Meeting audio data summarization and visualization using ASR and NLP tools within the context of captured meeting data of the Shape Language

Ella Milinovic

Supervisor(s): Stephanie Tan, Edgar Salas Gironés

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Ella Milinovic Final project course: CSE3000 Research Project
Thesis committee: Stephanie Tan, Edgar Salas Gironés, Sole Pera

Abstract

Meetings are a vital part of discussions and negotiations. Unfortunately, individuals often leave with a vague understanding of the topics covered during the meeting and tend to forget even more of what transpired as time goes on. Driven by previous research that attempts to solve the issue by using architectural shapes as a way of removing ambiguity along with recent advancements in Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) this research attempts to improve user understanding of key topics discussed in meetings by combining ASR models with NLP tools to create a visual summary that would improve user understanding of key topics covered during meetings. To achieve this the research utilizes the speech-to-text transcription and speaker identification capabilities of the WhisperX model with noun phrase extraction features provided by Spacy and key topic recognition functionality of Microsoft's DeBERTa model. Finally, the data is presented as a node-based graph utilizing the D3.js library. The results show that the system is able to identify between 33% - 58% of meeting key topics. This shows the potential of combining ASR models with NLP tools for creating concise meeting summaries but also raises new questions such as why some topics were missed, how the system performance can be improved, and how to design an optimal user interface for such a task.

1 Introduction

Meetings are a key component in discussions, planning, and negotiations. In addition, research has shown that workers sometimes spend more than half of their working hours in meetings [1] [2]. However, individuals often leave these meetings with a different or incomplete understanding of the topic/s discussed. This can sometimes lead to frustrations or some individuals even claim that meetings are a barrier to productivity [1] instead of a helpful mutual discussion. Additionally the more abstract the topic of the meetings is the less understanding people have of it when they leave [3].

Previous research has addressed this issue by using architectural shapes in order to improve understanding of complex discussions of questions/topics that have no definitive solution [4], these questions are sometimes referred to as wicked problems [5]. This approach named the Shape language (VoormTaal in Dutch)[4] leverages a set of geometric shapes such as spheres, pyramids, etc. as a tool for removing some of the abstractions in order to facilitate clearer communication and comprehension of the topics. The way these meetings were set up is as follows: Participants were given a topic such as "How should communities and real estate companies perceive and adapt to hazard risks?". They then needed to discuss this topic using shapes from the "Shape language " as visual and tactile communication aids. These shapes represent key topics such as: pyramids represent hazards, spheres communities, cubes real estate companies, etc. As shown in Figure1.

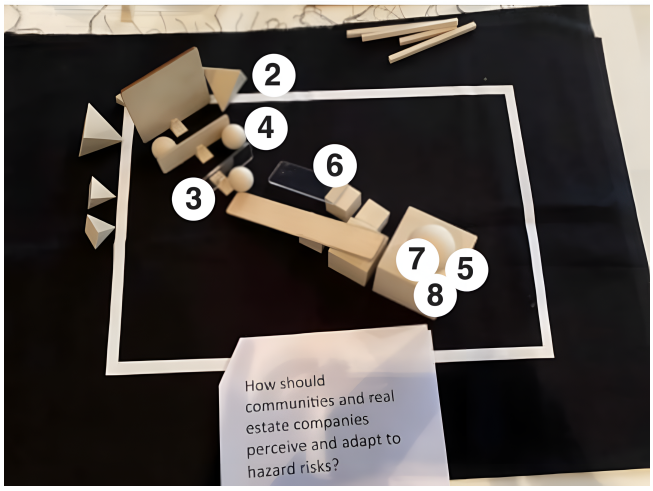

Figure 1: Meaning behind each shape used during the shape language meeting: 2 - pyramids represent climate hazards, 3 - plates represent barriers to what is known, 4 - spheres are different communities, 5 - cubes represent real estate companies, 6 - bridges represent pathways to reach adaptation, 7 - the large sphere represents the society, 8 - the large sphere on the cube represents the integration of the real estate companies and society

Furthermore, all the meetings conducted for the previous research explored questions and issues related to climate adaptation in the Netherlands. Although the abovementioned research provides a valuable visual aid and solution for a better understanding of complex discussions during a meeting it does not address the issue that on average individuals forget over 60% of newly given information after only 1 day and this gets worse as time progresses [6].

