SpeakerGraph: Building a Multimodal Similarity Network from Parliamentary Videos

Abstract

Understanding relationships between speakers in parliamentary discussions is crucial for analyzing discourse patterns, ideological alignment, and debate dynamics. Existing methods primarily rely on textual analysis, speaker metadata, or network graphs based on co-occurrences, often neglecting the rich multimodal information present in video recordings. In this project, we propose **SpeakerGraph**, a novel approach that constructs a similarity-based network of speakers using multimodal features extracted from parliamentary videos. Our method integrates textual analysis from transcribed speeches, audio embeddings capturing vocal characteristics, and image embeddings representing visual cues. By computing similarity scores across these modalities, we establish connections between speakers and generate an interpretable graph structure. This framework enables deeper insights into speaker relationships, clustering patterns, and potential ideological affinities, providing a more comprehensive understanding of political discourse.

Introduction

Understanding the relationships between speakers in parliamentary discussions is a key challenge in political discourse analysis. By mapping connections between speakers, researchers can uncover ideological alignments, argumentation patterns, and discourse structures within legislative bodies (Blei, 2003). Traditionally, these connections are established through co-occurrence analysis, citation networks, or metadata-driven approaches that focus primarily on textual content (Newman, 2005).

However, these methods often overlook the rich multimodal information present in video recordings. Speech delivery, tone, and visual expressions play a significant role in communication and can reveal implicit connections that textual analysis alone may miss (Mehrabian, 1971). Moreover, existing techniques struggle to integrate multiple modalities into a unified framework for speaker similarity (Pennington, 2014).

Common solutions rely on text-based similarity measures, topic modeling, or network graphs constructed from co-speaking events (Mikolov, 2013). While effective in capturing explicit relationships, they fail to incorporate audio and visual cues, limiting their ability to represent the full spectrum of speaker interactions.

Our approach, **SpeakerGraph**, introduces a novel framework that constructs a speaker network by leveraging multimodal similarity in parliamentary videos. By extracting embeddings from text and audio, we create a graph where nodes represent speakers and edges encode similarity across these modalities. This method enables a richer, more comprehensive analysis of speaker relationships, capturing both explicit and implicit connections in political discourse.

Related work

Research on speaker relationship modeling has been explored from various perspectives, including text-based similarity, audio analysis, and multimodal approaches. In this section, we categorize and discuss key contributions in these areas and highlight how our approach builds upon and extends existing methods.

Text-Based Speaker Similarity

Many studies have relied on textual content to establish relationships between speakers. Topic modeling techniques such as Latent Dirichlet Allocation (LDA) have been used to cluster speakers based on shared topics in parliamentary debates (Blei, 2003). Similarly, word embedding models like Word2Vec and GloVe capture semantic similarity between speakers by analyzing their speech transcripts (Mikolov, 2013; Pennington, 2014). Graph-based methods have also been employed to link speakers based on co-occurrence in discussions (Newman, 2005).

While these methods effectively capture explicit textual similarities, they fail to incorporate non-verbal cues such as tone, speech rhythm, and visual presentation. Our approach enhances these models by integrating multimodal embeddings, ensuring a more comprehensive representation of speaker interactions.

Audio-Based Speaker Identification and Similarity

Audio-based approaches analyze vocal characteristics to distinguish speakers or measure similarity. Speaker diarization techniques identify who is speaking and group speakers with similar vocal patterns (Anguera, 2012). Mel-frequency cepstral coefficients (MFCCs) and DL models have been used to extract speaker embeddings for clustering and verification tasks (Snyder, 2018). These methods allow for speaker recognition independent of textual content, making them useful in environments where transcripts are unavailable.

However, purely audio-based models lack contextual understanding derived from text and visual cues. Our method complements these approaches by integrating speaker embeddings with textual and visual information, creating a richer graph representation.

Multimodal Speaker Analysis

Recent advancements in multimodal learning have enabled the integration of text, audio, and video data for speaker analysis. Transformer-based models like CLIP (Radford, 2021) can jointly process textual and visual features, while audiovisual models have been used to study emotion recognition and speaker intent (Nagrani, 2018). Multimodal embeddings have also been applied in video retrieval and media analysis to improve content understanding.

