Project proposal

● Graded

Group

ELOUDRHIRI Rayane
DELPORTE Guillaume
LOUIS Arthur
View or edit group

Total Points

20 / 20 pts

Question 1

Accept

20 / 20 pts

✓ **– 0 pts** Accept

– 20 pts Reject

Your project proposal has been accepted.

We advise you to target speech-to-text as it is already a very hard task. Note that generating text is hard, so you should consider using a per-trained language model and condition it on audio. Take a look at CLIP (https://openai.com/research/clip) and Whisper (https://openai.com/research/whisper).

Your assistant is François Rozet (office I.112). Consider looking deeper into the literature. Don't hesitate to ask questions and start early: deep learning takes more time than you expect.

INFO8010 Project Proposal: Closed Captioning

Delporte Guillaume,[1] **Eloudrhiri Rayane**,[2] and **Louis Arthur**[3]

[1] *guillaume.delporte@student.uliege.be (s191981)*
[2] *rayane.eloudrhiri@student.uliege.be (s193009)*
[3] *alouis@student.uliege.be (s191230)*

I. INTRODUCTION

In this project, we would like to implement a model that we called `DeepSpeakCC` capable of enabling dynamic close captioning and then apply that to children's animation movies.

In recent years, the advancement of deep learning techniques has revolutionized various fields, including natural language processing (NLP) and computer vision. Leveraging these advancements, this project proposal aims to introduce an innovative solution for close captioning and character localization in children's animation movies.

The ability to provide accurate and dynamic subtitles for audio content, coupled with the capability to identify the speaker within the video frame, holds significant potential in enhancing accessibility and user experience for diverse audiences, including those with hearing impairments or language barriers.

Additionally, such technology could greatly benefit educational settings, where children's comprehension and engagement with audiovisual content play a crucial role in learning outcomes.

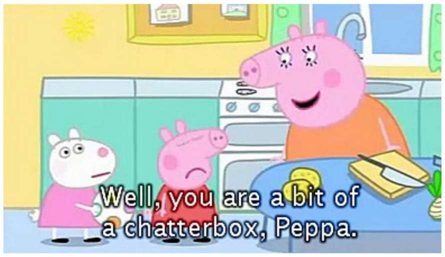

FIG. 1. Frame from the children show *Peppa Pig* with exemplified closed captions.

II. DATASET

The dataset chosen by our team will be drawn from Mozilla's Common Voice [1], which amasses an extensive collection of voice recordings from thousands of contributors worldwide. This project distinguishes itself by capturing voices across a lot of languages and accents, contributing to one of the most diverse and inclusive voice datasets available (this is their claim). We settled on focusing on English, the richest and most extensive language provided by the dataset.

As described in the paper on Common Voice [2], the dataset is organized into a collection of six Tab-Separated Values files, alongside a designated sub-directory for audio clips. These TSV files are instrumental in segmenting the voice data into distinct categories, each supplemented by: `[client_id, path, sentence, up_votes, down_votes, age, gender, accent]`. The first three features corresponds to a unique speaker identifier through an anonymized ID, specifies the audio file's storage path, and a text record of the sentence. The additional data can be used to assess the quality of the narrated sentence as well as some optional demographic features that could be used to specify the type of accent or dialect.

Concerning the English division of this dataset, as of March 2024, The age distribution among contributors is varied, with a significant portion falling within the 20 - 29 age range at 25%, followed by 14% within the 30 - 39 bracket, 9% aged 40 - 49, 6% under 20, 5% between 50 - 59, 4% within the 60 - 69 category, and 1% aged 70 - 79. The gender distribution among contributors is predominantly male at 45%, with females representing 17% and other genders making up 2% of the total.

Although these numbers have to be investigated as each person contributing is not forced to give this data and thus a big part of the distribution remains unknown. The Common Voice Corpus 16.1 spans a substantial 80.8 GB, comprising 3,438 recorded hours of which 2,586 hours have been validated. The dataset is open-source, evidenced by its CC-0 license, making it a good choice for research purposes.

III. HARDWARE REQUIREMENTS

As the size of the dataset (≈ 80GB), the depth of the needed ConvNet as well as the number of inputs being substantial, a lot of GPU computing power will be needed to train the model as well as a good amount of disk space.

We therefore would be keen on using the Alan cluster to be able to train our model efficiently. Another option would be to be allowed to use Google Collab Pro to be able to leverage the special space and computing power features in order to be able to complete training. Both of these approaches, at first glance, seem to be excellent for collaboration purposes as the team can access the data and the notebooks simultaneously.

IV. NICE TO HAVE

In this section, we will briefly cover some *nice-to-have* additions to the considered problem:

1. Automatic Text placement: Identify the speaker within the video frame to be able to create a bonding box and automatically place the caption at the right place on the screen
2. Automatic Translation: The use of a transformer network or some API to translate and adapt our captioning to multiple languages.
3. Automatic Content Filtering for Age Appropriateness: Implementing a feature that automatically adjusts the captions based on the viewer's age or a specified content sensitivity setting.
4. Character Recognition and Naming: Beyond identifying the speaker within the video frame, developing an algorithm that can recognize recurring characters and automatically tag their dialogues with their names or identifiers.

V. BRIEF REVIEW OF RELATED WORK

The advancements in the field of speech processing have been thoroughly sped up thanks to the steady breakthroughs in deep learning. Indeed, the use of multiple processing layers led to models capable of understanding intricate features from human speech. Understanding speech is a topic sought after and studied by many which have shown the proficiency of CNN and Transformers-based deep learning model.

Mehrish et al. [3] highlight the state-of-the-art deep learning techniques used in speech-processing applications and offer a comparative description of each (CNN and Transformers).
Furthermore, Kheddar et al. [4] provide an exhaustive survey of Deep Learning approaches spanning the years 2016 to 2023. It provides insights on the use of different deep learning approaches such as Federated Learning, Deep Reinforced Learning, Transfer Learning as well as Transformers.

Thus, our focus will primarily be on CNN and the Transformers as it is course material and widely mentioned in the literature.

[1] Mozilla Foundation. Common Voice Dataset. https://commonvoice.mozilla.org/en/datasets, 2024. Accessed: 2024-03-18.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4218–4222, Marseille, 11–16 May 2020, 2020. European Language Resources Association (ELRA).

[3] Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869, 2023.

[4] Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. Automatic speech recognition using advanced deep learning approaches: A survey, 2024.