38.0%, Lee 28.9%, Park 31.3%, Moon 27.1%) are consistent with the substantive interpretations offered in the analysis: the variation across presidents reflects differences in rhetorical strategy rather than differential model noise. A chi-square test confirms that these differences are statistically significant ($\chi^2(3) = 36.89$, $p < 0.001$), indicating that mismatch rates vary systematically across presidents rather than randomly.

One concern with cross-modal comparison is that the two models employ different categorical frameworks (7-class text, 7-class audio with partially overlapping but non-identical labels). To address this, all analyses use a common 3-class mapping (negative, positive, neutral) that harmonizes the two label sets: KoELECTRA's 7 classes (happiness, neutral, sadness, anger, surprise, disgust, fear) and Emotion2Vec's training labels. This 3-class reduction resolves the cross-modal category-alignment concern while preserving the primary dimension of interest: the valence structure of presidential speech.

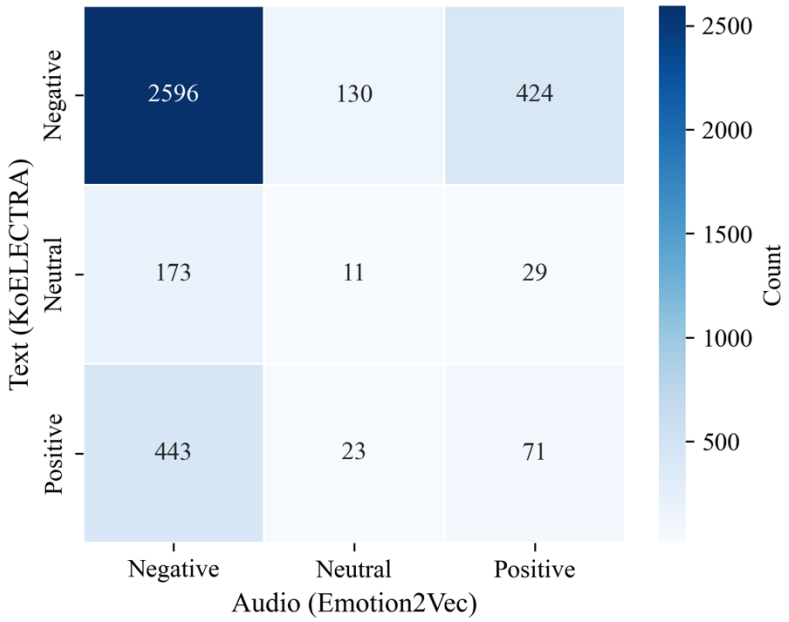


***Figure 6.*** *Text–Audio Emotion Classification Confusion Matrix*

An important political discovery is that text-based and audio-based models tend to produce divergent classification outcomes for specific presidential addresses to the National Assembly. For example, statements such as *"If increasing the number of National Assembly members would help overcome regional divisions, then we must do so to resolve them"* (President Roh Moo-hyun, February 25, 2005); *"At present, countries around the world are demonstrating political maturity by responding to the financial crisis in a bipartisan and agile manner"* (President Lee Myung-bak, October 27, 2008); *"To revitalize domestic demand, we must steadily expand corporate investment through regulatory reform"* (President Park Geun-hye, October 29, 2014); and *"We will place even greater emphasis on protecting vulnerable groups and investing in people"* (President Moon Jae-in, October 28, 2020) were classified as negative by the text-based model but as positive by the audio-based model. While the linguistic content of these statements centers on diagnosing crises or identifying policy challenges, the delivery was marked by confident and assertive vocal cues that led the audio model to interpret them as positive signals. In the Emotion2Vec framework, 'positive' classifications reflect vocal patterns, including energetic delivery and forward prosodic momentum, that the model associates with non-distressed, assertive speech (Ma et al. 2024). When a sentence with negatively-valenced content (crisis diagnosis, problem identification) is delivered with these prosodic features, the text-audio divergence signals that the speaker is projecting confidence and resolve even while acknowledging difficulty. Prior research on political communication similarly finds that leaders strategically deploy assertive nonverbal delivery to project authority during adverse conditions (Grebelsky-Lichtman 2016). This pattern, in which negative content is delivered with assertive prosody, is consistent with a rhetorical structure in which presidents simultaneously signal awareness of problems and communicate their capacity to resolve them.

