Real Time Transcription and Subtitles

Eli Kersey
University of Michigan
Grand Rapids, MI
ekersey@umich.edu

Joseph Adelman
University of Michigan
Birmingham, MI
jadels@umich.edu

Ari Singer
University of Michigan
West Bloomfield, MI
arising@umich.edu

Tony Tang
University of Michigan
Ann Arbor, MI
tonytang@umich.edu

ABSTRACT

In the post-Covid era, virtual meetings have become a mainstay of professional and personal communication. However, the integration of accessibility features, especially for the deaf and hard of hearing (DHH), remains inadequate. Traditional subtitling systems often do not indicate who is speaking, which can significantly hinder comprehension and participation in conversations. To bridge this accessibility gap, our team has developed an application that dynamically associates transcribed speech with its respective speaker.

The application uses live camera feed and microphone input, and leverages state-of-the-art computer vision algorithms to detect and track speakers' faces and lip movements. Concurrently, OpenAI's Whisper API transcribes the spoken words accurately. The resulting captions are then displayed in real-time adjacent to the speaker, ensuring clarity and enhancing communication.

Preliminary user testing has demonstrated the system's potential to significantly improve accessibility, with users reporting greater satisfaction and engagement compared to traditional captioning methods. These results underscore our technology's potential as a transformative tool for inclusivity in digital communication environments. By directly addressing the crucial need for speaker-specific captioning, our solution sets a new standard for accessible virtual interactions.

INTRODUCTION

The digital age has transformed the way we communicate, collaborate, and convene. Virtual meetings, a niche convenience pre-pandemic, have ascended to become a primary medium for professional and social interactions since the start of the Covid-19. This shift, while revolutionary, has spotlighted significant gaps in accessibility, particularly for individuals who are DHH. Traditional virtual meeting platforms often feature subtitling systems that are static and non-interactive, typically displaying captions in a continuous scroll at the bottom of the screen without indicating which participant is speaking. This method can be disorienting and inadequate for following conversations meaningfully, as it strips away the nuances of dialogue and speaker dynamics.

Recognizing this deficiency, our team set out to devise a system that not only enhances the accessibility of virtual meetings but also enriches the communication experience for all users. Our development—Real Time Transcription and Subtitles—integrates advanced computer vision and speech-to-text technologies to create a dynamic, speaker-associated captioning system. By utilizing OpenCV for facial recognition and tracking lip movements, coupled with OpenAI's Whisper API for real-time speech transcription, our system uniquely overlays captions directly adjacent to the respective speakers in the video feed. This approach ensures that each participant's spoken words are immediately identifiable, fostering a more inclusive and engaging interaction for users with hearing impairments.

The importance of this innovation extends beyond mere convenience. For the DHH, the ability to easily identify who is speaking and what is being said in real-time is not just an enhancement, it's a fundamental requirement for equitable participation. Our team recognizes the importance of being heard in a conversation, and we believe that eye contact and focus are two main factors in this. Our technology aims to set a new standard for accessibility in digital communications, addressing a critical need that has been largely overlooked by existing platforms.

Initially, our team envisioned implementing this technology using Microsoft Hololens, which would have allowed users to see real-time captions directly in their field of vision through augmented reality. This approach promised an even more immersive and intuitive user experience, integrating digital text into the physical world in a seamless manner. However, technical and developmental challenges prevented the project from being successfully implemented on the Hololens platform within our project timeline.

Faced with these hurdles, our team made a strategic decision to pivot the application towards a more feasible platform. This shift allowed us to focus on perfecting the core functionalities of facial recognition, lip movement tracking, and live captioning on standard computing devices. While this change required us to adjust our initial goals, it ultimately enabled us to create a robust and scalable solution that could be readily integrated into existing virtual meeting platforms. This adaptation was necessary to ensure we could pursue our foundational vision of enhancing accessibility and interaction in digital communications with the limited time we were allotted.

In the following sections, we will delve deeper into the technical architecture of our system, the outcomes of our studies, and the results of our work in the field of human-computer interaction (HCI) and accessibility. By bridging the gap between technological capability and

user-centric design, Real Time Transcription and Subtitles enhances the functionality of virtual meetings.

RELATED WORK

There are many solutions for generating closed captions on the market, however none of them address the issue of identifying speakers. The best example of this is YouTube's automated captioning algorithm. While YouTube's powerful deep neural networks do well generating traditional captions, it simply transcribes speech without associating a speaker to their words.

