# Project 1 Report
# Multi-speaker Identification

Hubert Lin
EC601 Section A1

## I. Introduction

In today's society, communication is critical to human interaction in professional, educational, and social environments. New technologies emerge as powerful tools used to help facilitate and aid in effective conversation; they connect people from around the world and address a growing need for wireless and online communication methods. In many cases, software must be able to identify who is speaking in a conversation. This creates a need for multi-speaker identification for situations where communication requires distinction between two or more participants. Situations like these can be found in the workplace, where team meetings or client interactions require software that is able to track speakers in real time. Other examples include class lectures, where students and teachers must be able to participate simultaneously in discussions, or voice calls between family and friends who are not able to meet in person. This report aims to explore the need for multi-speaker identification and analyze various approaches that have emerged to satisfy user needs. Section II explores the purpose and use case for better products, and section III compares existing approaches and tools that have been explored in research and industry. Finally, section IV explores one implementation in an attempt to better understand the challenges and approaches in creating speaker identification solutions.

## II. Purpose

Behind every approach and product lies an intent: a goal or vision to address a problem or need driven by user demands. In the case of multi-speaker identification, one must consider how users are using technology to communicate, what features they desire, and then develop a vision of how to improve the user experience. The following subsections explore how Zoom and Microsoft Teams, two popular video and voice conferencing platforms, have considered some of the most prominent user stories and crafted a product centered around improving user experiences.

### *A. Business*

Analyzing what users are using software for and their reasons for using that software are valuable in determining how to approach an issue. One target audience to consider is working professionals. Many businesses are shifting towards remote work, but wish to maintain collaborative efforts on team projects and leverage tools that can enable effective communication [1]. Users in these situations want a tool that can allow for video and audio calls and conferencing; software that is intuitive and can handle a large number of participants. Cloud-based communication platforms such as Zoom and Microsoft Teams have considered this particular use case. Both have easy to use visual

interfaces and have the ability to automatically detect a user's microphone and speakers. They enhance the user experience by giving users the ability to see multiple participants on one screen while still maintaining clarity by highlighting the person(s) speaking with special colored borders. Each participant's name is also included, minimizing confusion. In a rapid work environment, users want software that is fast and convenient. Both Zoom and Teams allow users to add participants in a matter of seconds regardless of location, addressing the need for remote communication while providing options for video and group inclusion that traditional means like telephone calls cannot offer. The result is a product that businesses like HubSpot, a software marketing company, will want to use and rely on for daily operations [2]. A company survey revealed that Zoom was the top ranked collaboration tool used by employees, for cases including: meetings not in the office, communicating across time zones, and showcasing products to customers.

### *B. Education*

Another field that demonstrates a strong need for multi-speaker communication is education. With a strong emphasis put on remote learning due to external factors, students and educators seek outlets for effective teaching and accessibility. In a classroom environment, instructors want to display course material such that it is visible to all students. Students want a learning experience that emulates that of a physical classroom, where they can ask questions, work in groups, and interact with their instructor and classmates. Students that are unable to attend class want the ability to access and engage in the material and discussions that they miss. The Zoom and Teams platforms have several features that are targeted towards the functionality that students and teachers seek. Both allow users to record calls for later reference, which enables instructors to save lecture recordings for students to access. Both also feature screen sharing, which allows instructors to share lecture slides and students to present during class discussions. Finally, a raise hand feature allows students to alert the instructor when they have a question, simulating the feel of a physical classroom without disrupting the flow of the video stream. The California State Park system has adopted such practices, utilizing the functionality of platforms like Zoom to give K-12 students the opportunity to learn about nature and wildlife from park staff even if they are unable to physically visit the parks [3]. Universities across the world have also utilized this technology to facilitate global learning and promote social distancing with hands off education; some notable examples include business programs at Columbia University and flexible degree programs at the University of Arizona [4].

## III. Analysis of Existing Contributions

In addition to the platforms mentioned in the previous section, there have been efforts in academia and industry exploring different advanced approaches to speaker identification. This section will compare and analyze the benefits and limitations of three approaches to different use cases and offer potential suggestions for future contributions.

### *A. Case 1: Speaker Identification During Live Conversation*

What happens when there are multiple speakers in the same meeting room? With many individuals sharing the same microphone input, how can one distinguish between who is speaking? Furthermore, how can one confirm that the person currently speaking is indeed the person being displayed and not an imposter or unknown? These are realistic scenarios that might occur in the workplace or classroom. Researchers from the International Computer Science Institute and the University of San Diego developed an application that attempted to tackle these challenges [5]. The application uses a training mode that records 60 seconds of speech from each new user. This is compiled into a database containing the data of each user. Then, a recognition mode is used for speaker identification during conversations. This process consists of two steps: speaker diarization and speaker recognition. Speaker diarization aims to split a stream of audio input into speech clusters. Using only the internal microphone on a laptop placed in a meeting room, recorded audio is divided into distinct regions containing speech from different conversation participants. This is done through feature comparison using Mel Frequency Cepstral Coefficients (MFCC), where audio is converted to MFCC features that are compared with Gaussian Mixture Models (GMM) for speech and non-speech created during the training mode. For data classified as speech, features are then compared against speaker models for each speaker registered in the database to determine the best match. As an additional check, the likelihoods for the two most similar models are compared to see if the speaker is really a speaker in the database or if the speaker is potentially unknown and not in the database. The final application presents the results to users via a Java interface showing a picture and name of the current speaker. The application is also accessible online; users not in the room are able to monitor the conversation and keep track of who is speaking when.

