Automatic Captioning Software from any Audio Source

A Technical Report
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Edward J. Clark

May 10, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Edward Joseph Clark

Primary Technical advisor: Lu Feng, Department of Computer Science
Secondary Technical advisor: Jim Cohoon, Department of Computer Science

Automatic Captioning Software from any Audio Source

Empowering Deaf and Hard-of-Hearing People in a Virtual World

Edward Clark
University of Virginia
ejc7re@virginia.edu

ABSTRACT

The deaf and hard of hearing (DHH) community has often struggled with finding accessible ways to use modern software. In an increasingly digital world, DHH individuals must adapt to interfaces that expect them to hear by default, and try to adapt such interfaces to their own needs. While there are accessible interfaces available, they either work poorly or are quickly outdated as technology continues to advance. Our reliance on technology has only increased due to the COVID-19 pandemic, and fully virtual workplaces and school environments are ubiquitous; there must be adequate tools available so that the social and intellectual gap between abled people and DHH people doesn't widen. I am proposing automatic captioning software to combat this. Using natural language processing, the application would act as an overlay that can caption any computer audio currently playing on the user's machine (Zoom Meetings, YouTube videos, etc.). This removes the need to download specific add-ons for each platform as the software will attempt to caption any audio that is currently playing. With this tool, DHH individuals would be able to seamlessly read what is being said on their computer without assistance or extra devices.

INTRODUCTION

The deaf and hard-of-hearing (DHH) community has always struggled with using devices and programs that heavily rely on audio or speech input. This is exacerbated as remote work and learning skyrocketed due to the COVID-19 pandemic, and popular tools such as Zoom and Google Meet do not have first-party support for automatic captioning [3]. Now, these tools are essential parts of everyone's lives.

The DHH community is a very large demographic in the United States, and accounts for 1 in 20 Americans [11]. Supporting this group with accessible tools is a paramount issue so that the social and intellectual gap between DHH users and abled individuals doesn't widen. Many online informational and entertainment videos, livestreams, and podcasts are all only sometimes captioned, but closed captioning is needed more than ever with American video consumption up 660% from 2011 to 2012 [10]. Requesting specific accommodations to view such content, like third-party ASL interpretation or captioning, can elicit a social stigma from coworkers and peers. ASL grammar can also be difficult with the pandemic, as face masks cover facial expressions, and video quality is often not good enough to clearly discern ASL signs [9].

While there is major work to be done, there is still progress being made in accessibility for the DHH community. The tech giants Google, Apple, and Facebook have made strides in deaf accommodations through better captioning and better speech recognition. Microsoft has also introduced chief accessibility officers and hired people with disabilities to "all kinds of roles" [2]. IBM's Watson AI and Google's Cloud Speech-to-Text API are both powerful tools that developers can leverage to build more robust captioning services that will only get better with time. However, we must ensure that companies keep up with accessibility. Hugely popular social media apps such as Snapchat, Instragram, and Twitter, as well as the quickly growing public audio chatroom app Clubhouse all have no support for automatic captioning and therefore have features that are unusable for many DHH users [8]. With these apps all implementing their own short-video-clip sharing tools pioneered by TikTok, the need for captioning on this content is essential. If large companies don't take advantage of the huge strides made in AI captioning tools, then the DHH community will become even more marginalized. This is why software that can handle captions on a computer would be incredibly helpful. DHH users would not have to rely on hoping that a particular platform or service has support for real-time captions, and they would no longer need to request special third party accommodations. A DHH person could fully engage in a virtual society with complete independence.

BACKGROUND

The primary technical aspect of my project uses Natural Language Processing (NLP), which is a branch of AI that studies how computers can make sense of human language, including semantics, tone, proper nouns, and context. NLP is used in tools such as Google Translate Microsoft Word, and Grammarly, but my project primarily deals with the understanding and translating spoken text.

I will not discuss the low-level implementation of NLP as I am proposing to use an API to handle the heavy-lifting of translating speech. There are currently powerful speech-to-text APIs available, such as Google's Speech-to-Text API, or IBM's Watson API.

RELATED WORK

Automated captioning tools and software exist on several platforms today. For example, many Google products include automatic real-time captioning, such as YouTube and Google Hangouts. However, the quality of YouTube's auto-generated captions is usually poor, so the video author must take the time to write their own captions on their videos. Google Hangout's live captioning system appears to be better, but these are only available if you are using the platform. The meeting application Zoom, popularized due to the COVID-19 pandemic, has recently introduced a live transcription tool; however, this is inadequate because they must be turned on by the meeting host, and there is no way to scroll back to see previous captions. If the meeting is on Discord, Skype, or any other service without full caption support, a deaf user will not be able to view captions without a third-party translator or ASL interpreter who is in the meeting as well.

