PROJECT REQUIREMENTS SPECIFICATION

"News on the Go" - A News Video and Text Summarization and Translation Service

UE20CS390A – Project Phase – 1

Submitted by:

Name	SRN
Rohit V Shastry	**PES2UG20CS282**
Rushil Ranjan	**PES2UG20CS288**
Sai Hardik Sriram Talluru	**PES2UG20CS296**
Niranjan Rao SS	**PES2UG20CS226**

Under the guidance of

Prof. Shilpa S
Professor
PES University

January - May 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

TABLE OF CONTENTS

1. Introduction

In today's information age, keeping up with the latest news and developments is more important than ever. However, with so many news sources available and so much content to sift through, staying informed can be a daunting and time-consuming task. That's why we're developing a phone app that aims to make staying up-to-date with the latest news faster, easier, and more convenient than ever before.

Our app takes a unique approach to summarizing news content. Using advanced algorithms and artificial intelligence, our app analyzes news videos and generates concise and accurate summaries of their key points. This makes it easy for users to get the gist of a news story quickly and efficiently, without having to spend time watching a full video or reading a long article.

Additionally, our app transcribes the videos into text and then generates a summarized text summary, providing even more convenience and accessibility to our users. And for those who prefer to read news in their native language, our app also includes a translation feature that can translate both the video summaries and text summaries into multiple languages.

But our app isn't just about making it easier to stay informed. We're also committed to making the app intuitive, user-friendly, and accessible to as many people as possible. Our phone app has a sleek and simple interface, making it easy to navigate and use. Additionally, our app is designed with accessibility in mind, ensuring that users with disabilities can use it with ease.

Overall, our app represents a revolutionary new way to stay informed in today's fast-paced world. Whether you're a busy professional who doesn't have time to keep up with the news, a student who wants to stay informed on the latest developments, or anyone in between, our app makes it easy to stay up-to-date with the latest news, all in one convenient location.

1.1. Project Scope

Purpose

The purpose of our project is to develop a phone app that summarizes news videos into short and concise summaries, transcribes the videos into text, and summarizes the text into an even shorter summary. Our app also includes a translation feature that can translate both the video summaries and text summaries into multiple languages. The app aims to make it easy for users to stay informed and up-to-date with the latest news from around the world, all in one convenient location.

Benefits

- Helps users save time by providing them with concise and accurate summaries of news stories.
- Provides accessibility to news content in multiple languages.
- Convenient and user-friendly interface for easy navigation and use.

Objectives

- Develop advanced algorithms and artificial intelligence to generate accurate and concise summaries of news videos.
- Integrate a translation feature to make the app accessible to users in multiple languages.
- Develop a sleek and user-friendly interface for the phone app.

- Ensure accessibility for users with disabilities.

Goals

- Create an app that provides users with fast and convenient access to news summaries in multiple languages.

- Ensure that the app is intuitive, user-friendly, and accessible to as many people as possible.

- Continuously improve the app through user feedback and updates to the underlying technology.

Coverage

Our app covers a wide range of news topics and sources, providing users with access to the latest news from around the world. The app aims to provide coverage that is relevant and accurate, with an emphasis on concise and accurate summaries.

Limitations

- The accuracy of the news summaries may vary depending on the quality of the source material and the algorithms used to generate the summaries.
- The app may not be able to cover every news topic or source, depending on the availability of relevant content.
- The translation feature may not always provide perfect translations, depending on the complexity of the language and the accuracy of the underlying translation technology.

2. Literature Survey or Existing System

1. Multimodal Video Summarization via Time-Aware Transformers

29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China. ACM

With the growing number of videos in video sharing platforms, how to facilitate the searching and browsing of the user-generated video has attracted intense attention by multimedia community. To help people efficiently search and browse relevant videos, summaries of videos become important. The prior works in multimodal video summarization mainly explore visual and ASR tokens as two separate sources and struggle to fuse the multimodal information for generating the summaries.