Google Meet allows for closed captioning on calls, and their subtitles are differentiated by each person. This works by listing a person's name in front of their speech in real time. This works as each participant has their own microphone input, so Google can easily differentiate and isolate each speaker's audio.

SYSTEM DESIGN AND APPROACH

The goal of our project was to develop a real-time captioning system that not only transcribes spoken words but also associates these captions with the speakers in a virtual environment. This section outlines the system architecture, the technologies employed, and the methodologies used to achieve this goal.

Architecture

The system is designed to handle three main functionalities: facial detection and tracking, speech recognition, and caption rendering. The architecture is divided into two major components: facial detection and caption rendering, and then the transcription server. The client-side application captures video and audio streams from the user's device and sends the audio data to openAI to be transcribed.

Facial Detection and Tracking

At the heart of our system lies the integration of OpenCV, a powerful library used for computer vision tasks. OpenCV processes the incoming video feed to detect faces in each frame and tracks these faces across the video. To facilitate this, we implemented a Face class, which encapsulates each detected face. This class stores key metrics such as the positions of facial landmarks, with a particular focus on the lips, and any spoken words.

For each video frame, the system measures the distance between key points around the lips within each Face object, calculating the amount of movement over time to determine the likelihood that the person is speaking. In addition to detecting speech, our system must continuously track identified faces across frames, which is critical for maintaining accurate captioning when individuals move or when new faces appear.

To manage the tracking of faces from frame to frame, we employ a nearest neighbor approach. Each new face detected in a frame is compared against existing Face objects, and is associated with the previous face that is closest in terms of spatial positioning. This method ensures that the identity of each speaker is maintained consistently throughout the video feed, allowing the captions to be accurately associated with the correct speaker as they move or as the camera angle changes. This dynamic tracking is key to our system's ability to deliver real-time, speaker-associated subtitles effectively.

Speech Recognition

Our system processes the audio stream captured from the computer's microphone in its entirety, rather than isolating individual speakers. This continuous audio feed is sent to the Whisper API, which provides robust speech-to-text capabilities. The audio is continuously processed by Whisper, allowing the transcription to be completed in real time, rather than waiting for an entire phrase to be spoken.

The continuous stream of audio is processed in chunks by the Whisper API as it is received. Our system employs an algorithm that tracks the spoken phrases and attributes them to the correct speaker based on the visual data provided by the facial detection component. This algorithm continuously updates the text displayed based on the detection of who is currently speaking. If a new speaker begins talking, the system adjusts to display only the newly spoken words above the correct speaker's head. This is done by having each face track which words they have spoken, and then displaying each face's words on the screen independently. Our initial approach was to edit the phrase, but as this changes over time due to the nature of the Whisper API, that was not a realistic approach.

However, since all audio is processed together, our system faces limitations when multiple people speak simultaneously. In such cases, the ability to accurately transcribe the audio and assign text to the right speaker is compromised, reflecting a challenge in handling overlaps and interruptions. The system resets the transcription phrase after a brief pause—approximately half a second—although this mechanism is not immune to errors from sudden noises or rapid conversational exchanges. These complexities highlight the challenges of real-time transcription in dynamic and unpredictable environments.

Caption Rendering

Once transcribed, the text is sent back to the client application, where it is rendered onto the video feed using OpenCV. This integration of OpenCV allows for precise placement of captions in real time, ensuring the transcription appears above the user's head. The system calculates the spatial coordinates of each speaker detected in the video frame and dynamically adjusts the placement of the text as the speaker moves or as the camera angle changes. Since people tend to speak in large bouts, we limit the number of characters on the screen at once to forty, and as more words are spoken the words carousel from right to left in a readable manner.

OpenCV's powerful image processing capabilities are leveraged to overlay the text directly onto the video feed.

This ensures that the captions are not only accurately positioned relative to the speaker but also move seamlessly with them, maintaining clarity and coherence in the visual presentation. One issue we faced with this is that if a user is close to another speaker, their words may overlap, causing difficulties in reading either transcription. This method of on-screen text rendering helps in keeping the audience engaged by providing a clear visual association between the

- OpenCV: Utilized for its advanced video processing capabilities, especially in detecting and tracking facial features in real-time.
- OpenAI Whisper API: Selected for its state-of-the-art performance in speech-to-text transcription, crucial for ensuring the accuracy of captions.

spoken words and the speaker, enhancing both accessibility and the overall viewing experience.