To address the issues this research will explore a more technical solution, namely combining automatic speech recognition (ASR) models with natural language processing (NLP) tools to create a node graph-based key topic summary of the gathered meeting audio data.

Existing research in the domain of speech-to-text conversion, such as OpenAi's Whisper model provides a model for accurately converting audio recordings into text [7]. Additionally, speaker identification tools, also known as speaker diarization tools such as Pyannote [8], can be combined with the Whisper model to facilitate accurate speech-to-text conversion with speaker identification (WhisperX model [9]). Moreover, a tool such as spaCy [10] can extract noun phrases. Finally, the noun phrases can be passed to the Microsoft DeBERTa model [11] along with pieces of the meeting text. The model can then determine which of the noun phrases are most likely the focus of the provided text also known as key phrases. While individually these tools are able to create key components that can contribute to meeting summaries combining them to create a visual key topic summary and how this would contribute to individual understanding of meetings remains unexplored.

The research aims to answer the following question:

How can automatic speech recognition and natural language processing tools extract key meeting topics to create a visual summary of meeting audio data?

The paper is structured as follows: In Section 2 an overview of related work is provided. Section 3 describes the dataset of climate adaptation meetings along with the models and tools used to address the research question. Section 4 presents the experimental system workflow and describes the experimental setup. In Section 5 the research results are presented. Section 6 analyzes the results along with their implications and relevance for the research goals. In section 7 the conclusions of the research and potential directions for future work are addressed. Finally, Section 8 describes responsible research practices used during the research.

2 Related Work

Similar research has been conducted in implementing a sliding window approach to create a summary of meeting minutes [12]. While this approach gives a promising way of automatically creating meeting minute summaries it is not able to extract key topics for a shorter more concise way of showing meeting data. Additionally, it is not able to provide data that can be presented in a visually appealing way in a graph-based user interface.

Furthermore, with recent advancements in artificial intelligence researchers have investigated whether conversational artificial intelligence models based on Generative Pre-trained Transformers (ChatGPT) can be used for dialogue summarization [13]. Although this research utilizes a powerful state-of-the-art model its findings show that the models tend to produce overly long summaries. Although this improves when the models are given more specific prompts, this research still isn't able to create a short key topic-focused summary or visualize it in an engaging way.

Additionally, using unsupervised approaches for automatic keyword extraction from meeting transcripts has been explored [14]. While this research presents a good foundation for extracting keywords it only relies on traditional unsupervised methods such as term frequency, inverse document frequency (TFIDF), and word clustering, rather than leveraging state-of-the-art pre-trained models. Additionally, the research paper itself shows that approximately 30% of the extracted keywords were rejected by human evaluators.

Furthermore, due to an existing need for visualizing meeting data in an interactive and user-friendly manner researchers have proposed web user interface designs that show meeting data as a graph [15]. While this research proposes a node-based graph representation of meetings with a time slider component the user interface design is quite outdated. In addition, the research explores a conceptual design idea rather than implementing a working system that is able to summarize meeting data from meeting recordings.

3 Methodology

This section describes the dataset, models, and tools used in the research. It begins with an overview of the dataset collected from meetings conducted using the Shape language. The subsequent subsections present the models and tools used in the study, along with the reasons they were chosen.

3.1 The dataset

The research will focus on audio data gathered from meetings of the Shape language. This data is from a preexisting dataset gathered as a part of a three-part master's thesis that combines communication and architecture. Specifically, the Shape language meetings were collected for "Part III - Speaking through form" [4]. That is, four meetings were held with four participants, each of whom was required to discuss wicked problems/problems with no definitive solution. The question each group was given is shown in Figure 2.