In this paper, the team proposes to leverage the inherent time information inside video for better multimodal video summarization, which has been commonly ignored in the previous works. Therefore, the team proposed a Time-Aware Multimodal Transformer (TAMT) for multimodal video summarization. In order to explore the time information thoroughly, they introduce a novel short-term order-sensitive attention mechanism in the encoder of the TAMT. Extensive experimental analyses on YouCookII and How2 validated the superiority of TAMT over the state-of-the-arts, and the effectiveness of our idea to utilizing the associated timestamps to organize the multimodal signals in videos. For future work, it is potential to explore TAMT in learning better visual-linguistic representations in a self-supervised manner.

2. Unsupervised Extractive Text Summarization with Distance-Augmented Sentence Graphs

This paper proposes an unsupervised approach to extractive text summarization that does not rely on large quantities of human-generated summaries. Instead, the approach uses an automatically constructed sentence graph for each document to select salient sentences for summarization based on similarities and relative distances in the neighborhood of each sentence. The paper also extends the approach to the multi-document setting by aggregating document-level graphs via proximity-based cross-document edges. The experiments on benchmark datasets show that the proposed approach achieves competitive or better results than previous state-of-the-art unsupervised extractive summarization methods in both single-document and multi-document settings, and its performance is competitive with strong supervised baselines. Overall, the paper aims to provide an effective and practical unsupervised approach for extractive text summarization that does not require human-generated summaries.

Advantage - The advantage of this paper is that it proposes an effective and practical unsupervised approach for extractive text summarization that leverages automatically constructed sentence graphs and achieves competitive results without relying on large quantities of human-generated summaries.

Limitation - One potential disadvantage of this paper is that, while the proposed approach achieves competitive results, it may not perform as well as supervised methods that have access to large amounts of human-generated summaries.

3. Supervised Video Summarization Via Multiple Feature Sets With Parallel Attention – IEEE, 2021

In this paper, Multi-Source Visual Attention (MSVA) model is used. The task of the model is to produce an output that represents the importance score of frames. In order to extract features to represent frames in videos, pre-trained models like GoogleNet or content-based image and motion features are used. Once the features are extracted, an attention mechanism is employed followed by a couple of linear layers with different features and fuse them to obtain a common embedding space to represent frames. After fusion, linear layers, normalization, activation functions, and the importance score of given input frames is applied. Then the visual feature extraction, the attention mechanism, and fusion techniques to incorporate visual features from multiple sources for video summarization are described.

Advantages:

It can be adapted to different types of video content and summarization tasks. It allows for multiple feature sets to be attended to simultaneously. This improves the quality of the generated video summary.

Limitations:

Requires a large amount of labeled data to train the neural network. Use of multiple feature sets and attention mechanisms makes it computationally intensive.

4. Unsupervised Video Summarization via Multi-source Features

ACM , 2021

In this paper, MultiSource Chunk and Stride Fusion (MCSF) model is used. The frames in videos are subsampled at the rate of two frames per second where each frame's features are extracted using the respective visual encoder model.

The input features are fed through different layers and finally fused. Chunks and strides are fed to a Bidirectional Long-short Term Memory (Bi-LSTM) and linear layers and the output is summed up with the difference attention.

The end results are probabilities to select frames in the summary.

There were three fusion techniques used:

1)Early fusion: The fusion is applied at the feature-level

2)Intermediate fusion: The fusion is applied after the different chunks and strides are fed through Bi-LSTM and linear layer.

3)Late fusion: The fusion is applied after both resulting importance scores from each computed dataset. The paper assesses the performance of compared methods using F1 scores on different splits of datasets.

Advantages - The use of chunk and stride fusion allows the algorithm to capture both global and local information in the video. This results in more accurate and representative video summaries.

It is scalable and can be applied to large volumes of video content.

Limitations - Cannot work well with complex videos that have multiple subplots or scenes with different visual or audio characteristics. It may be difficult to identify the most representative parts of the video.

5. GPT2MVS: Generative Pre-trained Transformer-2 for Multi-modal Video Summarization.

ACM, 2021

The aim of this paper was to create a multi-modal video summarization using an explicit text-based query.

To handle this task, a new strategy was presented that uses a specialized attention network and contextualized word representations.