These mismatches provide valuable insight into the unique rhetorical strategies used in presidential speeches. Political addresses often present a complex structure that simultaneously introduces the problems and the signal to solve them. This leads to different interpretations by text- and audio-based models. This divergence highlights core features of political persuasion that are difficult to capture with a single analytical approach, suggesting that future research should focus more closely on these mismatches as a

central object of analysis. Furthermore, in this context, relying on a single analytical approach to classify individual utterances or entire speeches entails a serious risk of misinterpretation in political analysis.

Presidential addresses to the National Assembly are political in nature, rather than mere policy explanations or one-way communication. Therefore, their meaning can only be fully understood by comprehensively considering the signals conveyed through both the text and audio channels together. Most importantly, the findings of this study confirm that these concerns are observable realities, not just theoretical assumptions. Numerous instances of mismatches between text and audio were identified in actual cases, further reinforcing the need for a multimodal approach. Thus, this study not only points out the limitations of existing approaches but also justifies a new analytical framework for studying political speech.

Furthermore, this analytical approach is closely related to the broader direction of technological development. First, the construction of high-quality training datasets that reflect the specific characteristics of Korean political discourse is essential. Customized text and audio datasets capable of capturing the complex nature of political speech are required. Second, future research should use fusion-based approaches, such as late fusion or integration frameworks based on Large Language Models (LLMs), to more precisely combine and interpret the distinct text and audio outcomes. This involves advancing beyond simply improving the accuracy of emotion classification to a deeper level of analysis. This detailed analysis will explore the underlying reasons for mismatches and how they relate to political rhetorical strategies.

Together, these studies expand beyond the application of machine learning techniques to propose a new approach to analyzing presidential speeches as political strategies. By integrating text and audio data, it extends the analytical horizon for studying political discourse and persuasive strategies, offering a more comprehensive understanding of how political meaning is constructed and conveyed.

## V. Discussion

This study analyzes presidential addresses to the National Assembly from textual and audio perspectives. This approach sheds light on forms of political interaction and rhetorical strategies that single-mode approaches do not capture. Text-based analysis revealed changes in presidential policy agendas and rhetorical structures over time. In contrast, audio-based analysis visually demonstrated variations in speaking styles, including intonation and emotional tone. Furthermore, a multimodal analysis combining these two approaches reveals that discrepancies between text- and audio-based sentiment classifications are significant indicators of the persuasive strategies employed in presidential speeches. These findings, when considered together, suggest that presidential addresses to the National Assembly should be reinterpreted as political actions involving complex decision-making processes and strategic interaction, rather than simply as venues for policy proposals.

Future research should advance beyond basic forms of late fusion, such as averaging text- and audio-based prediction probabilities. Instead, it should adopt approaches that dynamically adjust the importance of text and audio data based on political context. Let $P(y_i)$ denote the probability that a political utterance is classified into category $y_i$. This relationship can be expressed as:

$$P(y_i) = \alpha_i \cdot P_{text(y_i)} + (1 - \alpha_i) \cdot P_{audio(y_i)}$$

Here, $\alpha_i \in [0,1]$ is not a fixed constant but a context-dependent weight that may vary according to the type of speech (e.g., policy address, press interview, or speech by a party leader), the broader political situation, and the immediate rhetorical context of the statement. For example, textual content may be more critical when record-keeping and policy presentation are priorities, such as when presenting long-term policy visions ($\alpha_i \uparrow$). In contrast, vocal cues may play a more substantial role in situations involving partisan conflict or where persuasive tactics are necessary, thereby increasing the importance of the audio signal ($\alpha_i \downarrow$). Therefore, identifying and validating context-dependent weighting schemes empirically is a core task for future research on political speech.