Groups:	Question:
Group 1	How do physical climate risks affect Dutch residential real-estate markets?
Group 2	How to study how residents and policy actors make sense of subsidence in Bloemhof in the particular institutional environment of Rotterdam?
Group 3	How should communities and real estate companies perceive and adapt to hazard risks?
Group 4	Flood risk adaptation, and climate adaptation in general, is a combination of measures on various scales. How can we design flood adaptation strategies so that measures at different scales complement each other and prevent maladaptation?

Figure 2: Discussion questions assigned to groups

Each of the four meetings lasted between twelve to fifteen minutes. Right after each meeting was concluded, all the participants of each group were asked to give one list of key topics and their associated shapes. The key topics from each of these lists are considered the most relevant key topics for the associated meeting and will be used to evaluate the performance of the proposed system.

3.2 Speech-to-text

For transcribing the audio data into text along with identifying which part was spoken by which participant this research utilizes the WhisperX model [9]. WhisperX is a model that combines OpenAi's Whisper model [7] and pyannote [8] to facilitate audio transcription with speaker identification, also known as speaker diarization. This model was chosen due to the Whisper model itself performing well across various datasets, as shown in Figure 3.

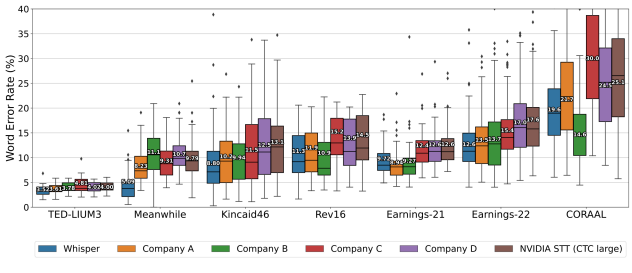

Figure 3: Whisper word error rate (WER) performance compared to other commercial ASR services [7]

Additionally, the model also performs similarly or even better than professional human transcribers [7]. Unfortunately while high performing the model itself does not support speaker diarization. On the other hand, Pyannote is an open-source toolkit designed for speaker diarization that does not support speech-to-text but performs well for speaker identification across various datasets [16] [8]. After attempting to combine Whisper and Pyannote, it was discovered that transcribed audio data timestamps have a tendency to not match and that additional preprocessing is required. Therefore the WhisperX model was chosen as a high-performing modified version of Whisper that supports speaker diarization. [9].

3.3 Parts of Speech Identification

Parts of speech identification is an area of Natural Language Processing (NLP) that deals with assigning grammatical categories such as nouns, verbs, etc. to words in a given text. Assigning these grammatical categories enables NLP tools/-models to identify the semantic meaning of a word in a sentence. This feature will be used to identify potential key topics in segments of text. For parts of speech identification, this research will use SpaCy.

SpaCy is an open-source Natural Language Processing (NLP) library [10]. This library was chosen for its pre-trained language models, more specifically for its English large model (en_core_web_lg) [17]. The specific model was chosen for its part of speech tagging (POS), static word vectors, and similarity comparison capabilities. Due to being trained on part of speech tagging (POS), the model is able to extract noun phrases/noun chunks accurately. A noun phrase is a group of words that consist of a noun and words that further describe/define the noun.

This research will utilize the extracted noun phrases to construct lists of potential key topics that can be incorporated into a meeting summary. Additionally, spaCy's similarity function which uses static (pre-trained) word vectors, will be used for comparing the extracted words with a list of key topics for each meeting in order to evaluate the performance of the system.