A contextualized video summary controller, multi-modal attention mechanisms, an interactive attention network, and a video summary generator were all part of the suggested model.

Advantages:

1) A new end-to-end deep model for multi-modal video summarization is introduced, based on a specialized attention network and contextualized word representations.

2) One biggest advantage it received was using contextualized word representations over the Bag of Words method.

3) The experimental results showed that the proposed model was effective with an increase of +5.88% in accuracy and a +4.06% increase in F1-score, compared with the state-of-the-art method present.

Limitations:

1) Since video data does not only have a visual channel but also a speech channel, does not use speech input.

2) Scalability: The proposed approach requires significant computational resources, making it difficult to implement in resource-constrained environments.

3) Generalizability: While the proposed approach is effective across a wide range of video genres and styles,

it may not be as effective for certain types of videos, such as those with complex visual content or those with non-standard language use.

6. Speech Recognition and Machine Translation Using Neural Networks

The paper discusses the study of using deep recurrent neural networks LSTM for English-Russian translation from speech to text. The study involved compiling a neural network learning dataset from

English audiobooks and Russian text, extracting features from audio files, and using TensorFlow algorithms for learning and validating the neural network. The study found that deep neural networks

were effective for machine translation of language.

The paragraph discusses the use of modern deep learning approaches such as long short-term memory network, encoder-decoder architecture, attention mechanism, and ray search in a translation model.

The model was trained on English audio files and corresponding Russian texts, but the correct translation was not revealed due to the small size of the learning sample, indicating the need for further

development with a larger learning sample and different parameter values. The calculations were carried out on GPUs using the TensorFlow software environment.

Advantages

Text to text and Speech to text both translations can be done using the model.

Limitations

The need for learning in artificial neural network is time consuming as the number of hidden layers and neurons grows, the time it takes to train and evaluate a new instance increases exponentially

3. Product Perspective

Market Need:

People want a quick and easy way to stay informed about the news. This product addresses that need by providing summarized news videos and text.

Target Market:

News consumers who are short on time, multilingual users, and busy professionals who want to stay informed.

Unique Selling Proposition (USP):

The product offers both video and text summarization, along with translation into different languages, making it a comprehensive news solution.

Competitive Advantage:

The product's comprehensiveness gives it a competitive advantage over other news summarization products in the market.

Revenue Model:

Subscription-based model or targeted advertising/sponsorships.

3.1. Product Features

Video Summarization:

- Automatically summarizes news videos into concise and accurate summaries.
- Utilizes advanced algorithms and artificial intelligence to ensure accurate reporting.
- Provides users with a text-based transcription of the news video.

Text Summarization:

- Generates a text summary of the news video based on the transcribed text.
- Ensures that the summary is concise, accurate, and easy to understand.

Translation:

- Allows users to translate both the video and text summaries into multiple languages.

- Provides accurate translations using advanced machine learning algorithms.
- Helps users to stay informed about global news, regardless of language barriers.

Accessibility:

- Designed to be accessible to users with disabilities.
- Includes features such as text-to-speech and high contrast mode.
- Ensures that everyone can benefit from the app's features.

User Experience:

- Intuitive and user-friendly interface.
- Fast and reliable performance, with minimal loading times.
- Continually updated based on user feedback, with new features and functionality added over time.

Overall, our product provides a comprehensive set of features that allow users to access news summaries quickly and conveniently in multiple languages.

3.2. User Classes and Characteristics

1. General News Consumers
    - Regularly watch news videos and read news articles to stay informed.
    - May not have strong technical capabilities, but are comfortable using smartphones and apps.
    - Are interested in accessing news summaries quickly and conveniently.
2. Non-Native Speakers
    - May have difficulty understanding news videos and articles in their non-native language.
    - Are interested in using the translation feature to access news in their own language.

- May not have strong technical capabilities, but are comfortable using smartphones and apps.