Benefits to this approach include an easy to use interface and low latency. In the study, the application was tested by users who did not have familiarity or prior interactions with the system or its development. The test group was able to use the application successfully with only basic instructions and prompts from the interface. Users benefit from products with simple yet effective interfaces that do not require specialization to navigate. The testing group was also able to run the application on a basic dual core MacBook laptop and output results to the interface with delays of only 1.5 seconds [5]. While this could be optimized in the future, the current iteration is a good proof of concept and would benefit users by providing the desired functionality of speaker identification without having large delays disrupt user experience. A potential area for improvement would be flexible noise sensitivity. Different meeting environments will have different ambient noise levels. Natural disruptions such as laughter,

coughing, and typing may interfere with normal operation. In addition, users may want additional features such as compatibility with mobile devices and the ability to adjust input and output settings. Future work could be done on the algorithm optimization side to reduce latency and improve accuracy or on the user interface to offer more options for user customization.

### *B. Case 2: Blacklisting Speakers*

With an increase in available communication methods and the number of conversations initiated every day at home and at work, how can one maximize safe and meaningful interactions while avoiding potential threats? In home or work environments, one may receive calls from unknown callers. These can be spam calls, automated advertisements, or messages requesting personal information; all are generally undesired by the typical recipient, but in some cases the caller information is unavailable or does not reveal enough information to raise alarm. In cases like these, it would be useful to have software capable of filtering out all unwanted correspondence. Researchers at the MIT Computer Science and Artificial Intelligence Laboratory examined the specific case of customer-agent interactions in call centers [6]. In this scenario, it would be beneficial to maintain a blacklist of known scammers and fraudsters. By treating this list of speakers as a target group, speaker identification methods can be used to determine whether unknown callers belong to the blacklist. Then, some type of alarm can be triggered that notifies the call recipient. The researchers define an approach consisting of two main tasks: open-set detection and closed-set identification. Open-set detection attempts to determine whether or not a speaker's input falls into a known target set, while closed-set identification assumes an input belongs to a known set and attempts to identify the specific set. The researchers issued a public challenge to develop a system that followed the proposed approach. Using a provided dataset containing training, development, and test data for blacklist and non-blacklist speakers, challenge participants would try to develop systems that would accurately determine both the number of blacklist speakers in the test set and whether each individual input recording came from a member of a blacklist or non-blacklist group. System performance is evaluated by comparing each submission against a baseline system utilizing multi-target score normalization (M-Norm) to rank decision scores. The top performing submission is found to be a system fusion of Probabilistic Linear Discriminant Analysis (PLDA) and Neural-Network (NN) models, which showed performance improvements of over 30% compared to the baseline. Most submissions used completely different training models and approaches, with only 40% of all submissions performing better than the baseline system. The challenge results suggest that while machine learning and artificial intelligence have the potential to be leveraged for multi-target speaker identification, at present the best methods and approaches for doing so are still very much unknown.

Trends observed during the study hint at possible benefits and drawbacks to current speech technology models. One advantage of these systems is that they can be trained to become more accurate over time as the system sees more real data. Another advantage lies in versatility,

where models can be fed training data specific to the desired user cases and scenarios. Such systems will be able to take advantage of patterns and trends in human speech. In work environments that have consistent daily routines, these models might be extremely effective. However, there are limitations that still need to be addressed before these types of systems can be refined into consumer products. In the challenge, the researchers noticed that dialects caused issues with the performance of many systems and led to mismatches of data to correct target groups. In real world situations, one does not have control over many user characteristics. Thus, in order to be accessible to a wide variety of users, a successful product must maintain consistent performance independent of the user's speech qualities. Future work can be done to further improve identification models and algorithms, and more testing must be done with consideration towards realistic speech attributes and user groups.