A shoddy solution to needing captions on the computer is when DHH individuals hold their smartphone up to the computer's speaker and then read the words that appear on the screen using the phone's dictation ability [7]. However, this process is especially difficult if the user must multitask or appear engaged on a video call. Having the captions appear right on the screen would make the experience better for both parties.

There are also real-time closed captioning services that are similar to what I am proposing. One example is Adobe Closed Captioning offered through their Adobe Connect platform [1]. This service offers many of the same features that my software would have, such as customizable font, skip back to view old captions, and multi-language support. The problem is this powerful captioning service is only available on the Adobe Connect meeting software, and meeting organizers must hire the specific professional captioning service they want to use before the meeting begins.

Powerful automated captioning exists on some mobile phone apps too. One notable case is TikTok, which very recently implemented full captioning for users who turn it on in their settings in April 2021 [6]. This is an important step for such a popular app, and it should set an example for other large social media apps to do the same; however, support for this content is only on phones and not on computers. While the mobile app environment is moving forward with accessibility support, traditional computers are lagging behind.

Many non-captioning solutions for the DHH community, such as ASL interpreters or providing a transcription of content, work well to support the deaf, but have different issues. Requiring separate devices and requesting an ASL interpreter for a virtual meeting may elicit a social stigma from coworkers and peers. Transcriptions often leave DHH people feeling disconnected from the content, especially if they are provided after a live event. If there are no captions during an event and a transcription is provided afterward, the DHH community is left out of collective cultural moments.

A system where the DHH user is in control and can always get good, real-time captions regardless of platform or service would alleviate many of the inadequacies with current accessibility options.

SYSTEM DESIGN

The project itself is an outline of how the software would be implemented along with user interface mockups to show example use cases. I also did background research detailing why the tool is needed, which I will summarize here.

1 Use Cases

The initial primary use case for the software would be to help DHH people be better equipped to use only a computer to handle online meetings, lectures, or media. Virtual events such as online meetings are often a source of anxiety for the DHH community, as they often don't know if the platform they are using supports captioning or if they will need to plan ahead with the meeting organizer to provide accessibility options (ASL interpreters, 3rd party captioners).

Through working on the project, other use cases arose, such as helping people who are unfamiliar with a particular

language. Since the tool would have multi-language support, the software could be used to translate speech in real-time, which would give people much more autonomy in an environment where they don't know a particular language.

Ideally the software would be free to all users, but because the speech-to-text API costs money to use, there would have to be some way to pay this expense. To make the tool accessible to as many people as possible, everyone would be able to download and use the software for free, and then there would be a paid upgrade to get access to extra features. These extra features might be multi-language support, and the ability to view and save the entire transcript of a captioning session.

2 Background Research

I performed background research to gain insight on why a tool like this would be useful. This included learning about the deaf experience with regards to technology and how they have historically struggled with it. It appeared that there are three large issues when using technology that I will outline below.

The first issue is that current technology options are inadequate. I went into this in detail in the introduction and related work sections, but I will add that a study conducted on the quality of Deaf assistants labeled many of them as "wholly inaccessible" as actual DHH individuals did not have a say when they were being developed [12].

The second large issue was that there is a social stigma caused when using separate deaf accessibility tools. This is largely caused by the hearing world viewing being deaf as a disability, rather than an identity that beings personal growth and community. Separate special tools that DHH individuals must carry around may highlight their "otherness" amongst peers.

Lastly, the rate at which deaf accessibility tools are iterated upon is much slower than the rate that standard technology progresses. Captioned Telephone Services with an operator on the call typing captions are still commonplace, as are FM radio systems where a speaker wears a special microphone and the listener wears a receiver linked to a hearing aid.

Each of these issues contribute to the struggle that DHH individuals face when using technology. The captioning software I am proposing aims to alleviate them.

3 User Interface

The goal for the user interface of the system was for it to be familiar, unintrusive, and easy-to-use. The software is simply an overlay that appears on top of whatever program is on the screen. Figure 1 displays a mockup of how this might look in practice.

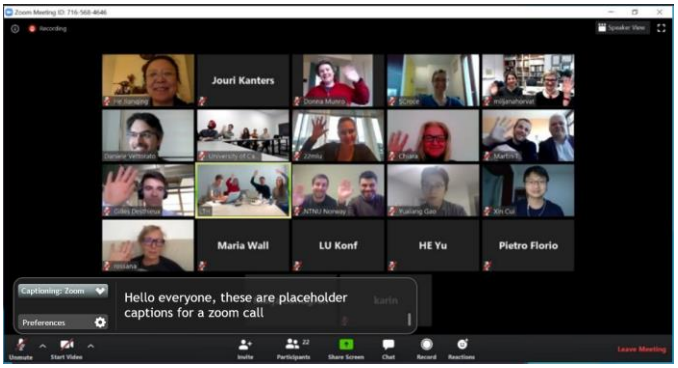

Figure 1: UI Mockup of software overlay shown on bottom left of a Zoom call.

