CALTRANSCENSE: A REAL-TIME SPEAKER IDENTIFICATION SYSTEM

Antonio De Lima Fernandes

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-55
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-55.html

May 11, 2015

Acknowledgement

- I thank my team: Helene Hillion, Mathilde Motte, Leonard Berrada and Nigil Valikodath for being the best capstone team ever!

- Thank you Transcense for giving us a fun technical challenge to solve

- Thank you to my thesis signers - Don Wroblewski and Bjoern Hartmann

- Many thanks to Alex Beliaev, IK Udekwu, and the MENG staff for thier support

- Finally, I thank God through whom I do all things.

CALTRANSCENSE: A REAL-TIME SPEAKER IDENTIFICATION SYSTEM

By Antonio Rohit de Lima Fernandes

Team: Helene Hillion, Mathilde Motte, Nigil Valikodath, Leonard Berrada, Antonio de Lima Fernandes

A report submitted in partial fulfillment
of the requirements for the degree of
Master of Engineering
in
Electrical Engineering and Computer Science
in the
Graduate Division
of the
University of California at Berkeley
Spring 2015

University of California, Berkeley College of Engineering

MASTER OF ENGINEERING - SPRING 2015

Electrical Engineering and Computer Science

Robotics and Embedded Systems

CalTranscense: A Real-Time Speaker Identification System

ANTONIO ROHIT DE LIMA FERNANDES

This Masters Project Paper fulfills the Master of Engineering degree requirement.

Approved by:

1. Capstone Project Advisor:

Signature: _____ Date _____

Print Name/Department: DON WROBLEWSKI/FUNG INSTITUTE

2. Faculty Committee Member #2:

Signature: _____ Date _____

Print Name/Department: BJÖRN HARTMANN/EECS

ABSTRACT

Speech is possibly the most efficient way humans can communicate with a machine. If acoustic waves could perfectly map to words, a variety of exciting applications could be realized. However, speech recognition is difficult in reality. One reason is that there is much variability in people's voices. In addition, many applications of speech-recognition require voice authentication or speaker-recognition in order to be practically useful. Speaker identification systems can account for voice variability, and also can provide voice authentication, and hence are an interesting area of research.

This report describes a real-time speaker identification system developed for a particular application – transcribing group conversations. Our capstone team worked with a startup – Transcense – which has an app that transcribes a group conversation in real time for the benefit of deaf and hard-of-hearing people. In this application, our speaker identification system allows the deaf person to know the identity of the current speaker, and hence follow the conversation better.

The speaker identification system we developed uses a dual approach to identify speakers: (1) Using voice features and machine learning (2) Using sound-source localization. The system achieves ~80% accuracy on a rigorous custom test-protocol with 4 speakers. This makes it a solid proof-of-concept that Transcense could use as a platform to develop their own speaker identification system.

Table of Contents

INTRODUCTION AND CONTEXT

In collaboration with Transcense, a one-year old startup, our capstone team is developing an app which helps deaf people follow group conversations through real-time captioning. By developing the app's user interface and a real-time speaker identification system, we are helping Transcense break communication barriers between the deaf and hard-of-hearing community and the rest of the world. The final goal of the capstone project is to provide Transcense with a useful proof-of-concept of a speaker identification solution optimized for their needs, and designed with a user-centered approach.

Participating in group conversations can be a daily challenge for the deaf and hard-of-hearing. Our studies have shown that the deaf often avoid situations with groups of people because of the difficulty of following the rapid exchanges of conversations in groups. Thus there is a real need for a solution such as Transcense's app, but there exists no comparable product, and substitutes are either too expensive or provide inadequate functionality.

Figure 1. The Transcense App

Figure 1 shows how Transcense captions a conversation between 4 speakers (the 4 differently colored text bubbles) by displaying both the text and the person who said it. Currently, in order for the app to work, every participant in the conversation needs to have their own microphone-equipped mobile device (smartphone/laptop/tablet) logged into Transcense. The participants are recognized by the device on which they are logged in. Even so, there often is the problem of nearby

devices picking up the wrong person's voice, and thus mixing-up or repeating parts of the conversation. Ideally, Transcense would require as few devices as possible, and yet be able to identify speakers robustly. As a Capstone team, our objective is to provide Transcense with a speaker identification solution that meets this challenge. In our case 'solution' includes both speaker identification (ID) technology as well as the user-interface design, though my contributions are mainly to the speaker ID technology.

This paper is divided into 5 distinct sections. This section introduces and provides context on our capstone project. The next two sections are devoted to Transcense's competitive and IP strategy respectively. The next section discusses my personal technical contributions to the speaker ID system we developed. The final section provides a summary of my learning and experience and concludes the paper.

Note: Throughout this paper 'Transcense' will be used to denote the product (the app) and the company interchangeably.

INDUSTRY AND STRATEGY

Overview

In the Introduction, we learnt that there is a need for a solution that would help the deaf participate in group conversations, but none exists on the market currently. However, this may not be the case for much longer: there are several large companies in related industries that could fill this gap on short notice. Consequently, Transcense wants to enter this market swiftly, assert itself as a leader in speech technology, and establish a loyal customer base. It intends to do so in 2 main ways: delighting customers and creating partnerships. Transcense's unrelenting focus on human-centered design is generating an app that is at once both completely novel and familiar; with it, communication is enhanced seamlessly. In addition, Transcense is building partnerships with strategic groups, notably a speech-recognition startup which will provide it the world's most accurate speech engine.

In order to be able to formulate a company's strategy, one must first understand the company, its industry and its competition. Hence, this section analyzes these topics in detail before explaining Transcense's corporate, competitive and marketing strategy. It is also important to note that since our capstone project is so closely tied with Transcense, a strong distinction is not made between Transcense and our capstone project.

Transcense's Industry

Though Transcense originally began as a 'smart glove' which converted sign language into text, through extensive user study the founders quickly realized that a bigger problem lies in the integration of the deaf into group conversations. Today, Transcense competes at the intersection of three industries: services for the deaf, mobile tech and voice-based productivity services.

The 'Deaf Services' industry broadly includes hearing aids, signers, note-takers, and any technology with the deaf as the primary customer. It is quite a large industry given that there are about 360M [1] deaf and hard-of-hearing people in the world. A related industry, the translation industry for instance, is $4.9B [2]. Mobile tech includes any software whose primary platform are mobile phones. It is a $9.7B industry projected to grow 30% over the next 5 years [3]. Voice-based productivity services are a relatively new class of industry which provide value to customers using

their voice as input. One familiar example is Apple's voice-activated assistant Siri. Each of these three industries are large and hence bring along with them an aggressive competitive landscape.

Competitive Landscape

This section familiarizes the reader with Transcense's competitors and industry trends, so as to form a basis of defining the company's competitive strategy.

Transcense's competitors are of 3 types:

1. Legacy services for the hearing impaired
2. Established companies in speech technology
3. Smaller companies focused on the deaf or on meetings

Legacy Services for Hearing Impaired

One of the major competitors to Transcense is legacy solutions targeted at the deaf and hard-of-hearing. These include hearing aids, listening systems, and real-time captioning services. These solutions are usually expensive: hearing aids costs thousands of dollars, while captioners are upwards of $100/hour [4]. In addition, hearing aids or listening systems don't usually provide enough of a hearing improvement to be an effective solution for deaf people trying to communicate in a group.

In order to get a flavor for this industry, we have picked one competitor to study in depth. CART (Communication Access Realtime Translation) is a service proposed by Speech-to-Text Reporters - also called Captioners - to help out deaf people following meetings, presentations, or lectures. Captioners listen to what is being said, type it verbatim onto an electronic shorthand keyboard and the text instantaneously appears on a screen for users to read and follow. Transcription includes dialogues, identification of the speaker when known and description of the sound where possible. Captioners can be either present in the room where the discussion takes place or be in a remote location. The only difference between live and remote captioning is that in the second situation, the audio of the meeting or lecture is captured by a microphone used by the speaker and transmitted to the captioner through a phone line.