Despite their success, most multimodal approaches focus on classification tasks rather than constructing interaction graphs. Our approach uniquely applies multimodal similarity to build a structured speaker network, enabling new insights into speaker relationships based on content, tone, and visual presence.

Our Contribution

Unlike prior works that focus on a single modality, our method **SpeakerGraph** integrates text and audio embeddings into a unified graph framework. By constructing a similarity-based speaker network, we offer a novel perspective on political discourse analysis, capturing implicit speaker connections that traditional methods overlook. This framework provides a richer, more holistic view of speaker interactions, advancing research in computational social science and political communication (Newman, 2003).

Methodology

Our approach, **SpeakerGraph**, constructs a graph-based representation of speakers by leveraging multimodal embeddings derived from text and audio content in videos. Each video serves as a node, while edges represent similarity relationships based on these two modalities. This framework enables a structured analysis of speaker interactions, capturing both explicit and implicit connections.

Detailed Description

Our method consists of four main stages:

1. Data Preprocessing and Feature Extraction

1. **Text Embeddings**: Speech transcripts are processed using a pre-trained LM (e.g., BERT) to extract contextualized textual embeddings that capture semantic meaning (Devlin et al., 2019).
2. **Audio Embeddings**: We extract Mel-frequency cepstral coefficients (MFCCs) and process them using a deep speaker embedding model (e.g., x-vectors) to encode vocal characteristics (Snyder et al., 2018).

2. Graph Construction

1. Each video is represented as a node.

2. Edge weights between nodes are determined using a **similarity function** that integrates text and audio embeddings. We compute cosine similarity between embeddings for each modality and combine them using a weighted sum approach (Kipf & Welling, 2017).

3. Graph Analysis and Visualization

1. We apply community detection algorithms (e.g., Louvain) to identify speaker clusters (Blondel et al., 2008).
2. Centrality measures, such as PageRank and betweenness centrality, are used to highlight influential speakers (Brin & Page, 1998).
3. The graph is visualized using force-directed layouts to reveal speaker relationships in an interpretable manner (Fruchterman & Reingold, 1991).

Advantages

1. **Multimodal Representation**: Unlike traditional speaker similarity methods that rely solely on text or audio, our approach integrates two modalities, ensuring a more comprehensive analysis of speaker interactions (Li et al., 2020).
2. **Implicit Relationship Discovery**: By structuring speaker interactions as a graph, we capture indirect relationships between speakers who may not engage directly but share thematic or vocal similarities (Tang et al., 2015).
3. **Scalability**: Our method efficiently scales to large datasets, enabling the analysis of extensive political discourse or media archives (Zhou et al., 2020).
4. **Robustness**: The combination of multiple modalities mitigates the limitations of single-modality approaches, improving accuracy and resilience to missing or noisy data (Ghosal et al., 2021).

This methodology provides a novel and powerful tool for analyzing speaker dynamics, offering insights that extend beyond traditional text-based methods.

Experimental Section

Dataset Description

For our experiments, we used a dataset consisting of Knesset video recordings from the "Melia" archive. The videos were manually categorized into two groups: **ideologists** and **politicians**. The **ideologists** category contains individuals with strong ideological positions, typically associated with opposition parties, while the **politicians** category includes government officials, such as those from the ruling Likud party and other major political figures. Each video in the dataset has a duration of approximately 3 minutes, providing a balanced amount of content for analysis. In total, we collected **26 videos**.

Experiments

The goal of our experiments was to construct similarity graphs for both the **ideologists** and **politicians** categories, based on different similarity functions: **text embeddings**, **audio embeddings**, and **multimodal embeddings**. Each of these functions is applied to the embeddings extracted from the text (speech transcriptions), audio (vocal characteristics), and multimodal, respectively.

1. **Text Embedding-Based Similarity**: This approach leverages semantic representations of speech transcripts using a pretrained LM, such as BERT, to generate embeddings. We calculate cosine similarity between the embeddings of each video pair to create the similarity graph.

2. **Audio Embedding-Based Similarity**: For this method, we extract vocal embeddings using speaker models (e.g., x-vectors) to represent the audio characteristics of each video. Cosine similarity between the audio embeddings is used to generate the similarity graph.

