RINGR Podcast Analyzer with AI

Team 70 - Bhuvaneshkumar Radjendirane, Elliot Couvignou, and Sai Rajesh
ECE 445 Project Proposal - Spring 2020
TA: Shuai Tang

1 Introduction

1.1 Objective

To preface this proposal this project was sponsored by RINGR who came to us with their objective. Their service is focused on high-quality multi-agent podcast recording regardless of each person's separate location [1]. An issue about podcast recording is that the amount of time to edit audio takes too long and requires use of some external audio editing software. This makes people who want to get into podcast recording to have to be aware of the technicalities of editing audio which drives away a big audience. There are also many more issues that require even more advanced technology use from users, such as background noise, to further touch up their audio. Since RINGR as a service helps with the initial step of recording the audio, they have the potential to use audio processing with the raw audio to help bring down these technical barriers as well as saving time. Modern day audio processing and speech recognition through products such as Amazon Alexa and Google Assistant proves that these issues can be dealt with in a commercial setting. From this RINGR hopes to use artificial intelligence (AI) as their main approach to their issue as this is what the successful products use [2]. Since this company is a startup with limited technical knowledge on these specific topics they decided to allow this project to be sponsored to UIUC.

Our goal is to take in the raw podcast recordings and analyze them with the AI model to slice out unwanted portions as well as filtering out unnecessary noise. The scope of this project is large so our main focus is to focus on recognition of spoken words in transcript form as well as some recognition of setting. Once this component is solid then any other feature like filtering and context based slicing can be done since we know the context and semantics of the audio. This project also provides a side effect of transcripting audio which can be extremely useful for use as a new feature.

1.2 Background

Podcasting as a business has seen tremendous growth over the past couple of years with music streaming platforms like Spotify increasing their podcasting budgets to where their recent acquisition of The Ringer podcasting network "[brought] the cost of Spotify's podcast shopping spree up to $600 million"[3]. With the market for podcast streaming being higher than before, more people are wanting to get into the podcasting industry. The aforementioned technical barrier that exists with editing audio demotivates most people from trying to push their first cuts as more focus is placed editing the content than actually providing it.

Estimates for how long audio editing takes varies depending on the person or team doing it but it is commonly known to be magnitudes larger than the raw audio length. This process has become cumbersome to the point where companies like “We Edit Podcasts” have grown specifically focusing on editing podcasts for other creators [4]. From this, automating even a small portion of this tedious process can prove to be commercially desirable.

1.3 Physical Design

Since this is a software-only project we will go over how a user is expected to run the finished product. Below on Fig. 1 is an image of the app running during a normal podcast recording. We can use our AI model to start transcription at the start of recording so once the recording is finished we should prompt the user of our overall transcript with little to no delay. From there we will have further user inputs to configure our subject recognition AI to edit the inputted audio and keep the user’s ‘good’ parts. The app currently runs on all mobile platforms and on desktop browsers so we should expect ours to run on all of these.