Prices for CART services are very high. Live captioning costs between $150 and $400 per hour and remote captioning between $200 and $360 per hour [4]. In addition, the number of CART specialized Captioners is very low and is not enough to match the needs of people with hearing

disabilities. For instance, there are only 2,000 CART specialized captioners for about 60 million people with hearing disabilities in the US [5].

Transcense differentiates itself from the captioning industry by providing a service that is much cheaper (about $30/month), available 24/7 - while Captioners are only available at certain times and need to be booked well in advance - and that allows a real-time participation. With Transcense app, deaf people can not only understand what is being said but also participate in the conversation thanks to a text-to-speech feature implemented in the app.

Established Companies in Speech Technology

Though Transcense is targeted at breaking communication barriers, its core technology is in the area of speech. It therefore is necessary to study the industry related to speech technology as well. Input to smart devices is gradually moving from text-entry keyboards to speech-based methods of interaction. Google, Microsoft, Apple and Nuance are some of the largest companies active in the speech technology space with many other companies entering the fray. Facebook recently acquired Wit.ai - a company doing Natural Language Processing (NLP) for the Internet of Things.

As in the previous section, we have picked one competitor to study in depth. Nuance, arguably the most advanced speech recognition technology company in the world, is a public, $2B American company focused on productivity applications mainly using speech and imaging technology. Some of its better-known products include Dragon Dictation (a speech-to-text software for PC/Mac/Mobile) and Swype (a new text entry method for mobile devices). A bulk of Nuance's business also comes from providing business services – like customer self-service solutions, and medical transcription. In fact, Nuance claims to power self-service applications at 75% of the Fortune 100 companies [3].

At the heart of these solutions lies Nuance's natural language processing engine – a high accuracy and customizable speech-to-action and text-to-speech service. This system keeps getting better the more it is used, and has a number of flavors, each customized to industries like law and IT. Nuance is aggressively acquiring many smaller companies in speech or related technologies, including Dragon and Swype. This has particular significance for Transcense as they aim to be a technology leader in the domain.

<u>So why are these speech technology giants competitors to Transcense?</u>

Transcense's mobile captioning solution currently achieves its function by building a lightweight framework around Google's open speech-to-text engine. Even with its simplicity, Transcense provides a compelling use case – business meetings with deaf persons, and later possibly any sort of formal meeting. With its foot in the door, Transcense could organically add meeting-assistant functions – automatic meeting notes, action items, contextual suggestions, and even behave as a central knowledge repository for companies. Given just the business meeting space is so large in America (>3 billion meetings a year), there is a commensurately large revenue and data mining potential [6].

The revenue incentive is large by itself, but the potential access to data could be the killer application. With modern machine learning, particularly deep learning, more data means smarter systems, and Transcense is partnering with other companies to build speech technology in a few years that would rival the speech giants of today. In our capstone project, we are already developing speaker identification technology using a different approach than what these giants have established. The machine learning angle would put Transcense in competition with even more companies, for example Baidu Research, which specializes in artificial intelligence solutions, with speech as a focus area.

Smaller Companies Focused on the Deaf or on Meetings

The remarkable improvement of speech technology in the past few decades has given rise to a slew of voice-based companies recently. These companies directly or indirectly compete with Transcense. This section discusses three such startups of interest: ReMeeting, Gridspace and RogerVoice.

ReMeeting is a startup which has developed an app that serves as a smart meeting assistant. ReMeeting is a one-year old startup founded by speech recognition and speaker identification experts from UC Berkeley and Swiss Institutes. ReMeeting is targeting individuals and students – which is different from the professional business setting Transcense have chosen. Their product features include: record and archive meetings, summarize meetings in text, identify speakers, analyze meeting topics and keywords [7]. Even though Transcense and ReMeeting share some technologies and general objectives, their philosophies are very different. This is mainly due to the fact that they are not solving the same problem: as ReMeeting focuses on making meetings more

efficient with retrospective feedback, Transcense breaks real-time communication barriers for deaf and hard-of-hearing people.

Not much is known about Gridspace as they are a very new company. Gridspace started in mid-2014 and has only announced two products, Gridspace Memo (software) and Gridspace M1 (hardware), to "select early partners" [8]. Their goal is to "save and index meetings" [9]. They save a database of meeting archives in order to simplify communication and handoffs in the workplace. In that sense, Gridspace is similar to ReMeeting, but differs in two important regards: (1) They have hardware (Figure 2) (2) They have established strategic partnerships with leading businesses, rather than individuals.

Although Transcense isn't currently in this space, we fully intend on entering this space on the longer term. We appreciate the power of providing data analytics, and intend on enhancing communication not just for the deaf community but any space that needs it.

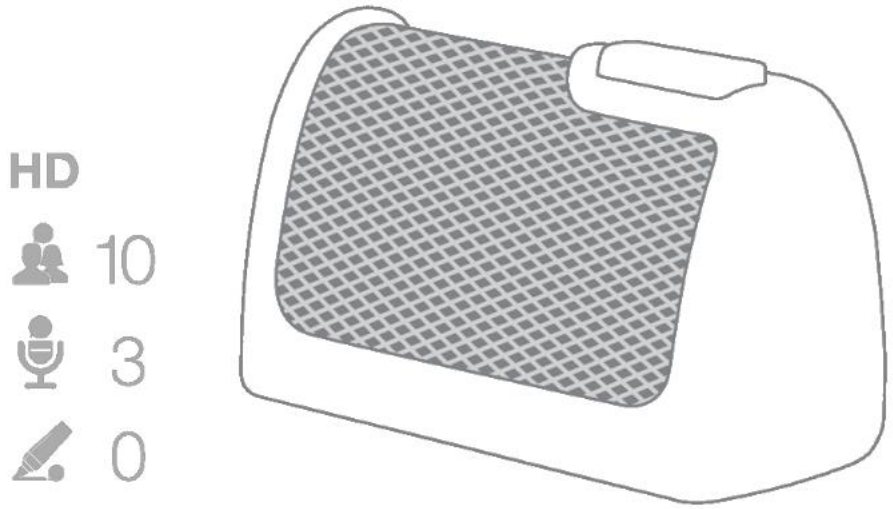

Figure 2. Gridspace M1 - supports 10 person meetings, 0 need for note-taking

RogerVoice is a French start-up which makes possible phone calls to deaf and hard of hearing individuals. They are developing an app that transcribes phone conversations and displays the text on the screen, in real-time. The solution is built on a Voice over Internet Protocol (VoIP) platform and automatic speech recognition. It is mainly intended for deaf and hard of hearing to make and receive phone calls, but it also provides archiving and searching functions. It allows you to get instant written minutes of phone calls, store conversations and re-read them at anytime. Although RogerVoice and Transcense have a very similar customer segment - people with hearing loss - and

use a similar technology - speech-to-text - the two companies offer different value propositions to their customers. On one hand, RogerVoice has a very specific use case: it is restricted to conversation over the phone. On the other hand, Transcense is focused on in-person conversations.

Overall Strategy

Now that we have understood the nature of Transcense's industry and its biggest rivals, we are in a position to map out a strategy to allow Transcense to differentiate itself and succeed. We used two frameworks - The 5 Porter forces model [10] and the SWOT analysis. Only the conclusions from these analyses are recorded in this document.

Conclusions from Porter's 5 Forces

Using Porter's framework [10], we concluded the following (see Appendix A for more details):

- The threat of new entrants is high
- The bargaining power of Transcense's suppliers (particularly Google speech API) is high
- There is a significant threat of substitute products being used instead of Transcense's app

Conclusions from SWOT

SWOT stands for Strength Weakness Opportunity Threats, widely used since 1989 to develop a company's strategy. We used this framework to conclude the following:

- Transcense has strong positioning, but no IP
- Transcense's value proposition is clear, but it currently has no technology protection
- It lacks in-house NLP expertise
- Transcense wants to be a leader in NLP over the next 10-years, but currently has no NLP IP or experts
- Its limited financial resources are stifling
- Transcense needs to raise money to be able to grow

Organizational Strategy

Organizational strategy is Transcense's strategy looking inwards towards the company. From the SWOT analysis, we realized that a lot of Transcense's momentum was derived from the implicit company DNA (a strength). Here are the company's mission, vision and core values in an attempt to make this DNA explicit:

Mission: To break communication barriers between the deaf and non-deaf through technology

Vision: Be the deaf community's go-to solution for any group conversation (5-year)

Be the technology leader in Natural Language Processing (NLP) in the context of group conversations (10-year)

Core Values: The only two defined so far are: 1. The customer, not technology is the focus of our design 2. Our products create social change and revenue.