Users can click and drag the overlay to position it around their screen so that it won't cover any important information. Using the dropdown menu on the left side of the overlay, users can select which audio source they would want to generate captions from.

The captions themselves take up the majority of the overlay, which is intentional as captions are the primary focus of the software. Figure 1 displays a scroll bar on the far right so that users can scroll back to see previous captions. There is a timestamp marking every minute so users know when something was said when scrolling back.

The preferences button on the bottom left of the overlay opens up a menu so that users can personalize the software to their liking. This includes choosing the font size, style, and color for the captions to achieve the best readability against the different backgrounds on different platforms and services. They can also change features such as the opacity of the overlay, how much text should be displayed at once, or small visual things such as choosing if the text scroll smoothly or jump straight to the next caption. The last feature of preferences is allowing users to choose which language they want the captions to be displayed in. The Speech-to-Text API can translate text from many languages in real-time, which makes the software much more accessible for those who may have trouble understanding English.

4 System Architecture

The desktop software client would be built with the Electron framework, which is an open-source framework built by GitHub. Electron is used to build desktop GUI applications with web technologies, and applications such as Discord, Visual Studio, and Slack code use it. It is also cross-platform, which is very important in order to support users with either a Mac or Windows machine.

Electron uses several processes to compose applications. The main process would run the application logic, which in our case is taking the speech input, generating captions, and outputting the captions as text. Separate rendering processes are then responsible for rendering what users see on their screen.

4.1 Front End

Electron lets desktop applications use HTML and CSS, which is what the front end of the tool would be built in. There would also need to be JavaScript to handle the buttons as well as grabbing the data from the back-end to actually display the captions.

4.2 Back End

The criteria for choosing a speech-to-text API was that it must be able to transcribe content in real-time, must be well supported by developers, and must have support for multiple languages. For these reasons, I am choosing the Google Cloud Speech-to-Text API. This API uses Google's deep learning neural network and has support for 125 languages. There is also very detailed documentation with tutorials on how to go about streaming speech input, which would help with implementation.

Both the Electron framework and the Speech-to-Text API have extended support with Node.js, which will be responsible for the back end translation.

The Sound Exchange (SoX) command line utility must also be used for implementation [13]. SoX can convert many different formats of computer audio files into other formats. This is essential because we will likely have differing formats from each audio source, and the Google Cloud Speech-to-Text API requires a WAV audio file. The with SoX utility, we must then install the Record dependency for Node.js so that we can record the audio from the microphone. With each of these building blocks, we would be able to implement functionality to translate captions in real-time.

5 Risk Analysis

There are several possible problems that may arise when using this kind of software. My initial concern was that the API would not be able to translate speech fast enough; however, this would not be an issue as the Google Cloud Speech-to-Text API can actually translate faster than real-time.

An actual issue that may arise is if the user chooses to caption audio that is not the spoken voice, such as non-lyrical music. What if the DHH user did not know that the audio they wanted to caption wasn't spoken language? In cases such as these, the overlay would display a simple message such as "[selected audio not able to be captioned]," which would indicate that some non-language audio is being played.

Another potential problem, which is common with many captioning tools, is when multiple people are speaking at the same time. Even if the speakers are speaking at separate times, a DHH individual wouldn't know there are two different people without visual cues. One powerful feature of the Google Cloud Speech-to-Text API is speaker diarization, which can detect when speakers change and labels each individual voice detected with a number. The overlay would display this number next to each spoken of line of text, so a user knows exactly how many people are speaking and what they are saying.

Other potential issues are if the speaker is speaking with a lot of background noise, such as if they are outside or if there is the chatter of an audience in the background. There is no way to avoid this issue, and any live captioning service would struggle with this problem. This problem will slowly improve over time as a result of smarter captioning APIs.

RESULTS

Results for the software proposal are purely anecdotal, based on how people perceive the usefulness of the software. I only know very few DHH people personally, so I posted a summary of the tool and its features on an online deaf community forum to gauge interest. There were few responses, but they were generally positive. The minimalism and ease-of-use of the tool seemed to be a strong positive point about the proposal. Nearly every respondent mentioned the struggle with making sure a particular service had support for real-time captioning. There was also a mention of the poor quality of some services that did offer live captions, such as YouTube.