3.4 Zero-shot classification

Zero-shot classification in Natural Language Processing (NLP) is a task where a pre-trained model is expected to classify previous unseen data (data it has not been trained on) [18]. This approach enables accurate data classification on small datasets without the need for training a model on the dataset. A high-performing model of this type is the DeBERTa-v3-Large-Zero-Shot-v2.0 model [11]. The DeBERTa-v3-Large-Zero-Shot-v2.0 is a model from the Microsoft zero-shot-v2.0 series of models. The models from this series are zero-shot versions of the Microsoft DeBERTa (Decoding-enhanced BERT with disentangled attention) model. The DeBERTa model is an enhanced transformers-based model able to capture the meaning of words within the context of the text they are in. This enables the model to extract likely key topics from the provided text.

The research will utilize the features that the zero-shot DeBERTa model offers to extract the most likely key topics from transcribed meeting audio data without prior training.

3.5 Data visualization

As previously stated when data is complex or abstract individuals have a harder time understanding it [3]. This is why this research motivated by previous research in using the Shape language as a tool for better understanding [4] along with studies that show that visualizing data can improve understanding and decision making [19] aims to present the summarized meeting data in a visual and appealing way. To do so this research utilizes the D3.js library. D3.js (Data-Driven-Documents) is an open-source and versatile JavaScript library used for data visualization [20]. The library was chosen due to its versatility, dynamic data presentation capabilities, and large user community. The research will utilize this library to create a node graph with a time slider component representing summarized key topics of the meeting data.

4 Experimental Setup

This section describes the experimental setup, starting with the data processing pipeline, which includes audio transcription with speaker recognition and key topic extraction, followed by data visualization. An overview of the entire system architecture can be seen in Figure 4.

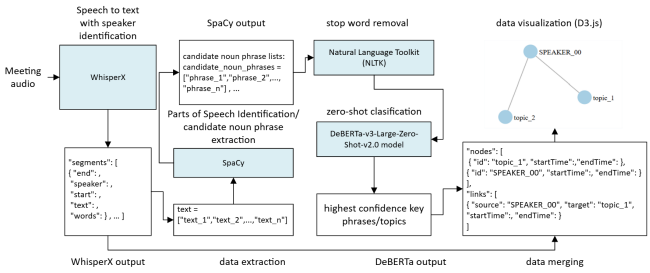

Figure 4: System architecture diagram

4.1 Data Processing

In order to visualize and summarize meeting audio data the data needs to be converted into a textual format and processed. This section describes how this is achieved:

Audio transcription with speaker diarization

The required first step is processing the audio data recorded during the Shape language-based meetings into a textual format. This is done by passing the recordings to the WhisperX model [9] which transcribes the meeting audio and assigns each generated text segment its respective speaker, start time, and end time. The speakers are not known due to WhisperX's internal logic and are labeled as SPEAKAER_00, SPEAKER_01, etc. The start time and end time are shown in seconds.

Key topic extraction

Once the audio data has been transcribed and speakers identified it needs to be further processed to acquire a dataset of key focus points/key topics of each speaker during the meetings.

To obtain a list of candidate words for key topics the transcribed data is passed through spaCy, which then facilitates noun phrase extraction [10]. To further improve the generated list of candidate noun phrases a stop word list from the Natural Language ToolKit (NLTK) [21] is applied to the dataset to remove stop words that add no value to the context. This is done to reduce the dataset and improve the performance of the system [22]. After the list of candidate noun phrases is obtained the list along with the respective data segment is passed to Microsoft's DeBERTa-v3-Large-Zero-Shot-v2.0 NLP model [11] which ranks how likely it is that each noun phrase is the key topic of that data segment and returns the highest ranking one.

4.2 Data visualization

The d3.js library is used to create a web-based node graph that shows key topics with their respective speakers across time frames. Once the audio data is transcribed into text and key topics and their speakers are identified within the set time frames the meeting data is transformed into a format that is accepted by the d3.js library.

```
"nodes": [
  { "id": "topic", "startTime":,"endTime": },
  { "id": "SPEAKER_00", "startTime":, "endTime": }
],
"links": [
  { "source": "SPEAKER_00", "target": "topic", "startTime":, "endTime": }
]
```

As shown above, the data is represented as follows: Both the speakers and key topics are represented as nodes with their respective speaker or topic, start time, and end time of the time frame in which they are mentioned. They are connected by edges that each have their source, destination, start and end time of their respective time frames. The final result is a web-based UI that shows key topics each speaker addresses during the meeting as shown in Figure 5.