The mission, vision and core values directly affect Transcense's short and long-term strategy. For example - Transcense's focus on group conversations puts it in a market niche. The emphasis on human-centered design over technology allows Transcense to have technology agility - pick the best technology to give the customer a better experience. Features that improve a deaf person's life (social change) are prioritized over strictly aesthetic concerns for example.

Another key finding from the SWOT analysis is that Transcense lacks NLP expertise. However, its long-term vision is to be an industry leader in NLP for group conversations. This has motivated Transcense's current strategy to prioritize hiring an industry expert in Natural Language Processing (NLP) and Machine Learning (ML).

Competitive Strategy

Competitive strategy is Transcense's strategy looking outwards from the company. It forms a response to the threats identified through the Porter 5 Forces model and SWOT analyses. There are three main thrusts for Transcense's competitive strategy:

<u>Transcense's strategy versus the threat of new entrants</u>

One of the critical forces is the threat of new entrants. However, Transcense benefits from the network effect - the phenomenon whereby a good or service becomes more valuable when more people use it. In order for the app to be useful, the deaf person needs to have all his colleagues using the app. The envisioned scenario is that the deaf person will ask his friends and family to download the app so that he can use it with them. Once his personal circle has it, even if a similar and even cheaper product were to come to market, the deaf person would need a very high incentive to change the app and ask all his friends to do so again. Because the cost of switching is psychologically high for its customers, Transcense's strategy is to have the product reach as fast as possible the market (by May 2015), so that virality can start.

Transcense is also using this - aggregating a customer database of voices - as its key IP strategy for protection, valuation and differentiation.

Transcense's strategy versus the bargaining power of suppliers

In order to mitigate the bargaining power of Google speech engine, Transcense has developed partnerships with a startup from Carnegie Mellon University named CAPIO. CAPIO has the most accurate speech recognition engine on the market and has agreed to give Transcense the technology in exchange of voice data. Transcense has also established a partnership with a leading speaker identification expert who wants to provide his technology to Transcense.

Differentiation to face the threat of substitutes

In order to avoid substitution, Transcense's strategy is to emphasize differentiation. To compete with legacy services like captioners (e.g. CART), Transcense uses differentiation both in price and flexibility. Transcense targets a price of $30/month for the deaf or hard-of-hearing person which is far below what currently exists on the market. To compare, it costs between $75 and $200 per hour to have real-time captioning [5] and around $4000 [11] to buy a hearing aid device.

Market Strategy

Naturally, after developing organizational and competitive strategy we need to solidify the go-to-market strategy. In other words, how can we effectively advertise and launch this product to the users that desperately need this solution. We use the four P's framework to properly form a coherent and logical go-to-market strategy. The four P's are product, price, place and promotion.

Product

The first beta and open launch will be on the Android operating system as well as a web application (for those that do not have Android). The public launch is scheduled for May 2015 and will launch on iOS in late-2015. The product development is, as discussed earlier, based on the human-centered design process. The deaf/hard-of-hearing community is truly building this app. We have our initial ideas/mockups and test weekly to validate or disprove those assumptions. This is constantly refined until a majority of the beta testers, as well as some fresh eyes, are pleased. The current strategy is to test the alpha app on 50 testers that use the app regularly. After the 50 are adequately pleased and the user experience/interface (UX/UI) is near final, we will move to 500 beta testers. iOS development will start around this time. Then Transcense will move to a public launch on Android and the webapp.

The marketing strategy for the product goes beyond these short term goals. As discussed earlier, there are immense possibilities with the underlying technology of Transcense. Many things can be utilized with a full transcription of a meeting or conversation. These moonshot ideas are also being tested throughout this user-centered design process. For example, with a transcription of an entire meeting, you can deliver metrics from the meeting as well as tips to improve productivity. Alternatively, this transcription could help not only the deaf/hard-of-hearing but also foreigners that have various levels of competency in conversation in another language. The goal is to give an incentive for all people to use this app, not just the deaf and hard-of-hearing.

Price

The pricing strategy of is a bit unique for Transcense as opposed to most mobile phone app. Generally, mobile phone apps have a one-time fee or in-app purchases. Our app will be free to download but will have a subscription model, similar to Netflix or Spotify. Transcense will rely on having a "free trial period" of 10 hours of transcription free group conversations per month to bring customers. After that, they will sell the transcription of the group conversation. You will still be able to download and contribute to other conversations with the free application. However, in order to see the conversation on the app (after the free trial period), you must pay $30 per month. There will also be an incentive for added hours per invitation to Transcense. This strategy allows users to try the product and love it before they make an investment.

Place

The question still stands, where will they make that investment? Well, the app itself will be available on the Google Play Store upon launch of the public beta. It will also be accessible as a web app, using a simple url such as: http://ava.me. However, the subscription model will not be charged through the Google Play Store since it will be offered as a free app. Rather, purchasing will be on an account basis internally within the Transcense app. This will prevent pirating as well as allow users to have a free trial of the app if they want it longer than the 2 hour trial period Google Play allows for paid apps [12]. We want users to love this experience and naturally desire for their friends to have it.

Promotion

This leads us to the final P, which stands for promotion. How will users know to even buy this product? As discussed in our competitive strategy using Porter's Five Forces, one of our weak

points (where there is strong bargaining power), is bargaining power of suppliers. Here, our strategy was to build partnerships to withstand the strong hold the suppliers currently have on us. i.e. Google, Amazon Web Services, etc This will also serve as a method of promotion. Having strong partnerships will undoubtedly lead to endorsements.

Promotion has already happened through the Indiegogo campaign that was launched last October 2014. Over 400 funders participated and were willing to pay for the cause even before it started! It was a great way to get funding but more importantly get the word out about the product. Transcense was also featured on several technology sites, including TechCrunch. This is the foundation to the primary method of promotion: word of mouth.

Since it is for a good cause and helping the disabled, there are several supporters of the mission. Furthermore, deaf and hard-of-hearing people can communicate orders of magnitude easier than before. If one person likes it, they'll recommend it to one or two people they know in their community, and them another two, and so on. Ideally, they will be compelled to share it because it is a intuitive and wonderful experience for them as well. This app is driven by the users, constantly receiving feedback from them and refining the design to make it a better experience. It is also a group conversation app, so multiple people in one use of the app will have knowledge to further increase the likelihood of promotion.

Conclusion

Though there is a real need for Transcense's app, there exists no comparable product, and substitutes are either very expensive or not good enough. However, this may not be the case for much longer: there are several large companies in related industries that could fill this gap on short notice. Consequently, Transcense wants to enter this market swiftly, assert itself as a leader in speech technology, and establish a loyal customer base. It intends to do so through 2 main ways: delighting customers and creating partnerships. Transcense's unrelenting focus on human-centered design is generating an app that is at once both completely novel and familiar; with it, communication is enhanced seamlessly. In addition, Transcense is building partnerships with strategic groups, notably a speech-recognition startup which will provide it the world's most accurate speech engine.

IP STRATEGY

Overview

This section assesses the value our work adds to Transcense's Intellectual Property (IP). None of the technologies we are using are patentable by themselves, but it could be possible to patent a combination of these techniques in the specific use case of deaf people. Having a patent would increase Transcense's valuation and give them competitive advantage. However, filing a patent costs time and money: resources that Transcense does not have at the current time.

Presentation of the technology and possible patent

As a reminder the use case for Transcense is the following: several people are gathered around a table for a meeting or for dinner. The number of microphone-equipped devices available is smaller than the number of people around the table and the app is installed on these devices. The devices continuously record audio and send this data to Transcense's servers, which in return send back the text transcribed.