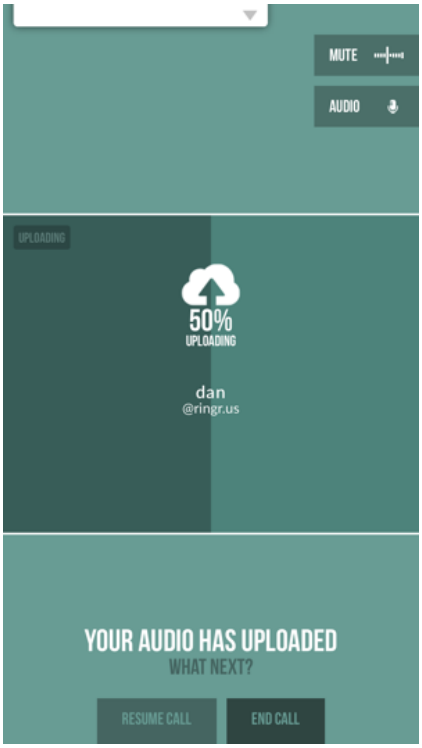


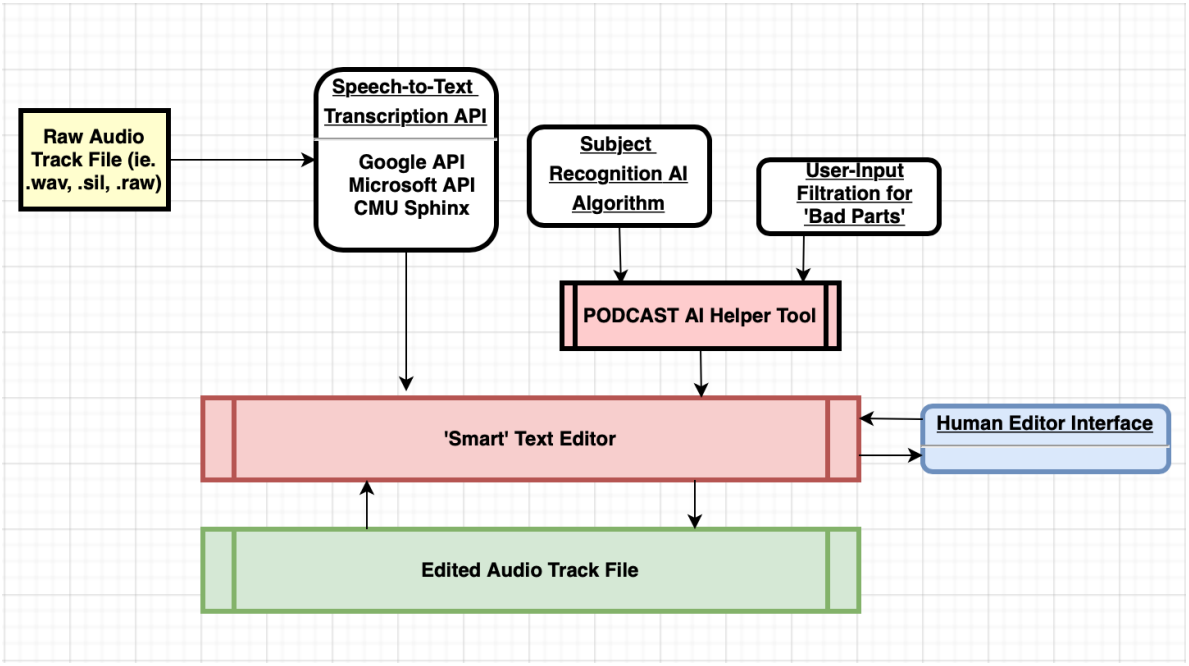
Fig. 1: RINGR mobile app during recording

1.4 High-Level Requirements List

- Size of AI models used should overall be below 1G to allow offline use of features through local storage. If size is too large to get actual results then we need to move this brain onto a server while making sure latency reasonably low.
- Runtime of transcription and analysis should at least be quicker than the length of time of the audio input. Our runtime should be more linear (O(n)) to audio input length than anything else.
- Transcription accuracy should be below 15% word error rate. Google currently has theirs at around 5% so we can't get too low. $WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$ [5]. *Definitions for these variables are in Reference.*
- If transcription happens during recording and in real-time, we need to make sure that our frame length is still enough to distinguish frequencies necessary to identify formants. From this we want a minimum frame length of 32 = 17 freq bins up to a max frame length of 4096 = 2049 freq bins.
- Our sampling rate for transcription should be set to make sure we can resolve human speech. High quality speech audio can be heard only from 50-8kHz range so we should sample around 16,000 samples per second for full resolution.

2 Design

2.1 Block Diagram



2.2 Functional Overview

2.2.1 Speech-to-Text Transcription API

Once we have our raw audio tracks, we will be utilizing a transcription API such as CMU Sphinx or Google's Speech-to-Text API. This will allow for a transcribed version of the raw audio to be available.

Requirement: Transcription accuracy should be below 15% word error rate.

2.2.2 Subject/Topic Recognition AI Algorithm [Podcast AI Helper Tool]

The algorithm we will develop will have to use some form of NLP to extract topics or subjects from the transcripts of our audio track files. LDA (Latent Dirichlet Allocation) for topic modeling is one of the more promising avenues we are exploring for this functionality [6].