METHODOLOGY

The development process followed an iterative approach, beginning with a prototype that focused solely on the detection of speech from a static speaker. We then attempted to connect this feature to the Microsoft Hololens two augmented reality glasses. We ran into development issues with this, and eventually switched to a pure software application that reads in camera and microphone data. We then built out the facial detection system, and combined this with our transcription model for a finished product. Each iteration involved:

Internal Testing: Continuous user testing was conducted to refine the approach. This included assessing the system's performance in different lighting conditions, with various background noises, and with speakers of different speech patterns and voices.

Performance Optimization: Key performance metrics such as latency, accuracy of facial tracking, and the synchronicity of captions were monitored and optimized. Challenges such as reducing the lag between speech and caption display and improving the accuracy of speaker identification in crowded scenes were addressed during this phase. One main issue we still are facing is if one person is speaking while another is continuously moving their lips. Since our application decides who is speaking based on lip movement, this can cause the transcription to be placed above the wrong person.

The approach was pragmatic, focusing on creating a functional system within the project's constraints and timelines. While the pivot from an augmented reality platform like Microsoft Hololens to a more traditional video conferencing setup changed the scope of deployment, it allowed us to concentrate on refining the core features of our application and ensuring their effective integration to work in unison in real time.

USER STUDY

The user study was structured into three primary experiments:

Single-Speaker Test: In this setup, one participant was recorded speaking directly to the camera for approximately 30 seconds. The participant was told they could speak about anything they wanted to, and to pretend as if they were on a video call.

Dual-Speaker Test: For this experiment, two participants were recorded while engaging in a dialogue for 30 seconds. Participants were instructed to remain facing the camera to simulate a video call environment, and to take turns speaking to mimic a natural conversation flow.

In both experiments, the recorded videos were processed through our captioning system to generate real-time captions associated with the speaker(s). These videos were then presented without sound to a separate group of observers. The observers were not aware of the original content of the dialogue but were tasked with understanding the conversation solely through the captions generated by our system.

Data Collection

Observers were asked to watch each video and then immediately complete a Google Form, which collected both quantitative and qualitative data. The survey included the following components:

Comprehension Questions Listed Below: Multiple-choice and short-answer questions that tested the observers' understanding of the content discussed in the video. These questions are based on the enjoyability, readability, and real world use of this transcription model.

Observers rated how much they enjoyed using this caption system on a scale from 1 (very poor) to 7 (excellent).

Observers rated the readability of the captioning system on a scale from 1 (very poor) to 7 (excellent), based on how easy it was to follow the conversation through the captions.

Observers rated their understanding of the conversation they watched on a scale from 1 (very poor) to 7 (excellent).

Observers rated if they would be able to use this type of captioning system in a real time conversation on a scale from 1 (strongly disagree) to 7 (strongly agree).

Observers compared this use of captioning to standard subtitles on a scale from 1 (much better) to 7 (much worse).

Would you like to see these subtitles in the real world, such as movies (Why or why not)?

If this was an option on a zoom call, would you use it? (Why or why not)?

Feedback Section Below: An open-ended section where observers could provide feedback on their experience,

including any difficulties they encountered or suggestions for improvement.

Are there any glaring features that are missing? What would you change about this program?

Based on these user videos, we found the Transcription Accuracy Score (TAS) and the Placement Accuracy Score (PAS). The TAS tests what percentage of words are correctly transcribed, and the PAS is to find what percentage of words are placed correctly above the speaker.

Users	TAS	PAS
Single	**99%**	**100%**
Double	**95%**	**88%**

Table 1. TAS and PAS scores of user testing, with either a single user of two users having a conversation

Results Analysis

The responses were analyzed to assess the overall effectiveness of the captioning system in providing a clear and understandable transcription of spoken dialogue.

Accuracy of Comprehension: The percentage of participants that felt as if they understood the conversation.

Usability Scores: The average usability scores provided by observers, reflecting the ease of use and effectiveness of the captioning in real-time communication.

Qualitative Feedback: Common themes and observations from the feedback section that highlighted areas of strength and opportunities for further enhancement of the system.