We make use of two technologies to know who is speaking: voice identification and sound localization. Voice identification is based on Machine Learning: we train our system to learn the discriminating features of each user's voice, in order to distinguish people's voices when they talk. This requires a learning step, which means that this system can only work if it knows features of this speaker's voice: we need audio data from this speaker in our system beforehand. The machine learning models we are using are found in speaker identification literature (see a review in [13] for example). As the US law states, one "can't claim what others have already published, either in patent or scientific article form, neither patent a design for an object which has been available for sale (hence the importance of a literature search" [14]. Therefore, the speaker identification process using machine learning to distinguish voice characteristics is not patentable.

We also identify the speakers with sound localization. Indeed, having several microphones receiving sound at (very slightly) different times allow us to approximate the spatial location of the audio sources. As for our first model, articles in the scientific literature describe the techniques we are using, which prevents us from patenting the source localization technology just by itself [15].

However, we might be able to patent the integration of the two technologies, in the specific use case of deaf and hard-of-hearing people in a group conversation. The American law is very specific on patents given to new use cases of an existing technology. As attorney Andrew P. Lahser explains: "If the patent claim only includes the old structure or composition, and, the "use" is simply a result of that old structure or composition, then the claim will not be allowed. However, if the "use" of the old structure or compound has an unexpected result, then, the patent claim can be granted. Also, the new property like must not be "inherent" to the existing product or old idea." [16]. In our case, the integration of two existing methods in order to build a more robust real-time speaker identification technology could be filed under the definition of an "improvement patent" that "can add something to an existing product, incorporate new technology into an old product, or find a new use for an existing product" [17]. The improvement patent would apply both to the merging of two existing technologies and the use case of deaf and hard-of-hearing people.

What is the added value of patenting for Transcense?

There are several reasons for Transcense to file patents. First of all, as seen in the strategy paper, barriers to entry are low in Transcense's industry and several big players, such as Microsoft, Apple and Google are already using speech-to-text technology in some of their products [18]. These companies are not targeting the deaf and hard-of-hearing market yet but could easily create new products that would. Patents would create a barrier to entry and give Transcense some competitive advantage. Secondly, patent ownership helps startups with raising funds. For investors and VCs especially, patents signify a smart team and a unique product. They are also a tangible asset that reduces an investor's downside risk. Finally, having several patents increases Transcense's valuation in the event that it is acquired.

Is it relevant for Transcense to patent today?

Transcense is a young and small startup, with only 3 full time employees. Patenting an idea takes time and money, the two scarcest resources of a startup. Indeed, to file a non-provisional patent, it is recommended to hire a patent agent or patent attorney that would cost between $200 to $400 per hour. The total cost to file a patent is more than $10,000 (between $5,000 and $10,000 just in attorney fees according to the United States Patent & Trademark Office Fees [19]). This explains why Transcense has chosen not to file a patent for the time being. This decision also makes sense with the global trends of the startup world: "Over the past several years, the average popularity of

patents has steadily declined among funded technology startups" [20]. However, Transcense is considering the option to file a provisional patent. It is cheap - $65 - and would give them a year to file the real patent. [21]

Conclusion

As seen in this section, the technology that we are developing as part of the Capstone project could be filed under an improvement patent and would bring value to Transcense. However, it costs money and time to file a patent and Transcense does not these resources yet. To mitigate the risks associated with lack of patented technology, Transcense's strategy is to go to market as soon as possible and develop a wide and loyal user base. By creating a network effect, Transcense will make it harder for new entrants to penetrate the market and this will give Transcense a strong competitive advantage.

TECHNICAL CONTRIBUTIONS

Introduction

As mentioned in the first section of this paper, the final goal of our capstone project is to provide Transcense with a useful proof-of-concept of a speaker identification solution optimized for their needs, and designed with a user-centered approach.

One of the biggest limitations of Transcense currently is that it requires one device per participant in the conversation. In addition, there often is the problem of nearby devices picking up the wrong person's voice, and thus mixing-up or repeating parts of the conversation. *It would be ideal that Transcense require as few devices as possible, and yet be able to identify speakers robustly*. As a Capstone team, our objective is to provide Transcense with a speaker identification solution that meets this challenge. It is useful to note that in our case 'solution' includes both speaker identification (ID) technology as well as the user-interface design, though my contributions are mainly to the speaker ID technology.

The technical team for this project comprises of Léonard Berrada, Hélène Hillion and me. Léonard and I co-developed the speaker ID subsystem based on machine learning, while Hélène and I co-developed the sound source localization (SSL) prototype described at the end of this section. The rest of this section of the paper discusses my technical contributions to these two aspects of the project.

Speaker Identification

Speaker identification[1] is the act of recognizing the person who is speaking. In the Transcense app, identifying the speaker is useful for (a) Performing speaker diarization (indicate who said what) (b) Ensuring speech-to-text robustness.

Every person's voice is unique. This uniqueness is due to two factors: a) The anatomy of the vocal chords and mouth b) Speaking style. Traditional speaker ID uses these features to recognize a person. However, in Transcense's use case: a group conversation, we brainstormed multiple ways to do speaker ID. Here are a few:

1 It is worthwhile distinguishing speaker identification from speech recognition, which is the act of understanding what has been said.

1. ‘Learning’ a person’s voice features (traditional method)
2. Spatial location of participants
3. A participant ‘tagging’ the conversation
4. By the phone accessing the app (current solution)
5. Using the cameras on the participants’ on the phone to identify participant’s faces

Currently, Transcense uses option 4 to identify speakers. However, this is not robust, because of the problem of nearby devices picking up the wrong person’s voice. Option 5 (the camera) is not feasible due to power and privacy constraints. Option 3 (tagging/correction) could be a useful feature to have on the app, no matter what the speaker ID method. We hence investigated options 1 (‘learning’ a person’s voice) and 2 (spatial location) and implemented these in our speaker ID solution.

Speaker ID using Voice Features

Speaker ID using voice features can be achieved through a machine learning system. Machine Learning (ML) is a field of computer science which uses algorithms to learn from data. Figure 3 provides a broad view into how this is achieved in the case of speaker identification. The process of speaker ID using ML can be divided into 2 phases: Training and Testing. During training, voice features are extracted from audio data taken from multiple speakers, and used to train a database. During testing (identification), previously unseen data (from any of the speakers used to train the system) is fed into the system, and the system predicts the speaker through comparison with the database. Machine learning is the primary way that our solution identifies a speaker.

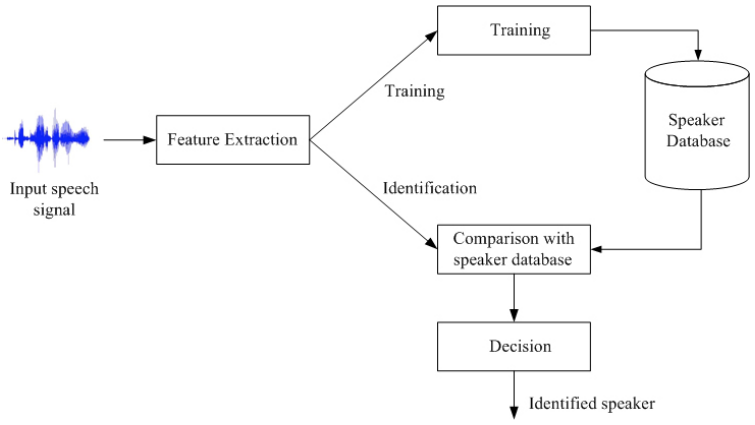

Figure 3. Speaker Recognition using Machine Learning

Figure 4 shows a little more detailed view of the testing phase of a ML system for speaker ID. Initially, I focused on the preprocessing stages in the signal chain - cleaning the audio data of noise and silence, and extracting features out of that cleaned data. In addition, to be able to tune this system fully, I implemented an exhaustive, parametrized test suite.