To illustrate how this weighting scheme might operate in practice, consider the mismatch patterns identified in the present analysis. During President Lee Myung-bak's October 2008 address (delivered in the context of the global financial crisis), 39 of 178 sentences (21.9%) were classified as negative by the text model but positive by the audio model. In this speech, where policy credibility and public reassurance were paramount, the audio signal may carry disproportionate interpretive weight ($\alpha_i \downarrow$). By contrast, in President Roh Moo-hyun's October 2003 address, where the political statement content of the speech was primary, the text channel's diagnostic value is likely higher ($\alpha_i \uparrow$). While these are illustrative inferences rather than empirically estimated values, they demonstrate that α is not an abstract parameter: it varies with the communicative purpose of each speech event. Systematically estimating α through external validation (such as correlating text-audio weights with audience reactions, approval rating shifts, or media framing) constitutes a concrete and tractable agenda for future research.

Furthermore, this dynamic weighting approach is not limited to plenary addresses delivered in the National Assembly. The relative weights assigned to text and audio when analyzing statements made in standing committee hearings may vary depending on the conversational context and surrounding political circumstances. More broadly, the weights may shift differently over time, particularly during specific political periods such as election campaigns. Substantively, this research contributes to the identification of the institutional and situational conditions under which political speech strategies are implemented. From a methodological aspect, it provides an important foundation for advancing multimodal approaches to the study of political speech.

To extend this line of research, securing adequate computational resources for data processing and model training is critical. Many researchers currently have limited access to high-performance computing resources, such as GPUs. This limits the adoption and systematic testing of computationally intensive approaches, including multimodal analysis. These constraints could be alleviated by expanding national-level research support or establishing shared research infrastructure, which would enable more extensive experimentation and validation of integrative methodologies. Such efforts would ultimately contribute to

methodological innovation in political science and accelerate the accumulation and transmission of scholarly findings.

Additionally, the infrastructure for systematically collecting and publicly releasing political speech data must be further developed. While textual records of presidential archives and legislative speeches are relatively well organized and accessible, audio and video data remain fragmented and difficult to obtain. A rigorous analysis of political speech requires the systematic, long-term construction of multimodal datasets, as well as researcher-friendly access to these materials (Barari and Simko 2025). Developing infrastructure in this way would improve research precision and significantly facilitate interdisciplinary collaboration.

Finally, even though contemporary research methods are evolving rapidly, it remains difficult for individual researchers to learn them and apply them to substantive and creative research questions in a short period of time. Thus, there is an increasing demand for structured research training programs and hands-on, practice-oriented workshops that facilitate the implementation of new approaches. Such initiatives would not only strengthen the methodological capacity of individual scholars but also contribute to the development of a more robust and sustainable research ecosystem within the academic community.

## VI. Conclusion

The study analyzed presidential speeches at two levels: *text-as-data* and *audio-as-data*. Focusing on twenty National Assembly addresses delivered by Presidents Roh Moo-hyun, Lee Myung-bak, Park Geun-hye, and Moon Jae-in, the analysis combined BERTopic-based topic modeling, audio-embedding-based clustering, and multimodal sentiment analysis. The findings demonstrate that presidential addresses to the National Assembly are not just opportunities to present policies but also highly strategic political actions that reveal the dynamics of executive-legislative relations in a multidimensional manner.

This study's contributions can be summarized in two main points. First, this research opens a new analytical avenue, *a politics of words and voices*, by presenting a systematic analysis that integrates textual and audio data in the study of Korean politics. Second, this study introduces a new dimension of evidence

into research on executive-legislative relations by demonstrating how presidential speaking styles and emotional signals operate as mechanisms shaping power relations and communication strategies. In doing so, it also illustrates how this perspective can serve as a foundation for a wide range of future studies.

This study is subject to several limitations. It focuses on a specific political context: presidential addresses to the National Assembly. Within its multimodal framework, it remains centered on relatively crude sentiment classification. At the same time, these constraints point directly to promising directions for future research. Developing models capable of systematically analyzing Korean political communication will require the construction of high-quality training datasets of Korean political text and audio, alongside sustained institutional and technical support. In addition, incorporating interactive elements such as legislative responses to presidential speech would allow for a more precise examination of the substantive dynamics of power relations between the executive and the legislature. Finally, future research may explore integrating visual modalities, such as gestures and facial expressions, to enrich the analysis of political processes.