I also discussed the tool with some of my friends, all of whom are hearing individuals. Many of them like to turn on captions whenever they are watching something simply so that they can understand exactly what is being said better. All of them thought a tool like this would be incredibly useful, and some surprised that a simple overlay captioning tool with powerful captions had not yet been built. I personally am hard-of-hearing and prefer to turn on captions if possible, but can still manage perfectly fine without them. The very simple and minimal nature of the overlay software itself is something that I would love to have – without all of the bloat that a lot of software seems to have these days.

CONCLUSION

This proposal of a simple software tool that can automatically caption incoming audio from any source currently playing on a user's computer was designed to help the DHH community better navigate fully virtual and school settings. As our lives become more and more reliant on technology, especially due to events such as the COVID-19 pandemic, supporting everyone with proper accessibility options is essential for their full inclusion in society.

My primary argument is that DHH individuals shouldn't need to buy special third party devices in order to use something like a computer, or need to request special accommodations from a virtual meeting organizer. Computers and virtual meeting software should be equipped with the necessary accommodations by default so that any person with disabilities can operate it autonomously, without third parties.

The results show that this software would be very valuable to a broader population than just the deaf and hard-of-hearing community. The simplicity of the software and the concept of "one tool for all," removing the need to rely on particular services with caption support, makes for a beneficial product. Giving people independence when using technology is necessary to close the social and intellectual gap between people with disabilities and their abled counterparts.

FUTURE WORK

Future work on this project of course includes actually building the software. I believe that the outline detailed here as well as the power of Google's Speech-to-Text API would be enough to implement the tool in a reasonable amount of time as an individual.

As speech-to-text and natural language processing become more versatile through artificial intelligence, smarter captioning may be possible. For example, providing captions through background noise or be able to decipher the speech of two people talking over each other and displaying the captions for both speakers would be a massive step in improving real-time captions. Lastly, if the software was successful, we could let meeting organizers use it at an enterprise level for any virtual events. This would be good for DHH individuals who do not know that the software exists as they could still get access to captions without it.

REFERENCES

[1] Adobe Connect Closed Captioning Homepage. https://www.adobe.com/products/adobeconnect/apps/closed-captioning.html

[2] Abrar Al-Heeti. 2019. Accessibility Tech has a lot of Unfinished Business to get Right. CNET (May, 2019). https://www.cnet.com/news/for-people-with-disabilities-accessibility-techs-still-not-all-it-could-be/

[3] Anderson et al. 2020. Deaf ACCESS: Adapting Consent Through Community Engagement and State-of-the-Art Simulation. Oxford University Press25-1 (Jan, 2020), 115-125. Web of Science.

[4] Dr. Michael J Garbade. A Simple Introduction to Natural Language Processing. Medium. (October, 2018). https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

[5] Google Cloud Speech-to-Text API Documentation. https://cloud.google.com/speech-to-text

[6] Stephanie Hind, 2021. Introducing auto captions. https://newsroom.tiktok.com/en-us/introducing-auto-captions

[7] Mikaela Lefrak, 2020. From Zoom Meetings to Face Masks, Deaf People Face Extra Struggles During Pandemic. NPR(April). https://www.npr.org/local/305/2020/04/30/848432236/from-zoom-meetings-to-face-masks-deaf-people-face-extra-struggles-during-pandemic

[8] Rachel Lerman, 2021. Social Media has Upped its Accessibility Game; but Deaf Creators say it has a Long Way to Go. The Washington Post (March, 2021). https://www.washingtonpost.com/technology/2021/03/15/social-media-accessibility-captions/

[9] Sarah Katz, 2020. The deaf community is facing new barriers as we navigate inaccessible face masks and struggle to follow news broadcasts and teleconferences – but the tools for accessibility are out there. Business Insider (April, 2020). https://www.businessinsider.com/deaf-people-face-new-barriers-amidst-pandemic-accessible-tools-2020-4

[10] Michella Maiorana-Basas and Claudia M. Pagliaro. 2014. Technology Use Among Adults Who Are Deaf and Hard of Hearing: A National Survey. The Journal of Deaf Studies and Deaf Education. 19-3 (July, 2014), 400-410. JSTOR.

[11] Ross E. Mitchell,.(2005. How Many Deaf People Are There in the United States? Estimates From the Survey of Income and Program Participation. The Journal of Deaf Studies and Deaf Education 11-1 (September), 112-119. Oxford Academic.

[12] Ryan Lee Romero et al, 2019. Quality of Deaf and Hard-of-Hearing Mobile Apps: Evaluation using the Mobile App Rating Scale (MARS) With Addition Criteria from a Content Expert. JMIR mHealth and uHealth (Oct.). US National Library of Medicine (PMC).

[13] Sound eXchange Homepage. (February, 2015). http://sox.sourceforge.net/