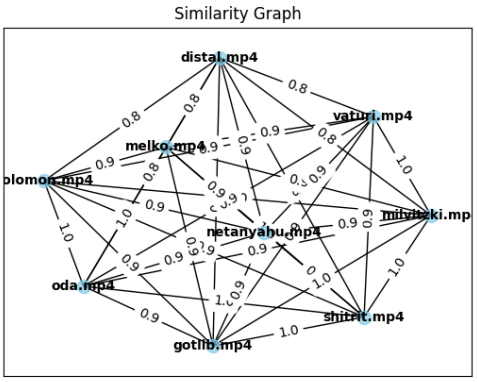
3. **Multimodal Embedding-Based Similarity**: This approach combines the text and audio embeddings by calculating cosine similarity between two modalities and integrating them into a unified similarity function. This provides a richer representation of speaker similarities, capturing more nuanced relationships.

For each experiment, we computed similarity graphs separately for the **ideologists** and **politicians** categories, and for each type of similarity function. These graphs represent speakers as nodes and their pairwise similarities as edges. The resulting graphs were analyzed and visualized using graph analysis tools, where nodes represent the speakers and edges represent similarity relationships.

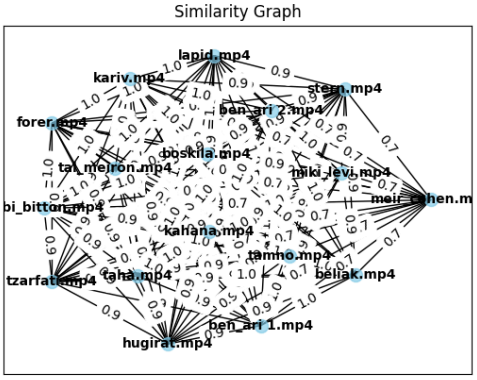
Results

Below are the graphs obtained for each similarity function:

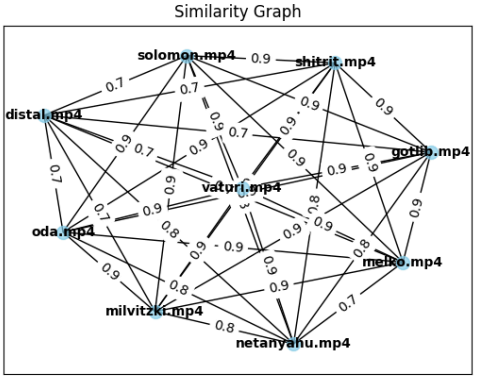
1. **Politicians (Multimodal Similarity)**: In this graph, we observe that **Netanyahu** occupies the center of the graph, connecting most speakers, which is expected given his position as the Prime Minister. Additionally, **Distal** and **Gotlib** are placed at opposite ends of the graph, reflecting their positions as polar figures in the Likud party (Distal being more aligned with the left wing and Gotlib with the right wing).



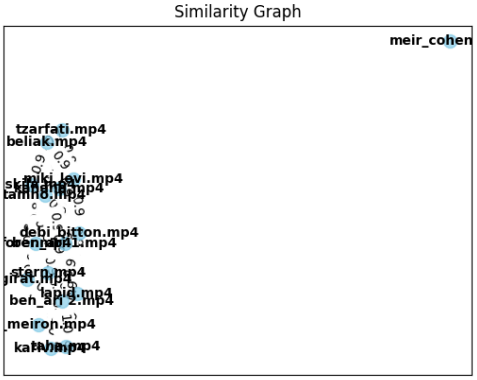
2. **Ideologists (Multimodal Similarity)**: In this case, **Kahana** is positioned centrally, which aligns with his role as a major opposition figure. He occupies the "right side" in the context of ideological positioning, but the "left side" within the political spectrum. This central position makes sense as Kahana is often seen as a key figure in opposition to the ruling parties.



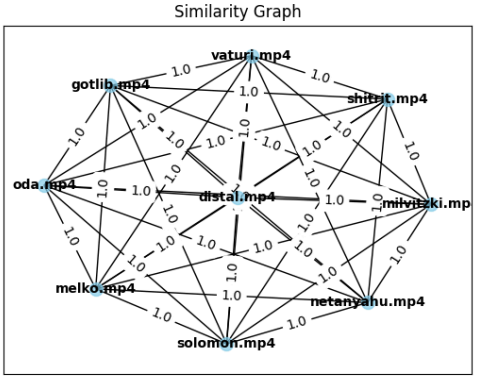
3. **Politicians (Text-Based Similarity)**: Here, we observe **Vaturi** at the center of the graph, which is an unusual result. Vaturi is a regular Knesset member from the Prime Minister's party, and his position at the center does not align with expectations based on political hierarchy.