In conclusion, presidential speech is the product of complex political calculation in which *what is said* and *how it is said* are inseparably intertwined. By uncovering these dual layers of communication through data-driven analysis, this study points toward a new direction for research on presidential rhetoric. By doing so, it helps open additional avenues for future research. More broadly, this study contributes to ongoing efforts to refine methodological approaches to political speech and to deepen empirical analysis of Korean political institutions.

## References


Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. "wav2vec 2.0: A Framework for Self-supervised Learning of Speech Representations." *Advances in Neural Information Processing Systems* 33: 12449–12460.

Barari, Soubhik, and Tibor Simko. 2025. "The Promise of Text, Audio, and Video Data for the Study of U.S. Local Politics and Federalism." *Publius: The Journal of Federalism* 55(2): 223–52.

Campbell, Karlyn Kohrs, and Kathleen Hall Jamieson. 1990. *Deeds Done in Words: Presidential Rhetoric and the Genres of Governance.* Chicago: University of Chicago Press.

Cho, Eunmi. 2024. "Determinants of Congressional Negative Discourse: Focusing on Legislators' Ideological Extremity, Legislative Experience, and Party Status." *Korean Party Studies Review* 23(4): 77–114. *(In Korean).*

Choi, Jihyun, Dong Uk Cho, Bum Joo Lee, Chanjung Kim, and Yeon-Man Jeong. 2016. "Identification of Advantages and Disadvantages Relative to Competitors of Politicians According to the Narrative Styles by Applying Voice Analysis." *The Journal of Korean Institute of Communications and Information Sciences* 41(5): 602–609. *(In Korean).*

Chung, Eunee, and In-Yong Nam. 2016. "A Study on the Vocal Images in Successive Presidential Speeches: From Former President Park Jeong-hee to Incumbent President Park Geun-hye." *Journal of Political Communication* 43: 323–349. *(In Korean).*

Chung, Eunee, and Sangho Lee. 2016. "A Comparative Study on the Public Speech Spectrum between ROK and USA Politicians." *Journal of Digital Contents Society* 17(3): 143–152. *(In Korean).*

Cohen, Jeffrey E. 1995. "Presidential Rhetoric and the Public Agenda." *American Journal of Political Science* 39(1): 87–107.

Dietrich, Bryce J., Matthew Hayes, and Diana Z. O'Brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review* 113(4): 941–62.

Grebelsky-Lichtman, Tsfira. 2016. "The Role of Verbal and Nonverbal Behavior in Televised Political Debates." *Journal of Political Marketing* 15(4): 362–87.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis for Political Texts." *Political Analysis* 21(3): 267–97.

Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." *arXiv preprint arXiv:2203.05794.*

Ham, Jihoon, Young Jin Choe, Janghoon Park, Inseop Choi, and Hyunsoo Soh. 2020. "KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding." *arXiv preprint*

*arXiv:2004.03289.*

Hashem, Ahlam, Muhammad Arif, and Manal Alghamdi. 2023. "Speech Emotion Recognition Approaches: A Systematic Review." *Speech Communication* 154: 102974.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 3451–3460.

Hur, Jaeyoung, and Kihong Eom. 2012. "The Perception of President Roh Moo-Hyun on Self-Reliant National Defense: Content Analysis of Presidential Speeches." *East and West Studies* 24(4): 37–56. *(In Korean).*

Jin, Youngjae, and Seongyi Yun. 2024. *Korean Politics*. 3rd ed. Seoul: Bubmoonsa. *(In Korean).*

Kang, Taewan, and Eunjeong Kim. 2004. "Rhetorical Characteristics and Role Framing in Presidential Speeches across Administrations." *Journal of Social Science* 30(2): 53–89. *(In Korean).*