3. Users with Disabilities
    - Have different needs when it comes to accessibility, such as text-to-speech and high contrast mode.
    - May not have strong technical capabilities, but are comfortable using smartphones and apps.
    - Are interested in using a product that is designed to be accessible to all users.
4. Researchers/Journalists
    - May require more detailed information from news videos and articles.
    - Are interested in using the text summary feature to quickly gather information.
    - May have stronger technical capabilities and require more advanced features.
5. Foreign Correspondents
    - Often need to stay informed about news in multiple languages.
    - Are interested in using the translation feature to quickly access news in different languages.
    - May have stronger technical capabilities and require more advanced features.

Overall, our product is designed to be accessible and useful to a wide range of users, from casual news consumers to researchers and journalists. We anticipate that users will have varying levels of technical capabilities, and have designed our product to be intuitive and user-friendly, while still offering advanced features for more advanced users. Our product is also designed to be accessible to users with disabilities, ensuring that everyone can benefit from its features.

3.3. Operating Environment

The system will operate on smart phones running on Android operating system. It will use APIs to access news videos and articles, and pre-trained machine learning models for video summarization, transcription, and translation. No additional hardware will be required.

3.4. General Constraints, Assumptions and Dependencies

Legal Implications

- Intellectual property infringement: This project will need to ensure that it is not infringing on any patents or copyrighted material when using APIs, pre-trained models, or Python libraries.
- Licensing: Some APIs, pre-trained models, and Python libraries may have specific licensing agreements that need to be followed.
- Usage limits: Some APIs and pre-trained models may have usage limits that need to be followed to avoid legal issues.
- Privacy laws: This project may need to comply with data privacy regulations, particularly if it collects and stores user data.
- Translation accuracy: The translations provided by the software tool may need to be accurate and not infringe on any intellectual property or trademark laws.
- Accessibility laws: The software tool may need to comply with accessibility laws to ensure that it is usable by people with disabilities.
- Liability: The software tool may need to address liability issues related to its usage and any harm that may arise from its usage.
- Data security: The software tool may need to ensure that any data accessed or stored is secure and not subject to unauthorized access or use.
- Local laws and regulations: The software tool may need to comply with local laws and regulations, which may vary depending on the location of the user or the developer.
- Terms of service: It is important to review the terms of service of APIs, pre-trained models, and Python libraries used in the project and ensure that they are being used in accordance with their respective terms of service.

Usage Limitations

- The accuracy of the video summarization and translation may vary depending on the complexity of the video and the quality of the translation algorithms used.

- The project may be limited to the types of news videos that can be effectively summarized and translated. Videos with highly technical or complex content may not be accurately summarized.
- The project may be limited to the languages that are supported by the translation algorithms used.
- The project may be limited by the usage limits of APIs, pre-trained models, and Python libraries used, which may require payment or have restrictions on usage.
- The app may have potential limitations in terms of the amount of user data that can be processed at a given time or the number of users that can access the tool concurrently.
- The app may have potential limitations in terms of the types of devices that can access it and the internet connectivity of those devices.

Dependencies

- **APIs**: The project may be dependent on third-party APIs to access news videos and perform text and video summarization, translation, and transcription. Examples of such APIs include the Google Cloud Translation API, Amazon Transcribe, and IBM Watson Speech to Text.
- **Pre-trained models**: The project may depend on pre-trained models for natural language processing, video summarization, and machine translation. Examples of such models include BERT, GPT-3, and OpenCV.
- **Python libraries**: The project may depend on various Python libraries for data processing, machine learning, and natural language processing. Examples of such libraries include NumPy, Pandas, Scikit-Learn, and NLTK.
- **Software development frameworks**: The project may be dependent on specific software development frameworks to build and deploy the app on different mobile operating systems, such as React Native or Flutter.
- **User feedback**: The project may depend on user feedback to improve the accuracy and quality of the video summarization, translation, and transcription. It may be necessary to collect and analyze user feedback to identify areas for improvement and adjust the algorithms accordingly.

Assumptions

- We assume that the news videos used in the project are available and accessible through third-party APIs or other online sources.