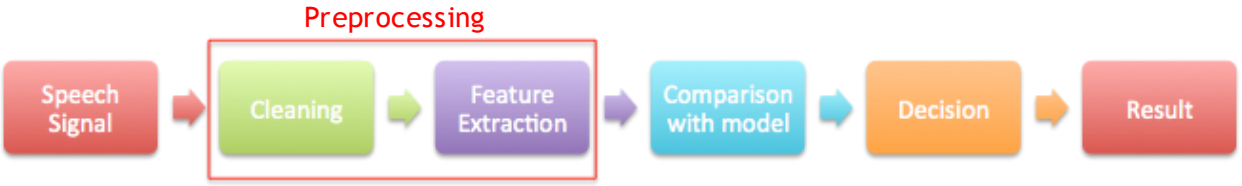

Figure 4. Signal Chain for Testing Phase in Speaker ID

Towards the end of the project, I began to contribute on the machine learning models as well, implementing a voice-based gender classifier which worked in parallel with the main speaker ID system to boost its accuracy.

Speaker ID using Sound Source Localization

In addition to recognizing speakers based on their voice signature, speakers could also be distinguished based on their position around the microphones. The theory behind this is that a speaker's distance from a microphone results in a time delay of arrival of the signal from the speaker on that phone. Using multiple phones, a speaker's position can be exactly triangulated. Even if the speaker's identity is not known, knowing the speaker's location could be used to achieve both the goals of diarization - person A said x and (a different) person B said y, as well as ensuring speech-to-text robustness - text is not mixed up or repeated due to multiple devices picking up the same voice. Hélène in her paper describes how we are using Sound Source Localization (SSL) to enhance speaker ID using machine learning.

The rest of this paper deals with the parts of the speaker ID system that I built: the preprocessing methods for feature extraction, the parametrized test suite, and also briefly discusses a physical prototype we built to demonstrate SSL technology. Léonard covers the machine learning aspects and Hélène the SSL technology in their technical papers [23], [24].

Overall Solution

Python was chosen as the language of choice for implementing the speaker identification solution given the team's familiarity with it, the Python ecosystem (availability of example code and

extension packages) and its ease of use. We used a private GitHub repository for source control, and Eclipse as the code editor. We drew a lot of inspiration from Spear - an open source toolbox in Python for speaker recognition [26].

Preprocessing

Signal Cleaning

Conversational speech contains repeated instances of speech followed by silence. Hence the first step we perform for feature extraction is Voice Activity Detection (VAD), i.e. identifying which parts of the audio data contain useful signal, and which parts are silent. This is a problem that is well documented in literature with a variety of methods. For use in Transcense, a VAD algorithm needs to be able to (a) Separate voice from silence (b) Adapt to varying background noise (c) Be computationally efficient (real-time). For our project, I chose one such method that meets these criteria: Adaptive Energy Detection (AED), documented in [30].

Adaptive Energy Detection

For AED, the signal is first broken into smaller chunks (frames) of k samples and the energy of each frame is calculated:

$$E_f = \frac{1}{k}\sum_{1}^{k} x(i)^2$$

Where E_f is the energy of the frame, $x(i)$ is the signal at the i[th] sample. Frame duration is taken to be ~20ms but is tunable system parameter.

The average energy of background noise is then estimated from a portion of the audio known not to contain speech (the first 200ms is a good approximation).

$$E_b = \frac{1}{v}\sum_{1}^{v} E_m$$

Where E_b = initial threshold estimate, and v = number of frames in prerecorded sample.

The decision rule for classifying a frame E_j as speech or silence is then

$$E_j > kE_b$$

where $k > 1$ is another tunable parameter of the system.

If TRUE, then the frame is considered speech, and if FALSE, the frame is considered to be silence.

Since background noise can vary, an adaptive threshold is more appropriate

$$E_{bnew} = (1 - p)E_{bold} + pE_{m\_noise}$$

Where E_{bnew} is the new background noise threshold, E_{bold} the old noise threshold, E_{m_noise} is the most recent noise frame, and p is a responsiveness parameter < 1

We implemented this algorithm in our system with varying frame lengths, values of p, and initial threshold, and ran tests to find the parameters that work best.

Figure 5 shows the output of the AED technique for p=0.1, frame length of 10ms, and an initial threshold taken from the first frame (clearly visible to be background noise in this case). The algorithm looks like it is working well, and this was verified by exporting the cleaned signal to a wav file and listening to it.

Periods without speech and only a little background noise are completely silenced (shown), and then deleted from the signal (not shown). Deleting the periods of silence is crucial so that the machine learning system is not overloaded with redundant data.

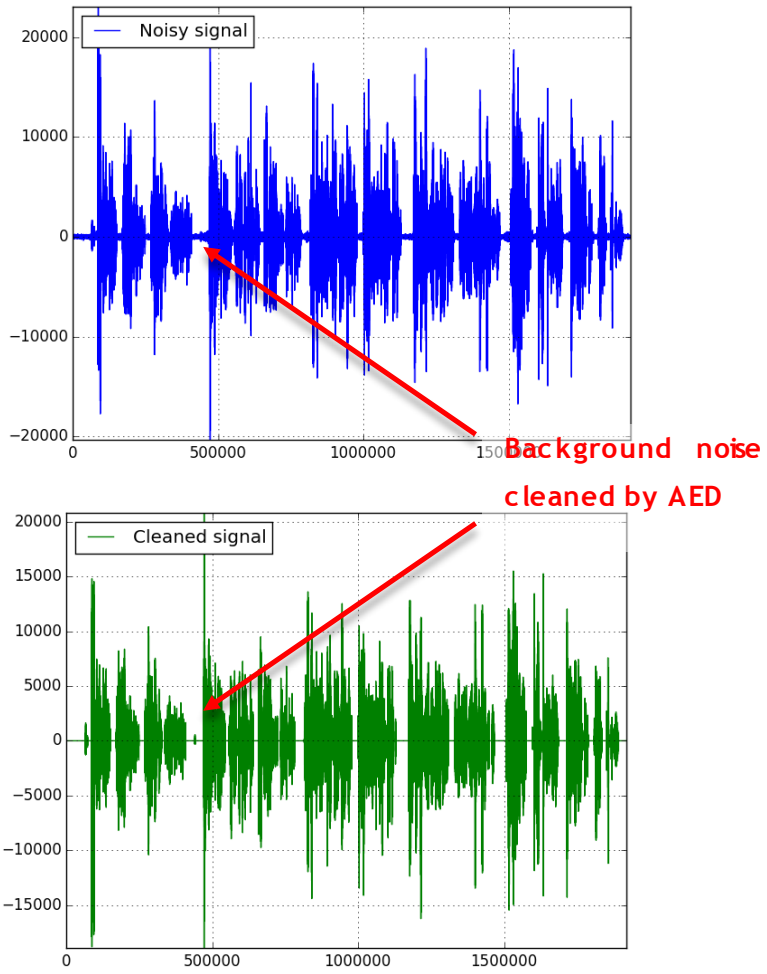

Figure 5. Results for AED

AED is chosen over other methods like Simple Energy Detection (SED) or Weak Fricative Detector (WFD) for its simplicity and performance. However, since it is a purely energy-based method, AED suffers from the drawback that weak fricatives like "to*w*", "of*f*", "lo*w*er" are completely silenced. In reality though, this may not be a major concern in speaker identification, because sufficient information is contained in high energy voiced speech segments such as vowels.

Other methods

Though energy-based methods are most popular for their simplicity and effectiveness, VAD can be done in other ways as well. Speech has a characteristic energy modulation peak around the 4-Hz syllabic rate usually considered as evidence of its presence [31]. This fact is used to create another popular speech detection technique called '4hz modulation', which is usually used in conjunction with a signal energy classifier. The classifier is a simple speech activity detector where

frame-level energy values are computed, normalized and then classified into two classes. The class with the higher mean (higher energy) is considered as speech, and corresponding speech segments are retained [26]. My first attempt to create such a classifier failed, and then we abandoned this VAD technique because it is likely too complex for a real-time implementation such as is needed for Transcense.

Kinnunen and Li [32], also suggest Long Term Spectral Divergence (LTSD) as an alternative to energy-based methods for VAD in real time systems. Though at present we have not implemented LTSD, we keep it in mind as an alternative, should AED fail.

Feature Extraction

Once the audio signal is cleaned, it is ready for feature extraction. Feature extraction refers to identifying the components of the audio signal that discriminate speakers, and discarding the rest. As we saw in the previous section, even silent periods in an audio signal are considered 'noise' and discarded since silence does not help distinguish speakers.