### *C. Case 3: Smart Home Control and Automation*

With a wide variety of smart devices available to consumers, can users utilize speech technologies to control and automate daily tasks? Smart devices can be found in many households and can be useful tools for monitoring or providing feedback on other home systems. A study on the long term behavior of smart speaker owners revealed that they commonly use their devices for streaming music, controlling other devices, and obtaining information about the weather forecast [7]. In addition, some of the most popular commands among users all inspire action: opening and closing doors, turning lights on and off, and setting alarms and timers. Users want the system to respond with an appropriate action based on the voice command issued. Multi-speaker identification can be useful in these systems by distinguishing between different users in a household. Researchers from Singapore created a smart home prototype that utilizes speaker recognition technologies to provide three different types of voice services that anticipate and respond to user needs [8]. The system consists of several blocks that are integrated to form a functional product. The user interacts with the system via a web service interface that can be accessed via a smartphone. The user's voice inputs are recorded through the phone's microphone and passed to the central processing block, where speech is decoded to determine the command issued and feature extraction is performed. Finally, the processed outputs are passed to the speaker recognition block, where they are matched to a specific user in a list of registered users. This gives the system the ability to verify whether certain commands should be executed based on who is speaking; this has applications in improving household security and allowing for user specific commands.

There are several benefits to including multi-speaker identification capabilities in smart home devices. One that was mentioned briefly in the previous case is security. Since home devices can be set to control other systems critical to home security, it is important to make sure that any individual issuing these types of commands is recognized by the system as a trusted person. Users would not want strangers or visitors to have control over certain actions. Another benefit lies in personalized profiles for multiple users using the same device. A command could

trigger different actions depending on preferences set by each user, such as their favorite genre of music or their daily wake up alarm time. One potential weakness lies in the integrated design; if one block fails, then the rest of the system does not work properly. This makes it easy to tell if the system is not functioning, but makes it harder to pinpoint what part or parts of the system are responsible for failure. As with many of the cases mentioned above, there is still room for improvement in recognition approaches and algorithms moving forward. In addition the design could be optimized further to minimize loopholes in security and add features to the user interface.

## IV. Exploring an Implementation

In an effort to learn more about some of the challenges surrounding speaker identification, I wanted to attempt some basic audio processing and analysis to see how different characteristics of speech vary between speakers and affect results. After considering various possible approaches, I tried to implement some basic speech classification in MATLAB using spectrogram analysis. I chose this example because I wanted to learn about some basic speech processing techniques while being able to produce some simple results I could visualize. I modified some basic open source speech processing functions on the MATLAB File Exchange and wrote some of my own functions to process some basic audio recordings that I took using a bluetooth microphone. For my approach, I tried to generate spectrograms for audio of recorded words spoken by myself and a friend. I took a few samples and generated a reference set of spectrograms for a few basic words, which I mapped to basic operating system commands. A few examples included opening Microsoft Word, Excel, and PowerPoint. I attempted to distinguish between different commands by recording a live voice command, generating a spectrogram of the live recording, and comparing it with the reference set. What I was able to discover was that with minimal noise present, the code was able to distinguish between different commands. However, identifying the speaker of the command proved to be more difficult and I noticed less accurate results when testing my basic implementation. The MATLAB functions, some figures, and more details are posted on GitHub. This was a fun project and I hope to play around with it more on my own time in the future.

## V. Conclusion

There is a growing need and user demand for speech and communication devices and platforms. Multi-speaker identification is an area that holds potential but remains relatively unexplored, with optimal methods and approaches yet to be discovered. Efforts are being made in academia and industry to test the feasibility of potential products and new systems. Examination of user stories and applications suggests that even current popular platforms can be

improved upon and benefit from new approaches. Moving forward, speech technologies will continue to play an increasingly large role in influencing how people communicate and interact.

**References**

[1] A. F. Martin and M. A. Przybocki, “Speaker Recognition in a Multi-Speaker Environment,” *NIST Speaker Recognition Evaluations*.

[2] “Zoom Helps HubSpot Stay Connected as it Grows Globally,” *Zoom Case Studies*, 2017. [Online]. Available: https://zoom.us/docs/doc/HubSpot.pdf.

[3] “PORTS Program Expands, Enhances Student Access to California State Parks with Zoom,” *Zoom Case Studies*, Jul-2019. [Online]. Available: https://zoom.us/docs/doc/case/Zoom-Case-Californiastateparks.pdf.

[4] “Columbia Business School Brings Zoom to the Ivy League,” *Zoom Case Studies*, Oct-2019. [Online]. Available: https://zoom.us/docs/doc/case/Zoom-Case-ColumbiaBusSchool.pdf.

[5] G. Friedland and O. Vinyals, “Live Speaker Identification in Conversations,” *MM’08*, Oct. 2008.

[6] S. Shon, N. Dehak, D. Reynolds, and J. Glass, “MCE 2018: The 1st Multi-target Speaker Detection and Identification Challenge Evaluation,” *INTERSPEECH 2019*, pp. 356–360, Sep. 2019.

[7] F. Bentley, C. LuVogt, M. Silverman, R. Wirasinghe, B. White, and D. Lottridge, “Understanding the Long-Term Use of Smart Speaker Assistants,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–24, Sep. 2018.

[8] K. A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, “Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home,” *Annual Conference of the International Speech Communication Association*, Aug. 2011.