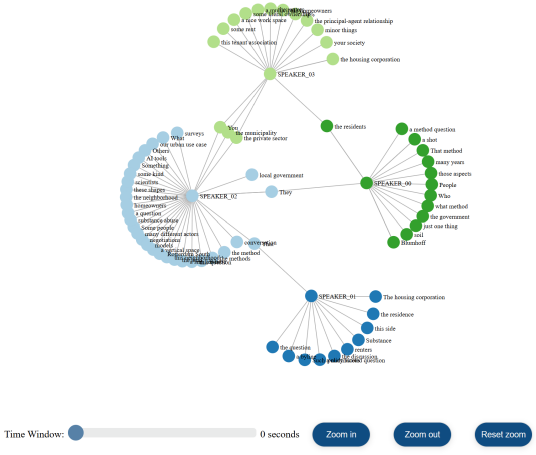

Figure 5: Data visualization graph

5 Results

To evaluate the results of this research the meeting audio data from the dataset of 4 meetings conducted with the Shape language is parsed through the experimental setup. The metrics this research focuses on are: How many topics are extracted from the meeting audio? How many key topics are discussed in each meeting? How many of the key topics does the experimental software recognize and with what frequency? After evaluating the data produced by the experimental setup the following results were obtained:

Group	Total number of extracted topics in summary	Number of key topics	Correctly identified key topics (%)	Total number of occurrences of key topics
Group 1	93	18	33.33%	9
Group 2	78	16	37.50%	10
Group 3	87	12	58.33%	12
Group 4	92	13	46.15%	11

Table 1: Key topics extraction results

As shown in Table 1, the experimental framework identified between 78 and 93 topics discussed during the meetings. Considering that each meeting lasted between 12 and 15 minutes this means that the framework identified approximately 6 discussed topics per minute. As the meetings had 4 participants each this is to be expected and is a good balance between identifying too many keywords and possibly missing entire discussion points. Each meeting also covered between 12 and 18 key topics. The amount of topics depended on the complexity of the question that needed to be discussed during a meeting. From a list of key phrases the framework identified between 33% and 58% of the key phrases. This indicates that the framework is able to successfully identify the key phrases. Which key phrases were identified out of a list of key phrases for each meeting is shown in Figure 6.

Groups:	Key topics:
Group 1	**climate risks**, **real-estate markets**, **house**, flood risk, urban environment, pole rot, **land subsidence**, **climate adaptation**, elevation, sea waves, dikes, **protection**, floating house, coastline, retreat strategy, inhabitants, migration, protected area
Group 2	**residents**, **policy actors**, subsidence, Bloemhof, **Rotterdam**, fragile situation, dynamic situation, environment, infrastructure, social housing renters, private homeowners, **municipality**, **financial sector**, **private sector**, real-estate developers, insurers
Group 3	**communities**, **real estate companies**, **adaptation**, **risks**, **hazards**, **climate hazards**, barriers, **perception**, pathways, society, integration
Group 4	**flood risk adaptation**, **climate adaptation**, **flood adaptation strategies**, maladaptation, **multilayer systems**, **houses**, subsidence, river, bottleneck, **tipping point**, decisions, black box, decentralized decision making

Figure 6: Table showing key topics per group. Identified topics are displayed in bold blue text.

Moreover, as shown in Figure 7, most key phrases were identified multiple times, with one topic being identified 7 times. This further supports the fact that the proposed speech

recognition and natural language processing system is able to extract key meeting topics well. It also confirms that the system not only identified the key phrases correctly and intentionally but can identify them consistently when spoken during the meetings.