Kim, Gi-duk, Mi-sook Kim, and Hack-man Lee. 2021. "Speech Emotion Recognition through Time Series Classification." *Proceedings of the Korean Society of Computer Information Conference, Jeju (In Korean).*

Kim, Hyok. 2014. "A Study on Presidential Leadership and Policy Agenda Setting Pattern: A Content Analysis on Korean Presidential Addresses." *Journal of Korean Politics* 23(2): 77–102. *(In Korean).*

Kim, Jounghee, and Pilsung Kang. 2021. "Korean Speech Recognition Using Wav2vec2.0." *Proceedings of the Spring Joint Conference of the Korean Institute of Industrial Engineers*, *Jeju*. *(In Korean).*

Kim, Ju-Hee, and SeokPil Lee. 2021. "Multi-modal Emotion Recognition using Speech Features and Text Embedding." *The Transactions of the Korean Institute of Electrical Engineers* 70(1): 108–113. *(In Korean).*

Kim, Su-Yun, and Sumi Han. 2024. "Analysis of Political Discourse in Presidential Speeches Using Text-Mining." *Journal of Digital Contents Society* 25(12): 3871–3884. *(In Korean).*

Kim, Sung-Sik, Jin-Hwan Yang, Hyuk-Soon Choi, Jun-Heok Go, and Nammee Moon. 2024. "The Research on Emotion Recognition through Multimodal Feature Combination." *Annual Conference of KIPS*, May 23, 739–40.

Knox, Dean, and Christopher Lucas. 2021. "A Dynamic Model of Speech for the Social Sciences." *American Political Science Review* 115(2): 649–666.

Kwon, Hyang Won, and Do Lim Choi. 2013. "How the Presidents in South Korea Strategically Use Linguistic Symbols for Institutional Legitimacy during Incumbency in Response to Crisis and Opportunities?" *Journal of Governmental Studies* 19(3): 285–320. *(In Korean).*

Lee, Byong-Kwon. 2025. “Design and Implementation of an AI-Based Emotion Analysis System through Music and Vocal Separation.” *Journal of The Korea Society of Computer and Information* 30(7): 33–39. *(In Korean).*

Lim, Jaemin, Kiyeon Kim, Sunghyun Cho, and Suk-Bok Lee. 2024. “Wav2vec2.0 Hidden Representations for Voice Conversion.” *Journal of the Institute of Electronics and Information Engineers*, 141–149.

Linz, Juan J. 1990. “The Perils of Presidentialism.” *Journal of Democracy* 1(1): 51–69.

Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer, and Dustin Tingley. 2015. “Computer-Assisted Text Analysis for Comparative Politics.” *Political Analysis* 23(2): 254–77.

Ma, Ziyang, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. “emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation.” In *Findings of the Association for Computational Linguistics: ACL 2024*, 15747–15760.

McFee, Brian, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, et al. 2025. “Librosa/librosa: 0.11.0.” *Zenodo*.

Moon, Kwangmin, and Dong Wan Lim. 2020. “A Comparative Analysis of the State Administration and Financial Management Directions of Previous Governments in Budget Speeches.” *Journal of Association for Korean Public Administration History* 50: 107-138. *(In Korean).*

Park, Janghoon. 2020. “KoELECTRA: Pretrained ELECTRA Model for Korean.” GitHub repository. https://github.com/monologg/KoELECTRA

Park, Jong Hee, and Keyeun Lee. 2024. “Analysis of Offensive Language Use by South Korea’s Two Major Political Parties: Using Party Press Releases from 2007 to 2023.” *Korean Political Science Review* 58(2): 73–104. *(In Korean).*

Park, Jungwon, and Gwangmin Yoo. 2020. “An Exploratory Study on the Policy Trends of Regulatory Reform based on Presidential Speeches: Utilizing Text Mining and Topic Modeling.” *Korean Policy Studies Review* 29(4): 87–118. *(In Korean).*

Park, Sanghyun, Minkyoung Jung, and Jiyoung Park. 2022. “Analysis of Topic Changes Appearing in Unification Speech of Presidents – Approach Using Structural Topic Model and Word2Vec.” *Korean Unification Studies* 26(2): 99–156. *(In Korean).*

Proksch, Sven-Oliver, and Jonathan B. Slapin. 2012. “Institutional Foundations of Legislative Speech.” *American Journal of Political Science* 56(3): 520–537.