Human speech has a number of discriminative features usually seen as different energies at different frequencies. Mel Frequency Cepstral Coefficents (MFCCs) are the most popular speech features used in automatic speaker recognition. They were introduced by Davis and Mermelstein in 1980 [25]. We use MFCC features in our system.

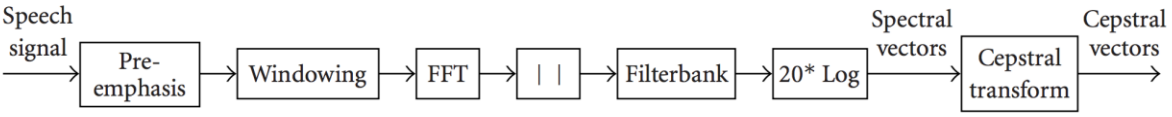

Figure 6. Modular representation of a filterbank-based cepstral parameterization

Figure 6 presents a block diagram view of the various steps involved in MFCC extraction.

- Filter the signal to emphasize the higher frequencies which are usually reduced by the recording device – Pre-emphasis
- Frame the signal into short frames – windowing.
- For each frame calculate the power spectrum (power at each frequency) estimate of the power spectrum. This is shown by the FFT and mod steps in Figure 6.

- Apply the Mel filterbank to the power spectra, sum the energy in each filter. The Mel filterbank allows us to clump the frequencies into bins based on working of the human ear.
- Take the logarithm of all filterbank energies. This is also done because the human ear does not hear sound on a linear, but logarithmic scale.
- Take the Discrete Cosine Transform (DCT) of the log filterbank energies – Cepstral Transform. DCT de-correlates the information in the bins and concentrates the information into the lower DCT coefficients.
- Keep DCT coefficients 2-13, discard the rest. Higher DCT coefficients represent fast changes in the filterbank energies and are smaller in magnitude. However, these fast changes actually degrade speaker ID performance. Coefficient 1 represents the overall energy contained in the signal, and this is also discarded because it usually does not help discriminate speakers.

For most of these blocks, we used James Lyon's open-source Python implementation of MFCCs available on GitHub [27], so I shall not go into more detail here. The results of the MFCC block are not intuitive to look at, so I have omitted these results from the report. The final accuracy of our system reflects the results of the feature extraction.

Gender Classifier

As a final project for a machine-learning class, I developed a voice-based gender identifier [33], re-using much of the intuition learnt from our speaker ID system. The gender identifier uses slightly tweaked MFCCs in addition to the pitch of speech as the features, and a Random Forest classifier to achieve ~90% accuracy on the test protocol. Given the accuracy of the gender classifier, its output is used to refine the possibilities of predicted speakers in our speaker recognition system.

Test

Given the complexity of the speaker ID system we developed, it became necessary to develop a dedicated test system to verify its performance. Towards this goal, Leonard initially developed a Graphical User Interface (GUI) as a way to visualize in real time who was speaking – see Figure 7. The GUI, running on a laptop, would listen to speech and display the picture of the person it predicted as the current speaker. However, this provided a qualitative appraisal of the system, but

not accuracy numbers or intuition for improving performance. Clearly, a more rigorous and robust parametrized test suite was necessary.

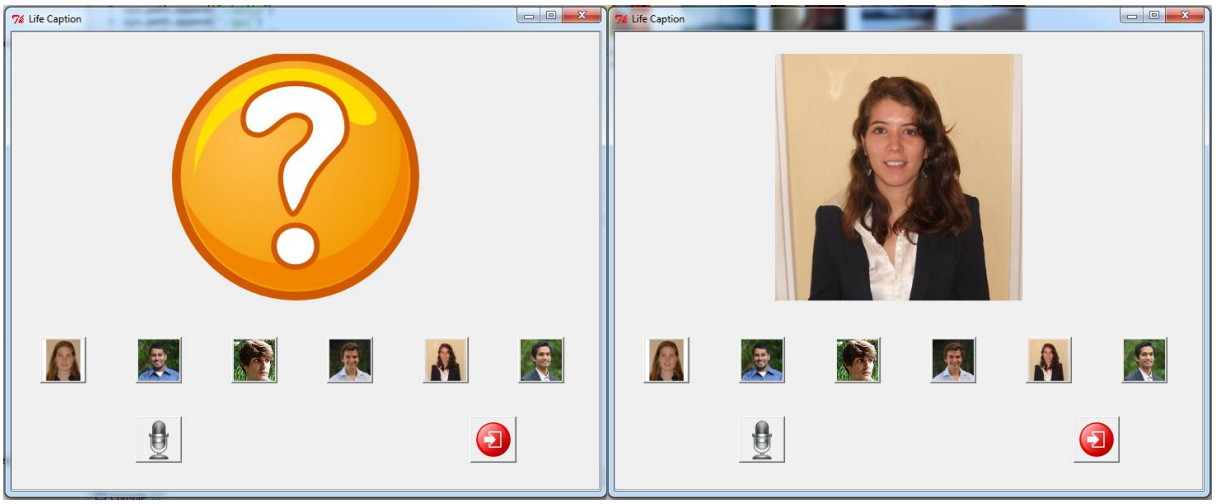

Figure 7. Speaker ID Graphical User Interface Showing the Current Speaker: Silence (Left), and Helene (Right)

Test Protocol

We developed a test protocol that described the audio files to be used to train and test the system. At a high level, we created 2 files, 24 minutes each, one to train and the second to test the system. The recordings were done in different controlled environments: clean (silence), with background noise (BGN), with background voices (BGV), and with echo. Recordings were captured on both an Android device as well as an Apple device. The recordings were then labeled for speakers and conditions using Audacity (software). Figure 8 shows one of the audio files recorded and labeled according to this protocol. For more details on the protocol, visit Appendix B. This protocol ensured that we had a repeatable, reliable and complete method of testing the system.

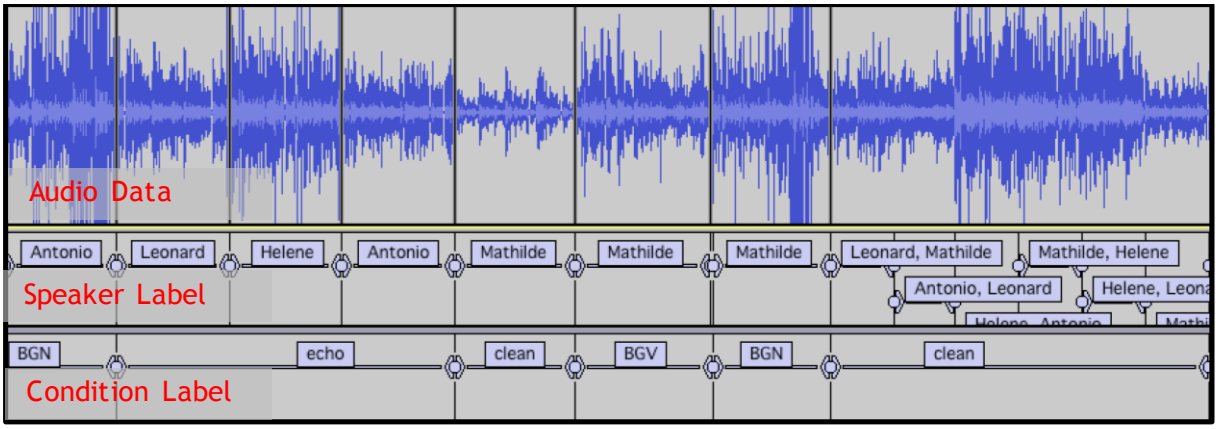

Figure 8. Sample of Audio File generated According to the Test Protocol

Test Suite

The test suite is a set of Python functions which take as input *.wav audio and *.txt label files, user-defined system parameters and what variables to test, and outputs the system accuracy for each of the cases tested, and records the best accuracy observed.

It performs the following steps in order

- Splits the test and training *.wav file into multiple wav files based on speaker and condition, using the labels in the corresponding *.txt file
- Assigns each of these files to a speaker
- Build a Gaussian Mixture Model[2] (GMM) for the speakers using the training audio files
- For each condition in BGV, BGN, Clean and Echo, test the speaker ID accuracy across speaker test files. Note that each of the test files (a few minutes long) is broken into smaller chunks to simulate real-time processing as happens with our GUI.
- Output the results of the test into a text file.
  See APPENDIX C: Test Output File for Android Device for an example of such an output.