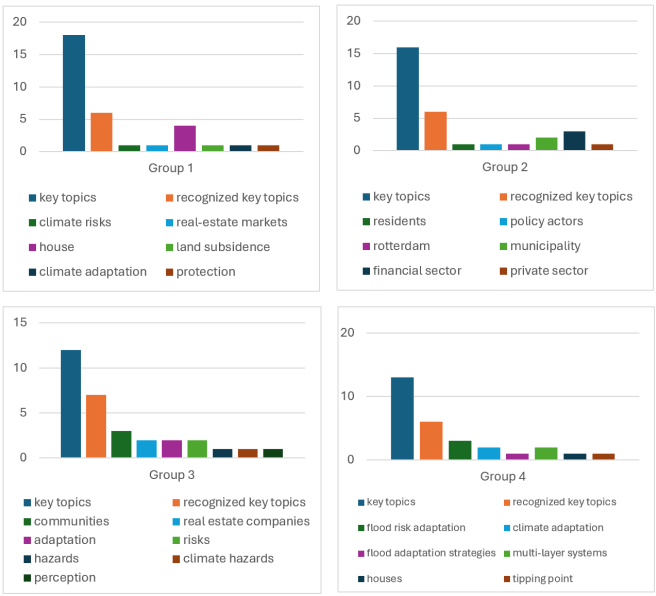

Figure 7: Frequency of correctly identified key topics per group.

6 Discussion

The results of the experiment demonstrate that the framework is able to extract topics spoken during meetings and is also able to identify 33-58% of the key topics multiple times. These numbers show that the experimental framework is able to retain meeting key topic information better than average individuals who tend to forget over 60% of newly attained information within 24 hours and even more as time goes on [6]. While this indicates a degree of success it also raises the question of why some of the key topics were missed.

One of the potential reasons for these results is the nature and context of the meetings themselves. Namely, the meetings were centered around open-ended complex questions that had no final answer. Additionally, there is the human factor to consider. Meeting participants usually do not know how to structure meetings well and tend to go off-topic during meetings [1]. This means there is a possibility that some of the key topics were never addressed during a meeting thus the framework had no way of identifying them. Another possible reason for the results could be due to people having a tendency to rephrase key ideas or use synonyms that were not recognized by the framework.

An additional angle to explore in an attempt to answer why some key topics were not identified are the limitations of the system itself. One such limitation is the Whisper model the system uses could have struggled with transcribing the speech from participants due to some of the audio data being noisy and all participants being non-native English speakers. Additionally, the used Whisper model is a pre-trained model with an average word error rate (WER) of 10% [7]. This means that some words from the dataset could have been incorrectly transcribed. Another observation during the experiment is that obscure words that rarely occur in the training data corpus were at times replaced with other similar words such as ”subsidence” (”soil subsidence”) being transcribed as substance. This is expected behavior due to the model not having enough training data for those specific words.

7 Conclusions and Future Work

This paper explores the possibility of combining Automatic Speech Recognition (ASR) models with Natural Language Processing tools for the purpose of extracting key meeting topics for a meeting summary. The experimental setup demonstrates that the proposed system can identify 33% to 58% key meeting topics confidently. These results show that the proposed system can be used to improve user understanding and retention of information acquired during a meeting. The abstract and complex nature of the meetings held shows that the system can be used for meetings covering various topics of various complexity.

While the presented results are quite promising they raise new questions such as how the system could be improved and how an effective user interface could be designed.

To address these questions, further research should be undertaken in several key areas. This includes research into system optimization to improve performance, conducting extensive user studies to gain a deeper understanding of user behavior and comprehension during meetings, and an exploration of web user interface design principles to create an effective interface for visualizing summarised meeting data.

8 Responsible Research

While conducting research it is important to consider the ethical implications of said research along with the integrity and reproducibility of the achieved results.