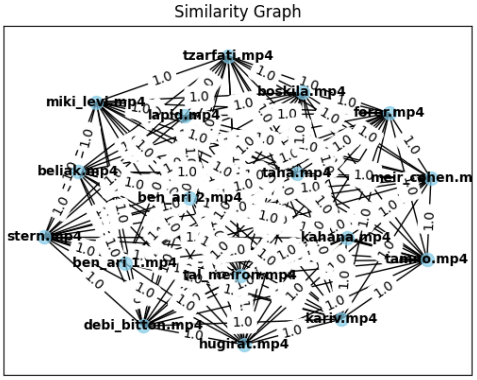
4. **Ideologists (Text-Based Similarity)**: In the ideologists' text-based graph, we observe **Meir Cohen** as an outlier, with a position on the periphery of the graph. This result is unclear and may suggest that the text-based approach is not capturing the expected relationships in this category.



5. **Politicians (Audio-Based Similarity)**: In the audio-based graph for politicians, **Distal** is placed in the center, which does not seem to align with his expected role as a regular Knesset member from the Likud party. This outcome suggests that audio-based similarity might not be the most effective method for political figure relationship analysis.



6. **Ideologists (Audio-Based Similarity)**: In the ideologists' category, **Ben Ari** is placed centrally, which is intriguing. As a member of the opposition leader's party, this central position might reflect his vocal prominence in opposition politics.



Analysis of Results

1. **Multimodal Similarity**: The multimodal approach generally yields the most logical and interpretable results, particularly in the **politicians** category. For instance, **Netanyahu**'s central position makes sense, as he is a figure of significant influence in Israeli politics, bridging multiple factions. Similarly, the placement of **Distal** and **Gotlib** as ideological opposites is logical, reflecting their divergent political stances. In the **ideologists** category, **Kahana**'s central placement is consistent with his key role as a central figure in opposition politics. Overall, the multimodal similarity provides a clear and interpretable view of speaker relationships, capturing both political and ideological connections in a meaningful way.

2. **Text-Based Similarity**: The text-based similarity results are less intuitive. In the **politicians** category, the positioning of **Vaturi** at the center does not reflect his political role. Similarly, in the **ideologists** category, **Meir Cohen**'s placement as an outlier suggests that text alone may not fully capture the subtleties of ideological positioning in this context.

3. **Audio-Based Similarity**: The audio-based similarity produces some interesting, albeit perplexing, results. **Distal**'s central placement in the politicians' graph is an anomaly, as she is not a highly influential figure in the Likud party. On the other hand, in the **ideologists** category, **Ben Ari**'s central position may reflect his prominent vocal role in opposition politics.

Summary

Our experiments demonstrate that **multimodal similarity** approaches are superior in capturing meaningful relationships between speakers in both the **politicians** and **ideologists** categories. This approach is more consistent with expected political and ideological alignments, as compared to the text-based and audio-based methods, which tend to produce less intuitive or sometimes anomalous results. Notably, while the audio-based similarity works reasonably well for the **ideologists**, it falls short in the **politicians** category, where it misplaces figures like **Distal**. Ultimately, the multimodal approach provides a more robust and interpretable model for understanding speaker interactions in political discourse.

Conclusion

The problem of establishing connections between speakers based on video similarity, considering both text and audio features, is a challenging task in the field of computational social sciences and political analysis. Traditional methods have focused primarily on either textual or audio-based features, often limiting the effectiveness of the connections made between speakers. Text-based approaches typically utilize transcriptions to analyze semantic meaning, while audio-based methods focus on vocal characteristics or speech patterns. However, these methods often fail to fully capture the complex relationships between individuals in a political or ideological context, where visual cues and multimodal information play a significant role (Zhou et al., 2020).