Test Results

We ran multiple tests with many different parameters, and here are some observations:

- **Overall**: The average speaker ID system test accuracy peaked around ~82%. State of the art speaker ID solutions have >90% accuracy.
- **Conditions**: The speaker ID system performed best in the case of clean data (by 5%) over BGN and BGV, and worst in the case with echo.
- **Devices**: The recordings done with the Android device (Nexus 9 tablet) and from the Apple (Macbook Pro) device had comparable accuracy, even though the Apple recordings were stereo and the Android recordings were in mono.

2 See Leonard's technical contributions paper for more details on GMMs

- **Chunk Size**: Tested chunk sizes from 0.25s to 2s. Framing the audio data into larger chunks before testing results in better overall accuracy. We achieved ~82% accuracy with 2 second long chunks, but ~60% accuracy with 0.25 second chunks.

  This makes intuitive sense because larger chunks provide the speaker ID system with a more complete picture before having to make a decision.

- **Number of Gaussians**: Tested with 1, 10, 20, 30, 40 gaussians. The more the gaussians, the better the accuracy, but slower the performance. The best performance was achieved with 20 GMMs – 82% (see Figure 9).

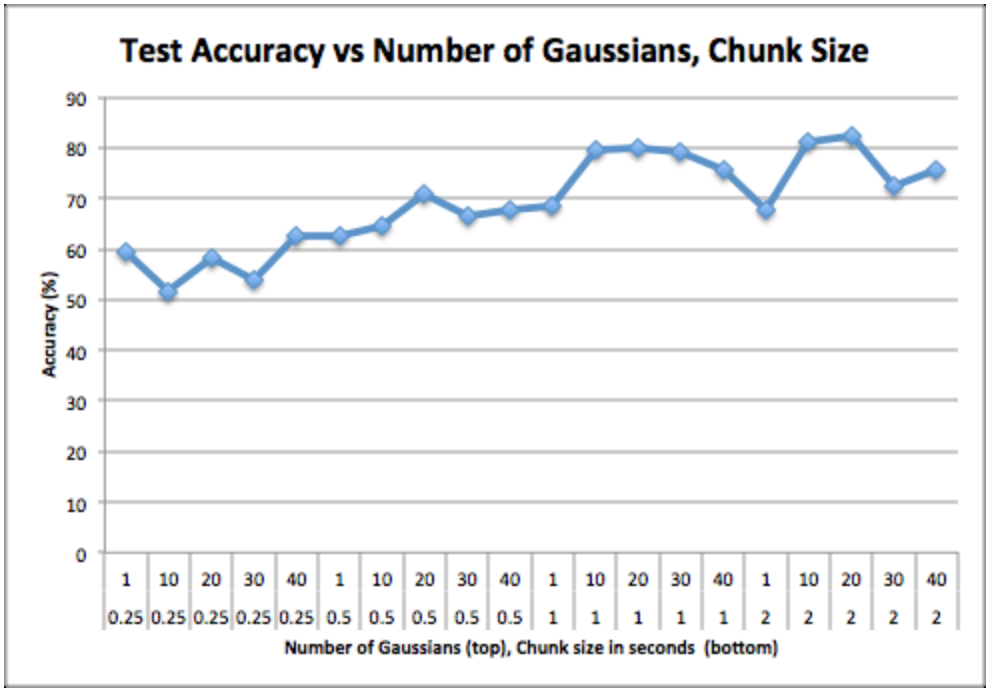

Figure 9. Test Accuracy vs Number of Gaussians, Chunk Size

- **With and without gender classifier**: We tested the contribution of the gender classifier, and found that it boosts overall accuracy of the system by about 5% on average (Figure 10).

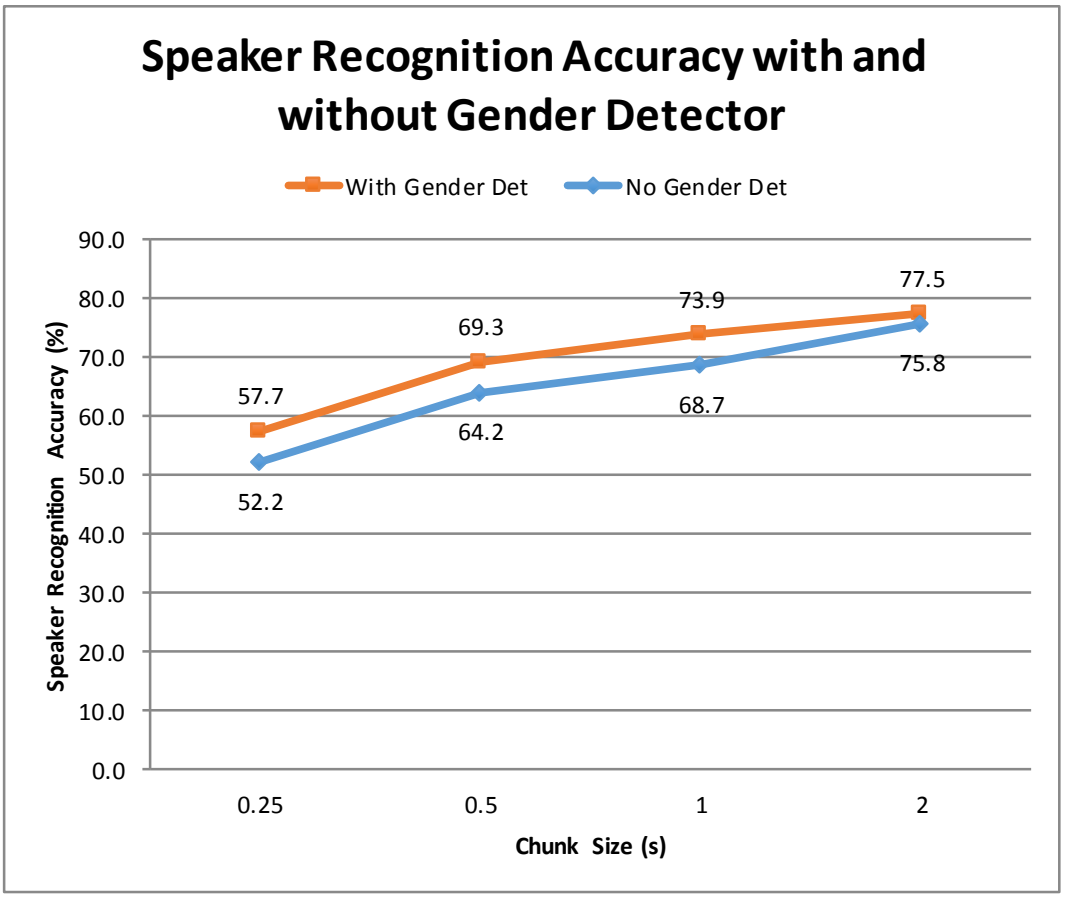

Figure 10. Effect of Gender Detector on Accuracy of Speaker ID

- **Cleaning parameters**: Setting p=0.1 and including a factor of 2 in the adaptive threshold works best:

$$E_{bnew} = (1-p)E_{bold} + 2pE_{m\_noise}$$

Physical Prototype for SSL

In addition to working on the pre-processing for the speaker ID system, I also designed, assembled and tested a physical prototype to demonstrate speaker ID using SSL in collaboration with Helene and Nigil.