- We assume that the videos used in the project are in a format that can be processed by the algorithms and technologies used, such as MP4 or AVI.
- We assume that the quality and accuracy of the video summarization, translation, and transcription algorithms used are sufficient for the purposes of the project.
- We assume that the text summarization and translation algorithms used are effective in generating accurate and concise summaries and translations.
- We assume that the video and audio quality of the news videos used is sufficiently high to be processed accurately.
- We assume that users have a reliable internet connection to access the app and use its features.
- We assume that users are able to understand and use the app's interface without significant difficulty.
- We assume that users are willing to provide feedback on the app's performance and functionality in order to improve its accuracy and effectiveness.
- We assume that the project will not infringe on any intellectual property or copyright laws, and that we have the necessary permissions and licenses to use the news videos and other resources used in the project.
- We assume that the project will comply with all relevant privacy laws and regulations, and that user data will be collected and stored in a secure and ethical manner.
- We assume that the project will not cause any harm or damage to users or their devices, and that appropriate security measures are taken to protect against malware, viruses, and other potential threats.
- We assume that the project will be developed and deployed in a timely manner, within the constraints of available resources.

3.5. Risks

There are several risks that could potentially pose obstacles to the final project delivery, including:

Technology failures: Algorithms and technologies used may not work as expected, leading to inaccurate results or errors in app functionality.

Hardware failures: Hardware used to develop or run the app may fail or malfunction, resulting in delays or loss of progress towards project completion.

Version compatibility problems: Different versions of libraries, frameworks, and technologies may not be compatible, resulting in errors or other issues that could delay project delivery.

Security threats: The project may be vulnerable to various security threats, such as hacking, data breaches, and malware attacks, leading to loss or theft of user data or sensitive information.

Legal issues: The project may violate laws and regulations related to intellectual property, privacy, or other areas, leading to legal action that could delay or halt project delivery.

Resource constraints: The project may require significant resources, such as time and personnel, which may be limited or unavailable, leading to delays or reduced functionality in the app.

4. Functional Requirements

Video Search:

- The app should provide a search functionality for users to find news videos based on news channels, keywords, genre, and other parameters.
- The search results should be a list of already summarized videos that match the user's search criteria.

Video Playback:

- The app should allow users to select and play the summarized video from the search results.
- The app should display the summarized text alongside the video.

Transcription and Translation:

- The app should transcribe the audio of the summarized video to text.
- The app should provide the option to translate the text summary into different languages.

User Interface:

- The app should have a user-friendly interface for searching and playing summarized videos.

- The app should allow users to interact with the summary text by highlighting or selecting specific sections.
- The app should provide the option to save or share the summary text and video.
- The app should allow users to listen to the summarized news as an audio by plugging in earphones.

Error Handling:

- The app should be able to handle errors or issues that may arise during transcription or translation.
- The app should notify the user if there is an error or issue with the video or summary generation.
- The app should provide suggestions or recommendations for resolving the issue.

5. External Interface Requirements

5.1. User Interfaces

The user interface for this project will be a mobile application designed to provide a simple and intuitive user experience. It will comply with standard GUI elements and styles.

The main screen of the app will provide options for the user to select their preferred language and category of news. The layout of the screen will include a video player with playback controls, a summary text box, and options to translate the summary text into different languages. The screen layout will have a consistent and organized design.

The app will provide a help option in the menu bar to assist users with navigating the application. The inputs will be immediate and provide real-time updates on video and text summaries. The outputs will be presented on the same screen, allowing the user to easily switch between video and text summaries.

The app will have a feature to store user preferences and provide a personalized experience. Additionally, it will display error messages in clear and concise language with suggestions for corrective action.

5.2. Hardware Requirements

- The system will require a mobile device with internet connectivity, such as a smartphone or tablet.
- The mobile device should have sufficient storage space to download and install the app.
- The app will be available for download from the relevant app store for the user's mobile device operating system (e.g. Google Play Store for Android)
- The app will be compatible with various mobile device hardware components, such as the microphone.
- The app will utilize cloud-based storage for video and text summaries, requiring a stable internet connection for seamless operation.
- The app will be designed to optimize battery usage to prolong the user's mobile device's battery life.

5.3. Software Requirements

• Name and Description:

- Python Programming Language: used to develop the app
- Flask Web Framework: used to build the backend of the app
- TensorFlow: used for video summarization and natural language processing tasks
- Google Cloud Platform/ Amazon Web Services: used for cloud-based storage of video and text summaries

• Version / Release Number:

- Python 3.x
- Flask 1.x
- TensorFlow 2.x
- Google Cloud Platform/ Amazon Web Services

• Databases:

MySQL: used to store user preferences and settings.

