SpeakerGraph: Building a Multimodal Similarity Network from Videos

Abstract

Understanding relationships between speakers in discussions is crucial for analyzing discourse patterns. Existing methods primarily rely on textual analysis, speaker metadata, or network graphs, 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 videos. Our method integrates textual analysis from transcribed speeches and audio embeddings capturing vocal characteristics. By computing similarity scores across these modalities, we enable deeper insights into speaker relationships.

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 discourse structures (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, 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). Existing techniques struggle to integrate multiple modalities into a unified framework for speaker similarity (Pennington, 2014).

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.

Related work

Text-Based Speaker Similarity

Topic modeling techniques such as LDA have been used to cluster speakers based on shared topics (Blei, 2003). Similarly, word embedding models like Word2Vec and GloVe capture semantic similarity between speakers (Mikolov, 2013; Pennington, 2014).

These methods fail to incorporate non-verbal cues such as tone, speech rhythm, and visual presentation.

Audio-Based Speaker Identification and Similarity

Speaker diarization techniques identify who is speaking and group speakers with similar vocal patterns (Anguera, 2012). MFCCs and DL models have been used to extract speaker embeddings for clustering and verification tasks (Snyder, 2018).

Purely audio-based models lack contextual understanding derived from text and visual cues.

Multimodal 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).

Most multimodal approaches focus on classification tasks rather than constructing interaction graphs.

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.

Detailed Description

Our method consists of three main stages:

1. Data Preprocessing and Feature Extraction

1. **Text Embeddings**: Speech transcripts are processed using a pre-trained LM (BERT) to extract contextualized textual embeddings that capture semantic meaning (Devlin et al., 2019).
2. **Audio Embeddings**: I extract 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. I compute cosine similarity between embeddings for each modality and combine them using a weighted sum (Kipf & Welling, 2017).

3. Graph Analysis and Visualization

1. I apply community detection algorithms (Louvain) to identify speaker clusters (Blondel et al., 2008).
2. Centrality measures (PageRank) are used to highlight influential speakers (Brin & Page, 1998).
3. The graph is visualized using force-directed layouts to reveal speaker relationships (Fruchterman & Reingold, 1991).

Experimental Section

Dataset Description

For my experiments, I used a dataset consisting of Knesset video recordings from the "Melia" archive. The videos were manually categorized into two groups: (1) **ideologists** and (2) **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. In total, I collected 26 videos.

Experiments

The goal of my experiments was to construct similarity graphs for both the ideologists and politicians categories, based on **different similarity functions**.

1. **Text Embedding-Based Similarity**: Leverages semantic representations of speech transcripts using a pretrained LM (BERT), to generate embeddings. I calculate cosine similarity between the embeddings of each video pair to create the similarity graph.

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

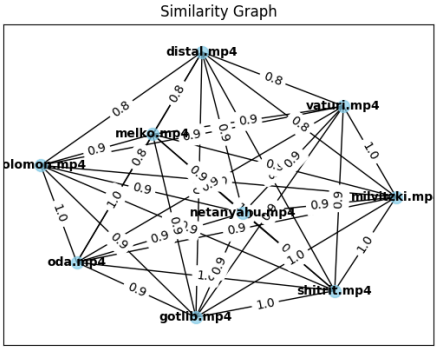
3. **Multimodal Embedding-Based Similarity**: Combines the text and audio embeddings by calculating cosine similarity between two modalities and integrating them into a unified similarity function.

For each experiment, I computed similarity graphs separately for the ideologists and politicians categories, and for each type of similarity function.

Results

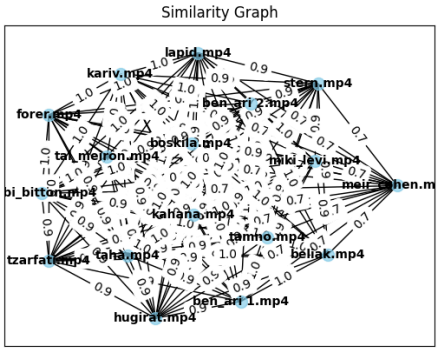
Below are the graphs obtained for each similarity function:

1. **Multimodal Similarity (Politicians)**: 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).