Proksch, Sven-Oliver, and Jonathan B. Slapin. 2015. *The Politics of Parliamentary Debate: Parties, Rebels and Representation.* Cambridge: Cambridge University Press.

Rabiner, Lawrence R., and Ronald W. Schafer. 2011. *Theory and Applications of Digital Speech Processing.* Pearson.

Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.” *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992.

Reimers, Nils, and Iryna Gurevych. 2020. “Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation.” In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing,* 4512-4525.

Rhyu, Sang Young, and Minjung Kim. 2021. “Two Paths of Nationalism in South Korea: A Text Mining Analysis of Official Speeches of Park Chung-hee and Kim Dae-jung.” *Journal of Contemporary Politics* 14(1): 87–130. *(In Korean).*

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58(4): 1064–82.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. “Exploring the Space of Topic Coherence Measures.” *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399-408.

Samuels, David, and Matthew Soberg Shugart. 2010. *Presidents, Parties, and Prime Ministers: How the Separation of Powers Affects Party Organization and Behavior.* Cambridge: Cambridge University Press.

Shin, Hyun-Sam, and Jun-Ki Hong. 2023. “Deep Learning-Based Speech Emotion Recognition Technology Using Voice Feature Filters.” *The Korea Journal of BigData* 8(2): 223–231. *(In Korean).*

Shugart, Matthew Soberg, and John M. Carey. 1992. *Presidents and Assemblies.* Cambridge: Cambridge University Press.

Ververidis, Dimitrios, and Constantine Kotropoulos. 2006. “Emotional Speech Recognition: Resources, Features, and Methods.” *Speech Communication* 48(9): 1162–1181.

# Appendices

## Appendix A: Model Fine-Tuning and Performance

This appendix provides full technical specifications for the two emotion classification models used in the multimodal analysis, including training procedures, evaluation metrics, and a discussion of the cross-modal performance asymmetry.

### A.1 KoELECTRA (Text-Based Emotion Classification)

The text-based emotion classifier was built on the *monologg/koelectra-base-v3-discriminator* pretrained model, which was fine-tuned end-to-end for 7-class emotion classification. The fine-tuning corpus was the NIA AI Hub "Conversational Speech Dataset for Emotion Classification" (N = 43,975 utterances), using the transcription column as input and the situation-based emotion label as the target.

Training configuration:

- Epochs: 5 (with early stopping, patience = 2, monitoring validation macro-F1)
- Learning rate: $3 \times 10^{-5}$
- Batch size: 16
- Loss function: weighted cross-entropy with balanced class weights
- Data split: 80/10/10 (train 35,192 / validation 4,399 / test 4,400), stratified by class, seed = 42
- Model selection criterion: validation macro-F1

Performance on the held-out test set (N = 4,400):

| **Metric** | **Score** |
|---|---|
| 7-class macro-F1 | 0.922 |
| Accuracy | 0.932 |
| Macro precision | 0.925 |
| Macro recall | 0.920 |

### A.2 Emotion2Vec (Audio-Based Emotion Classification)

The audio-based classifier used a two-stage pipeline: (1) extraction of frozen self-supervised embeddings from the emotion2vec_base pretrained model (Ma et al. 2024), and (2) classification via a logistic regression head with balanced class weights (saga solver). This architecture was chosen because the emotion2vec_base model produces high-quality speech representations without requiring GPU-intensive end-to-end fine-tuning.