• Operating Systems: Android for the mobile app.

• Source:

GitHub: the source code for the app will be stored on GitHub for version control and collaboration.

5.4. Communication Interfaces

The communication interfaces required for this project include:

- Internet connection for accessing and retrieving news videos.
- APIs for video summarization, speech-to-text conversion, text summarization, and translation.
- Local area network protocols for communication between the app and the server.
- HTTPS for secure communication and data transfer.
- Serial ports or USB for connecting to external devices such as headphones or speakers.

The communication standards requirements for this project include:

- High-speed internet connection for faster retrieval and processing of news videos.
- Large buffer size to accommodate the video and audio data during conversion and summarization.
- Standard protocols such as TCP/IP for reliable and efficient data transfer.

6. Non-Functional Requirements

6.1. Performance Requirement

- The system should be able to handle a minimum of 100 users simultaneously.
- The system should provide real-time summarization of news videos.
- The system should be able to process and transcribe the video summary into text within 10 seconds.
- The system should be able to summarize the text into a shorter summary within 5 seconds.
- The system should have an accuracy rate of at least 90% for summarization and transcription.

- The system should be able to translate the video summary and text summary into at least 5 different languages.
- The system should have a response time of no more than 2 seconds for user interactions.

Quality Attributes

- The system should be reliable and not produce errors more than once per week.
- The system should be robust and able to handle unexpected inputs.
- The system should be available 24/7 with no more than 2 hours of downtime per month

6.2. Safety Requirements

- The system should not produce any harmful outputs or content.
- The system should not store any user information or data without their consent.

6.3. Security Requirements

- The system should have secure authentication methods to ensure user privacy and prevent unauthorized access.
- The system should use encryption methods to protect user data and prevent data breaches.
- The system should comply with GDPR rules and other relevant data protection laws.

7. Other Requirements

Scalability: The system should be able to handle an increasing number of users and videos over time.

Compatibility: The system should be compatible with various devices, operating systems, and web browsers.

Accessibility: The system should be accessible to people with disabilities and meet relevant accessibility standards.

Maintainability: The system should be easy to maintain, update, and modify over time.

Availability: The system should be available and accessible to users at all times, with minimal downtime for maintenance or updates.

Appendix A: Definitions, Acronyms and Abbreviations

[Provide definition of all terms, acronyms and abbreviations required for interpreting this Requirements Specification.]

Appendix B: References

Multimodal Video Summarization via Time-Aware Transformers –
Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal Video Summarization via Time-Aware Transformers. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475321

Supervised Video Summarization via Multiple Feature Sets with Parallel Attention
Junaid Ahmed Ghauri, Sherzod Hakimov, Ralph Ewerth
https://arxiv.org/abs/2104.11530

Unsupervised Video Summarization via Multisource Features
Hussain Kanafani, Junaid Ahmed Ghauri, Sherzod Hakimov, Ralph Ewerth
https://arxiv.org/abs/2105.12532

GPT2MVS: Generative Pre-trained Transformer-2 for Multi-modal Video Summarization
ia-Hong Huang, Luka Murn, Marta Mrak, Marcel Worring1. 2021.
GPT2MVS: Generative Pre-trained Transformer-2 for Multi-modal Video Summarization. In Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan. ACM,
New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Unsupervised Extractive Text Summarization with Distance-Augmented Sentence Graphs - 44th ACM SIGIR CONFERENCE
Jingzhou Liu, Dominic J. D. Hughes, and Yiming Yang. 2021. Unsupervised Extractive Text Summarization with Distance-Augmented Sentence Graphs.

In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463111

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., and Norouzi, M. SpeechStew: Simply mix all available speech recognition data to train one large neural network. arXiv preprint arXiv:2104.02133, 2021.

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., et al. BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. arXiv preprint arXiv:2109.13226, 2021.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. arXiv preprint arXiv:2106.06909, 2021.