Key Findings

Results from the user study indicated that the system was generally effective in allowing observers to understand the content of the conversations, even in the absence of audible speech. 84% of participants had an enjoyable experience using these subtitles, felt as if they had at least a ‘good’ understanding of the conversations, and felt as if they could use this model in real time. Compared to standard subtitles, 53% of participants felt as if this model was better, with 31% feeling as if they were worse, and 16% of users feeling no difference.

However, challenges were noted in scenarios with overlapping dialogue in the dual-speaker test. This would occur due to how close the two speakers were. The main feedback on how to fix this issue was to place a background around the text box which acts as a bounding box to other text boxes. The tradeoff with this is that the text will then impede on more the listeners field of view. There were also occasional moments where the system struggled to accurately associate captions with the correct speaker, normally putting the last two to three words of one speaker on the other user as the beginning of their speech. Feedback from observers suggested that while the caption placement and timing were generally well-received, improvements could be made in handling rapid speaker switches and in reducing caption lag.

The outcomes of testing our applications accuracy are presented in Table 1, which summarizes the Transcription Accuracy Score (TAS) and the Placement Accuracy Score (PAS) for scenarios with a single user and with two users engaged in conversation.

For the TAS, our system displayed exceptional performance in single-user conditions, achieving a 99% accuracy rate. This near-perfect transcription accuracy signifies the system's capability in processing and converting spoken language into text with minimal error, a crucial feature for clear and reliable captioning.

In the more challenging double-user scenario, the system sustained a high level of transcription accuracy at 95%. Despite the overlapping speech and potential background noise, the Whisper speech to text algorithms have managed to maintain a high degree of precision.

The PAS was equally impressive in single-user situations, with an impeccable success rate of 100%. This was the expected result since there is only a single speaker on the screen, with the only possible errors being attributed to people in the background.

In the double-user testing environment, the system achieved an 88% PAS. While this is a slight reduction from the single-user case, which is expected, it represents a strong capability in differentiating and correctly placing captions between two interacting speakers. The main reason as for why this did not work with perfect precision due to participants speaking quickly after one another. This causes one speaker's last few spoken words to be placed above the new speaker as the start of their sentence.

These findings confirm the efficiency and effectiveness of our real-time captioning system in various user scenarios. The high TAS across both single and double-user tests underscores the reliability of the transcription service, which is fundamental for user comprehension and engagement. Meanwhile, the PAS results affirm the system's ability in caption placement for single-user interactions and point to areas for enhancement in multi-user exchanges.

CONCLUSION

This research embarked on addressing a significant gap in the accessibility of virtual communication platforms by developing a real-time, speaker-associated captioning system. Through integrating computer vision and speech-to-text technologies, we have created a tool that not only improves accessibility for DHH users but also enhances the overall user experience in virtual environments.

Our system, utilizing OpenCV for facial detection and tracking combined with OpenAI's Whisper API for transcription, represents a significant advancement in

real-time captioning technology. By overlaying captions directly above the speakers, our system ensures clarity and contextual relevance, helping users better follow and engage in conversations. The successful implementation of this system on standard computing devices, after pivoting from the initial augmented reality concept, demonstrates our team's adaptability and commitment to accessibility.

The user studies conducted as part of our research provided insightful feedback into the system's effectiveness and areas for improvement. The studies showed that our system enhances comprehension and user satisfaction compared to traditional captioning methods. Most participants reported a high level of engagement and expressed a preference for our dynamic captioning approach over static subtitles. However, challenges remain, particularly in scenarios with overlapping dialogues, and our system's ability to accurately associate captions with the correct speaker. These findings not only validate the utility of our system but also highlight the intricacies of real-time audio processing in complex environments.

Looking forward, the potential applications of this technology extend beyond virtual meetings to other areas such as education, live broadcasting, and public speaking, where inclusivity can be improved through better accessibility tools. Additionally, the integration of more advanced machine learning models and the exploration of spatial audio processing could further enhance the accuracy and reliability of speaker identification, especially in multi-speaker settings.

In conclusion, our project has laid a strong foundation for the future development of accessible communication tools. It also sets a new standard for how technology can be leveraged to make digital spaces more inclusive. With continued development and refinement, technologies like ours could soon become a staple feature in virtual communication platforms, ensuring that no participant is left unheard.