Requirement: AI should successfully identify >50% of total topics in a transcript

2.2.3 User-Input Filtration for 'Bad Parts' [Podcast AI Helper Tool]

We would like to offer the user a selection of filters (i.e stop words filter, 'um/uh/' filter, etc.) which they can toggle on their audio transcripts to filter out unwanted parts. We will be using the Python NLTK and creating NLP pipelines to develop these features.

Requirement: Filters should not affect 'valuable' parts of transcript and should successfully identify >75% of filtered item in a transcript.

2.2.4 'Smart' Text Editor

The editor will be filled out with the result of the transcriber, is presented to the user and is always referenced by the Subject Analyzing AI (SA-AI). Any changes made to this transcription in the editor will reflect in the final post production podcast. We will be using timestamps to reference word/phrase edits in the text editor to their appropriate times/locations in the audio file to reflect the changes made.

Requirement: The 'Smart' Text Editor must also ensure that the recording does not sound 'fragmented' after deleting segments of the podcast.

2.2.5 Human Editor Interface

The Human Editor Interface will be the UI through which users will be able to choose edits to add to the audio. Users will also be able to access and use the 'Smart' Text Editor via this component of our project.

Requirement: Design is left to the UI team at RINGR.

2.3 Risk Analysis

There isn't much quantitative risk analysis to be made but there are general ones that relate to our code of ethics. The main risk users face is the potential for private data being released without their consent. We must ensure that data is kept private to users and only made public once permission is given.

Our ears are pretty sensitive so it is also our responsibility to ensure that the volume of the output audio remains at the same level as the input. There are audio editing programs and streaming platforms that don't initially normalize imported audio which can lead to extremely loud and harmful noises. We want to reduce the risk of sudden abrupt jumps in volume as much as possible.

3 Ethics and Safety

Since this is a software only project, there really isn't much to consider for safety except for unethical use of audio data. Raw podcast recordings and the output generated by our project "should only use personal information for legitimate ends and without violating the rights of individuals and groups". [7]Likewise we also don't want to show any training data used as this goes with the privacy issue. If we do end up uploading the AI models to a remote server, another privacy issue arises since we need to ensure privacy of data on both ends.

It is also our responsibility to cite any resources and libraries used throughout this process in our efforts "to properly credit the contributions of others"[8]. It is apparent that we can't construct this project from scratch so it is in our best interest to ensure that we respect both the work and it's right for privacy. Any research that is used for some significant factor of our AI model and how we train it should also be cited when necessary.

Since an "essential aim of computing professionals is to minimize negative consequences of computing, including threats to health"[7], our project must serve the purpose of alleviating editing time and strain from the users. In other words, using our AI should be less tedious for editing audio data than using editing software.

RINGR as a company already holds their own code of ethics in regards to discrimination which follow in line with IEEE and ACM. The ACM ethics code stating that "Computing professionals should foster fair participation of all people, including those of underrepresented groups"[7] means that our AI model shouldn't place any extra focus on particular users. The ideal goal of this project would be to include recognition of all languages, however our primary focus for now is on english since this composes most of our training data and consumers.

References

[1] ringr.com, “How It Works”2020. [Online]. Available: https://www.ringr.com/

[2] towardsdatascience.com “How Amazon Alexa works” 2020. [Online]. Available: https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3

[3] fool.com “Why Spotify Is Buying Bill Simmons' The Ringer” Feb, 2020. [Online]. Available: https://www.fool.com/investing/2020/02/13/why-spotify-is-buying-bill-simmons-the-ringer.aspx

[4] weeditpodcasts.com, “Home Page” 2020. [Online]. Available: https://www.weeditpodcasts.com/

[5] martin-thoma.com, “Word Error Rate Calculation” 2020. [Online]. Available: https://martin-thoma.com/word-error-rate-calculation/

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (n.d.). *Latent Dirichlet Allocation* (Vol. 3). Journal of Machine Learning Research. Available: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[7] acm.org, “ACM Code of Ethics and Professional Conduct” 2020. [Online]. Available: https://www.acm.org/code-of-ethics. [Accessed: 13-Feb-2020].

[8] Ieee.org, "7.8 IEEE Code of Ethics" 2020. [Online]. Available: https://www.ieee.org/about/corporate/governance/p7-8.html. [Accessed: 13- Feb- 2020]