One of the ethical implications observed during this research is the fact that the meetings were conducted with human participants. However, the dataset used was collected for previous research so all of the participants have already given permission for the meeting data to be used. Furthermore, their faces and names are not visible anywhere in the dataset, and the dataset will not be shared with anyone outside of this research. Additionally, after the research is completed, the recordings will be deleted. Another ethical implication to consider is that this research builds upon preexisting software. The researcher was careful to only choose models and software with an appropriate open-source license that allows for them to be used in research. To comply with the conditions of these licenses, and therefore respect the wishes of the original authors, all required license files are included in the codebase.

Regarding integrity and reproducibility of the research results the results were obtained by passing the meeting audio data through the system without modifying or altering the data. Additionally, the system proposed in this paper builds upon several models and tools which are all publicly available and have an open source license as stated above. This ensures

that these models and tools can be accessed and used in future research. In order to ensure reproducibility the methodology and experimental design are described transparently and in detail in Section 3 and Section 4.

References

[1] Linda A. LeBlanc and Melissa R. Nosik. Planning and leading effective meetings. *Behavior Analysis in Practice*, 12(3):696–708, 1 2019.

[2] Joseph E. Mroz, Joseph A. Allen, Dana C. Verhoeven, and Marissa L. Shuffler. Do we really need another meeting? The science of workplace meetings. *Current Directions in Psychological Science*, 27(6):484–491, 10 2018.

[3] Caterina Villani, Matteo Orsoni, Luisa Lugli, Mariagrazia Benassi, and Anna M. Borghi. Abstract and concrete concepts in conversation. *Scientific Reports*, 12(1), 10 2022.

[4] A. Kamp. "Speaking: Part I - Speaking architecture / Part II - Speaking architecture / Part III - Speaking through form , 2018.

[5] Johanna Lönngren and Katrien Van Poeck. Wicked problems: a mapping review of the literature. *International Journal of Sustainable Development World Ecology*, 28(6):481–502, 12 2020.

[6] Jaap M. J. Murre and Joeri Dros. Replication and analysis of Ebbinghaus' forgetting curve. *PLoS ONE*, 10(7):e0120644, 7 2015.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 23–29 Jul 2023.

[8] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH 2023*, pages 1983–1987, Dublin, Ireland, August 2023. ISCA.

[9] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio, 3 2023.

[10] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[11] Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Building Efficient Universal Classifiers with Natural Language Inference, December 2023. arXiv:2312.17543 [cs].

[12] Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. A Sliding-Window approach to automatic creation of meeting minutes, 4 2021.

[13] Yongxin Zhou, Fabien Ringeval, and François Portet. Can GPT models Follow Human Summarization Guidelines? Evaluating ChatGPT and GPT-4 for Dialogue Summarization, 10 2023.

[14] Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende, editors, *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 620–628, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[15] Yurdaer Doganata and Mercan Topkara. Visualizing meetings as a graph for more accessible meeting artifacts. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 1939–1944, New York, NY, USA, 2011. Association for Computing Machinery.

[16] Lea Fischbach. A comparative analysis of speaker diarization models: Creating a dataset for German dialectal speech. In Oleg Serikov, Ekaterina Voloshina, Anna Postnikova, Saliha Muradoglu, Eric Le Ferrand, Elena Klyachko, Ekaterina Vylomova, Tatiana Shavrina, and Francis Tyers, editors, *Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024)*, pages 43–51, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[17] ExplosionAI. en_core_web_lg: spacy models, n.d. Retrieved January 14, 2025.

[18] What is Zero-Shot Classification? - Hugging Face, 9 2023.

[19] Seungeun Park, Betty Bekemeier, Abraham Flaxman, and Melinda Schultz. Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. *Informatics for Health and Social Care*, 47(2):175–193, 9 2021.

[20] Mike Bostock. D3.js - data-driven documents, 2012.

[21] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[22] Marco Siino, Ilenia Tinnirello, and Marco La Cascia. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342, 2024.