Training configuration:

- Embedding extraction: emotion2vec_base (frozen, no gradient updates)
- Classifier: logistic regression, balanced class weights, saga solver
- Data: NIA AI Hub (N = 43,975 utterances), 80/20 stratified split, seed = 42
- Test set: N = 8,795

Performance on the held-out test set:

| Level | Metric | Score |
|---|---|---|
| 7-class | Macro-F1 | 0.584 |
| 7-class | Weighted-F1 | 0.611 |
| 7-class | Accuracy | 0.605 |
| 3-class | Macro-F1 | 0.692 |
| 3-class | Negative-F1 | 0.895 |
| 3-class | Positive-F1 | 0.634 |
| 3-class | Neutral-F1 | 0.548 |

### A.3 Training Data Composition

Both models were fine-tuned on the same underlying corpus: the NIA AI Hub "Conversational Speech Dataset for Emotion Classification," which contains recordings and transcriptions of everyday Korean conversational speech annotated with seven emotion categories. KoELECTRA used the text transcriptions (N = 43,991); Emotion2Vec used the corresponding audio files (N = 43,975). The 16-utterance difference reflects audio files excluded during preprocessing due to format or loading errors.

The distribution of the training data is as follows:

| Emotion (7-class) | N | 3-class mapping | Proportion |
|---|---|---|---|
| Sadness | 13,986 | Negative | 31.8% |
| Anger | 11,633 | Negative | 26.5% |
| Disgust | 4,660 | Negative | 10.6% |
| Fear | 4,131 | Negative | 9.3% |
| Happiness | 4,548 | Positive | 10.3% |
| Surprise | 1,755 | Positive | 4.0% |
| Neutral | 3,262 | Neutral | 7.4% |
| Total | 43,975 | | 100% |

Under the 3-class mapping used in this study, the training data is 78.2% negative, 14.3% positive, and 7.4% neutral. This imbalance reflects the emotional profile of everyday Korean conversational speech as captured in the NIA corpus, rather than any property of the political speech being analyzed. As discussed in the main text, this training-data composition directly contributes to both models' tendency to classify presidential speech as negative.

**A.4 Cross-Modal Performance Asymmetry**

The performance gap between KoELECTRA (7-class macro-F1 = 0.922) and Emotion2Vec (7-class macro-F1 = 0.58) reflects a fundamental difference in task difficulty rather than a deficiency in the audio model. Text-based emotion classification benefits from explicit lexical signals: words such as "crisis," "congratulations," or "regret" carry strong emotional connotations that a fine-tuned language model can readily exploit. Audio-based emotion recognition, by contrast, must infer emotional states from indirect acoustic cues, including pitch contour, speaking rate, and vocal tension, without access to lexical content. This asymmetry in task difficulty is well-documented in the speech emotion recognition literature (Hashem et al. 2023).

Critically, the 7-class macro-F1 understates Emotion2Vec's reliability for the analyses conducted in this study. All cross-modal comparisons use a 3-class mapping (negative, positive, neutral), under which Emotion2Vec achieves an F1 score of 0.90 for the negative class, the dominant valence category in the corpus. The model's weakness lies in distinguishing among specific negative emotions (e.g., sadness vs. anger vs. fear), a distinction that the present analysis does not require.

The performance asymmetry between the two models is, in fact, analytically productive. Because KoELECTRA and Emotion2Vec are strong on different dimensions, namely lexical content and vocal delivery, respectively, the cases where they disagree reveal something that neither model captures alone: the gap between what a president says and how they say it. If both models performed identically, the cross-modal comparison would be uninformative; it is precisely their complementary strengths that make the mismatch analysis interpretively valuable.

## Appendix B: BERTopic Coherence Validation

This appendix reports topic coherence scores for the BERTopic analysis described in the main text and provides validation that the extracted topics capture meaningful lexical patterns above chance.

### B.1 Per-Topic Coherence Scores

Topic coherence was assessed using the $C_v$ measure (Röder et al. 2015), computed via the *gensim* library. $C_v$ measures how frequently a topic's top keywords co-occur within a sliding text window, using normalized pointwise mutual information (NPMI) and cosine similarity. Among existing coherence measures, $C_v$ achieves the highest correlation with human judgments of topic quality. Higher values indicate more semantically coherent topics.