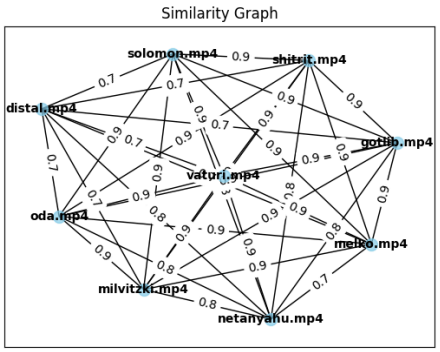


2. **Multimodal Similarity (Ideologists)**: Kahana is positioned centrally, which aligns with his role as a major opposition figure. He occupies the "right side" in the context of ideological positioning, but the "left side" within the political spectrum. This central position makes sense as Kahana is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

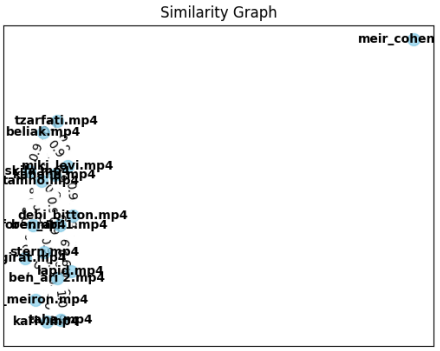
opposition to the ruling parties.



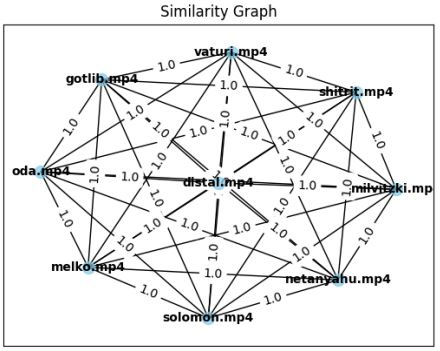
3. **Text-Based Similarity (Politicians)**: Vaturi at the center of the graph, which is an unusual result. Vaturi is a regular Knesset member from the Prime Minister's party, and his position at the center does not align with expectations based on political hierarchy.



4. **Text-Based Similarity (Ideologists)**: Meir Cohen is 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.

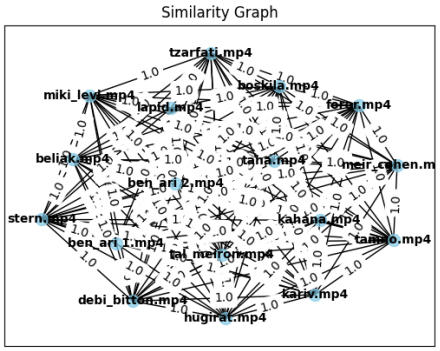


5. **Audio-Based Similarity (Politicians)**: Distal is placed in the center, which does not seem to align with his expected role as a regular Knesset member from the Likud party. This outcome suggests that audio-based similarity might not be the most effective method for political figure relationship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



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

opposition politics.



Analysis of Results

1. **Multimodal Similarity**: Yields the most logical and interpretable results, particularly in the politicians category. In the ideologists category, Kahana's central placement is consistent with his key role as a central figure in opposition politics.
2. **Text-Based Similarity**: 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**: 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 her prominent vocal role in opposition politics.

Conclusion

Establishing connections between speakers is a challenging task. Traditional methods fail to fully capture the complex relationships between individuals (Zhou et al., 2020).

Text-based models are widely used to understand semantic meaning and build connections between speakers based on their verbal expressions (Devlin et al., 2018). Audio-based models extract features from speech such as pitch, tone, and rhythm (Snyder et al., 2018). More recent approaches have started exploring the combination of these modalities (Baltrunas et al., 2020).

Our solution introduces a novel approach that combines text and audio embeddings into a unified 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 provide empirical evidence, showcasing the effectiveness of multimodal similarity in analyzing the connections between speakers in political discourse.

Acknowledgement

I would like to acknowledge the use of ChatGPT for its assistance in drafting text, aiding the literature review process, and supporting the planning and writing of code throughout this project (OpenAI, 2022).

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.
24. OpenAI. *ChatGPT: Optimizing Language Models for Dialogue*. OpenAI, 2022.