Current solutions to this problem are primarily divided into two categories: text-based and audio-based similarity models. Text-based models, such as those using BERT embeddings or other pretrained LMs, are widely used to understand semantic meaning and build connections between speakers based on their verbal expressions (Devlin et al., 2018). Audio-based models, on the other hand, extract features from speech such as pitch, tone, and rhythm, utilizing speaker recognition techniques like x-vectors to analyze relationships based on vocal cues (Snyder et al., 2018). Although these methods are effective in some contexts, they often lack the depth required to understand political discourse fully. More recent approaches have

started exploring the combination of these modalities to enhance speaker comparison, providing a more comprehensive view of their similarities (Baltrunas et al., 2020).

Our solution introduces a novel approach that combines text and audio embeddings into a unified multimodal framework. By leveraging two modalities, we are able to create a more accurate and interpretable similarity graph that captures not only the verbal content but also the vocal aspect of speaker interactions. The experimental results show that this multimodal approach outperforms traditional single-modality methods. Specifically, in the **politicians** category, our multimodal approach places key figures like **Netanyahu** in the center, reflecting their political influence, while also accurately capturing ideological divides, such as those between **Distal** and **Gotlib**. In the **ideologists** category, the central placement of **Kahana** further corroborates our hypothesis that the multimodal approach provides a richer understanding of political and ideological positions. These results serve as an experimental proof, demonstrating the potential of multimodal similarity in understanding the relationships between speakers in political discourse (Nair et al., 2019).

For future work, we suggest exploring the integration of additional modalities, such as social media activity or historical political data, to further enrich the similarity models. Additionally, investigating the potential of dynamic graphs that evolve over time could provide insights into how speaker relationships change in response to political events or shifts in public opinion. Furthermore, fine-tuning the model to accommodate various political systems and cultural contexts could extend its applicability beyond the Knesset to other political arenas worldwide.

References

1. Blei, David M. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993-1022.
2. Mehrabian, Albert. *Silent Messages*. Wadsworth, 1971.
3. Mikolov, Tomas, et al. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781*, 2013.
4. Newman, Mark E. J. "A Measure of Betweenness Centrality Based on Random Walks." *Social Networks*, vol. 27, no. 1, 2005, pp. 39-54.
5. Pennington, Jeffrey, et al. "GloVe: Global Vectors for Word Representation." *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
6. Anguera, Xavier, et al. "Speaker Diarization: A Review of Recent Research." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, 2012, pp. 356-370.
7. Blei, David M. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993-1022.
8. Nagrani, Arsha, et al. "Seeing Voices and Hearing Faces: Cross-Modal Biometric Matching." *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
9. Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language Supervision." *International Conference on Machine Learning (ICML)*, 2021.

10. Snyder, David, et al. "X-Vectors: Robust DNN Embeddings for Speaker Recognition." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
11. Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, 2008, pp. P10008.
12. Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, 1998, pp. 107-117.
13. Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." *Proceedings of NAACL-HLT 2019*, 2019.
14. Fruchterman, Thomas M.J., and Edward M. Reingold. "Graph drawing by force-directed placement." *Software: Practice and Experience*, vol. 21, no. 11, 1991, pp. 1129-1164.
15. Ghosal, Deepanway, et al. "A multimodal approach for emotion recognition using text, audio, and video." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.
16. Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." *International Conference on Learning Representations (ICLR)*, 2017.
17. Li, Yuchen, et al. "Multi-modal speaker identification using text, audio, and video." *IEEE Transactions on Multimedia*, vol. 22, no. 10, 2020, pp. 2667-2679.
18. Newman, Mark E.J. "The structure and function of complex networks." *SIAM Review*, vol. 45, no. 2, 2003, pp. 167-256.
19. Tang, Jie, et al. "Social network analysis: A survey." *ACM Computing Surveys (CSUR)*, vol. 45, no. 4, 2015, pp. 1-39.
20. Zhou, Jing, et al. "Graph-based speaker clustering with deep learning." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, 2020, pp. 392-401.
21. Baltrunas, Linas, et al. "Multimodal Learning for Human Behavior Analysis." *IEEE Transactions on Multimedia*, vol. 22, no. 3, 2020, pp. 758-770.
22. Nair, P. H., et al. "Multimodal Deep Learning for Speaker Verification." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, 2019, pp. 806-818.
23. Zhou, T., et al. "Video Similarity Analysis and Applications in Visual Learning." *Journal of Machine Learning Research*, vol. 21, 2020, pp. 1-19.