| Topic | $C_v$ | N (sentences) | Top keywords |
|---|---|---|---|
| 1 | 0.734 | 354 | Formal address conventions |
| 4 | 0.600 | 154 | Welfare and support measures |
| 0 | 0.583 | 355 | North Korea / inter-Korean affairs |
| 17 | 0.538 | 31 | Expressions of gratitude |
| 5 | 0.534 | 144 | Global economic context |
| 21 | 0.473 | 28 | Public safety |
| 18 | 0.437 | 31 | Culture and industry |

| | | | |
|---|---|---|---|
| 25 | 0.436 | 15 | Rhetorical emphasis |
| 3 | 0.431 | 199 | Innovation and competitiveness |
| 14 | 0.431 | 32 | Hedging and caution |
| 10 | 0.412 | 81 | Regional issues |
| 11 | 0.400 | 62 | Informal/personal narrative |
| 6 | 0.382 | 134 | Fiscal policy and budget |
| 12 | 0.382 | 49 | Economic conditions |
| 16 | 0.381 | 32 | Low-carbon / environment |
| 7 | 0.366 | 133 | Education policy |
| 2 | 0.347 | 222 | Economic crisis / COVID |
| 24 | 0.342 | 17 | Negation / denial |
| 23 | 0.339 | 23 | Trade and exports |
| 15 | 0.333 | 32 | Real estate policy |
| 13 | 0.329 | 49 | Government responsibility |
| 9 | 0.324 | 93 | Collective resolve |
| 19 | 0.308 | 30 | Market fairness |
| 26 | 0.302 | 13 | Energy policy |
| 20 | 0.294 | 28 | Historical references |
| 22 | 0.288 | 25 | Prosecutorial reform |
| 8 | 0.281 | 108 | Regulatory / civil service reform |
| 28 | 0.232 | 11 | (Residual) |
| 27 | 0.121 | 12 | (Residual) |

Note: Per-topic sentence counts are from the reproducibility re-run; the outlier rate reported in the main text (31.6%) reflects the original analysis.

Mean $C_v$ = 0.39 (median = 0.38, range: 0.12-0.73). The five highest-coherence topics are substantively interpretable and consistent with the thematic labels in Table 3. The two lowest-coherence topics (Topics 27 and 28) contain 12 and 11 sentences, respectively, representing residual heterogeneous content not captured by the main thematic structure. Their small size limits their influence on the cross-presidential comparisons.

**B.2 Random Baseline Comparison**

To assess whether the observed coherence reflects meaningful topic structure rather than an artifact of the coherence metric, I compared the observed $C_v$ scores against a random baseline. The procedure was as follows:

1. The topic assignments across all 3,900 sentences were randomly shuffled, breaking the association between sentences and topics while preserving the marginal distribution of topic sizes.
2. $C_v$ coherence was recomputed for each topic under the shuffled assignment.
3. Steps 1-2 were repeated 100 times to generate a null distribution of mean $C_v$ scores.

Results:

- Random baseline mean $C_v$: 0.32 (SD = 0.009)
- Observed mean $C_v$: 0.39
- Z-score: (0.39 - 0.32) / 0.009 = 7.64
- $p < 0.001$

The observed coherence significantly exceeds the random baseline, confirming that BERTopic's topic assignments capture meaningful lexical co-occurrence patterns rather than spurious groupings. The difference of 0.07 in mean $C_v$, while moderate in absolute terms, represents a 22% improvement over chance and is consistent with BERTopic's expected performance on short-text corpora where individual sentences provide limited lexical context for topic identification.

# Thesis scheduled confirmation document

| | | | |
|---|---|---|---|
| Document No. | THESIS-26-D000 | Manuscript ID | THESIS-26-004 |
| Title | Speech and Voice in Politics: Multimodal Analysis of Presidential Addresses to the Korean National Assembly | | |
| Author(s) | Kyusik Yang(1)<br>1.New York University | | |
| Corresponding Author | Kyusik Yang<br>Addr.<br>Tel.<br>E-mail. kyusik.yang@nyu.edu | | |
| Publish No. | Undefined | | |
| Publication | Undefined | Admission | 04, 2026 |

This manuscript has been submitted to this Journal and has been decided for publication, proving that it is currently under final editing and will be published.

June 6,2026

(IKS)