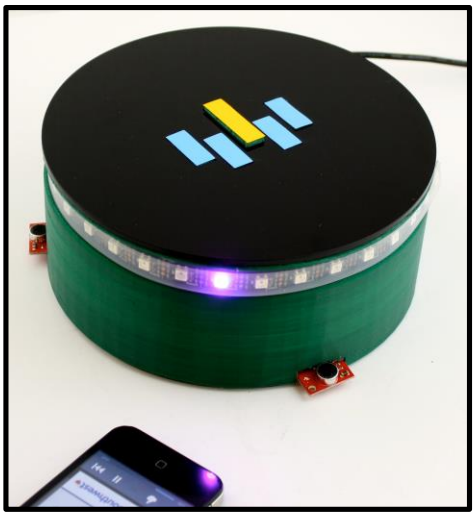

Figure 11. Physical Prototype to Demo SSL

The device shown in Figure 11 consisted of a KL25Z embedded board connected to 3 Sparkfun electret microphones and a strip of RGB 'neopixel' LEDs. The device was powered over USB, and the same USB cable was used to communicate with the device over serial. The KL25Z board has on it a Kinetis ARM-M0 48MHz processor with a 12-bit ADC. This was used to sample the 3 microphones at 30,000sps (samples per second) and continuously output these results over serial to a Python script. The script then utilized the algorithms described by Helene in her paper to triangulate the speaker's position using cross-correlation between the 3 microphones. This position was then relayed back to the embedded processor, and this information was used to light up an appropriately positioned LED.

The body of the prototype was 3D printed and laser cut and then hand-assembled by Nigil and me. The prototype worked reasonably well even when demoed in the noisy environment of the capstone expo at the end of Fall 2014. In Figure 11, it is shown responding to music played on the phone by lighting up an LED.

Conclusion

This paper introduced the Transcense app, and how our capstone project is meeting the challenge of reducing the number of devices required for the app to work robustly. The paper went on to explain how our team is solving this problem by implementing speaker identification using machine learning (ML) and sound source localization (SSL). In particular, it explained my contributions to the preprocessing stages of the ML implementations, the gender identifier, and the physical prototype we built to demo SSL.

Given that we already have a reasonable demo each for speaker ID using ML and SSL, the next step is to further develop each of these: tune the ML approach to gain optimal accuracy, and re-build the SSL solution using recent literature. The next step is to integrate the two methods into one and demonstrate that this increases the accuracy of the speaker ID system. We have handed off the entire project to Transcense who plans to use this system as the starting point for their own speaker identification IP.

CONCLUDING REFLECTIONS

In the beginning of the fall semester, we had very broad objectives for this project, and envisioned that our work would be integrated into Transcense by May. Since then, we have had multiple re-definitions of our goals and objectives, and finally have agreed upon developing a proof-of-concept of speaker ID technology and a user-interface optimized for Transcense's use case.

Our final deliverables to Transcense are twofold. From the technical aspect, it is a physical prototype which demonstrates SSL, while the laptop it is connected to is concurrently running the UI showing speaker identification using machine learning (plus all the code). From the UX side of things, it is the mock ups for the app which are already being used by Transcense, new mock-ups demonstrating how speaker ID features could look like in the Transcense app, and a detailed design spec showing a prioritized customer requirement list, along with all the design/tech related considerations related to each requirement.

Currently, our proof-of-concept works moderately well, 80% accuracy for 4 speakers, but there is much room for improvement. Some paths to investigate would be implementing feature selection, integrating the SSL model completely into the speaker ID system, comprehensive cross-validation for every parameter in the system, trying out a better VAD technique and implementing a more holistic strategy for prediction which depends on history of past predictions.

Since we did not have a clear objective to begin with, it is hard to evaluate the success of the project as a whole. However, this capstone project has taught me much – both technically and personally. Coming into the project, my knowledge of machine learning was minimal. Hence technically, my biggest achievement is applying theoretical machine learning concepts towards solving a practical challenge. Personally, I become a better team player and leader, and simultaneously developed close friendships with my project group. Because of these reasons, if asked if this project was a success, I would personally be inclined to say an emphatic yes!

REFERENCES

[1] World Health Organization. 2015. Deafness and hearing loss. World Health Organization. World Health Organization. http://www.who.int, accessed February 16, 2015.

[2] Kahn, Sarah. 2014. IBISWorld Industry Report OD5817: Smartphone App Developers in the US. http://www.ibis.com, accessed February 15, 2015.

[3] Diment, Dmitry. 2014. IBISWorld Industry Report 54193: Translation Services in the US. http://www.ibis.com, accessed February 10, 2015.

[4] "Becoming a CART Provider", 2015, Florida Court Reporters Association (FCRA) http://fcraonline.org/node/78, accessed April 15th 2015

[5] "General Information about Captioning and CART." CaptionMatch. http://www.captionmatch.com, accessed February 15, 2015.

[6] Horton, Mark. August 10 2010. "How 3 Billion Meetings Per Year Waste Time, Money and Productivity in the Enterprise." http://www.socialcast.com, accessed December 8, 2014.

[7] Berkeley-Haas. May 2, 2014. "16th Annual UC Berkeley Startup Competition Finals." YouTube. http://www.youtube.com, accessed February 16, 2015.

[8] "About Us". 2014. Gridspace. http://www.gridspace.com, accessed November 30, 2014.

[9] Lawler, Ryan. March 14, 2014. "Gridspace Uses Natural Language Processing To Make Your Meetings More Efficient." http://www.techcrunch.com, accessed on November 30, 2014.

[10] Porter, Michael E. "The five competitive forces that shape strategy." Harvard business review 86.1 (2008): 25-40.

[11] "Truth About Hearing Aids." Types of Hearing Aids, Sizes & Prices | Exposing Hearing Aids. http://exposinghearingaids.org, accessed February 16, 2015.

[12] "Return Paid Apps & Games." Google Play Help. http://www.support.google.com, accessed February 16, 2015.

[13] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1), 19-41.

[14] "Preserving Your Patent Rights - MIT-TLO." Accessed March 1, 2015.

http://web.mit.edu/tlo/www/community/preserving_patent_rights.html

[15] Hennecke, M. H., & Fink, G. A. (2011, May). Towards acoustic self-localization of ad hoc smartphone arrays. In Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on (pp. 127-132). IEEE.

[16] "Can You Patent a New Use of an Existing Product?" Accessed March 1, 2015.

http://howconceptual.com/patent-new-use-of-old-idea/ ,

[17] "What Are Improvement Patents and New Use Patents?" Accessed March 3, 2015.

https://smallbusiness.yahoo.com/advisor/what-are-improvement-patents-and-231516208.html

[18] "Where Speech Recognition Is Going | MIT Technology Review." Accessed March 3, 2015.

http://www.technologyreview.com/news/427793/where-speech-recognition-is-going/

[19] "How Much Does It Cost to Patent an Idea." Accessed March 1, 2015.

http://patentfile.org/howmuchdoesitcosttopatentanidea/

[20] "Do Patents Really Matter To Startups? New Data Reveals Shifting Habits | TechCrunch." Accessed March 3, 2015.

http://techcrunch.com/2012/06/21/do-patents-really-matter-to-startups-new-data-reveals-shifting-habits/

[21] "USPTO Fee Schedule | USPTO." Accessed March 1, 2015.

http://www.uspto.gov/learning-and-resources/fees-and-payment/uspto-fee-schedule

[22] "Network Effect - Wikipedia, the Free Encyclopedia." Accessed March 3, 2015. http://en.wikipedia.org/wiki/Network_effect

[23] Hillion, Hélène. "CalTranscense, Empowering People With Hearing Disabilities." (2015).

[24] Berrada, Léonard. "Design of a Speaker Identification System." (2015).

[25] Davis, S. Mermelstein, P. (1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366

[26] Khoury, E.; El Shafey, L.; Marcel, S., "Spear: An open source toolbox for speaker recognition based on Bob," Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on , vol., no., pp.1655,1659, 4-9 May 2014

[27] Practical Cryptography: MFCCs, Accessed on 3/14/15

http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[28] Python Speech Features, GitHub, Accessed on 3/15/15

https://github.com/jameslyons/python_speech_features

[29] Ronak Bajaj, Accessed on 3/14/15

http://ronakbajaj.in/speaker_id.shtml

[30] Sangwan, A.; Chiranth, M.C.; Jamadagni, H.S.; Sah, R.; Venkatesha Prasad, R.; Gaurav, V., "VAD techniques for real-time speech transmission on the Internet," High Speed Networks and Multimedia Communications 5th IEEE International Conference on , vol., no., pp.46,50, 2002

[31] Scheirer, Eric, and Malcolm Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator." Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. Vol. 2. IEEE, 1997.

[32] Tomi Kinnunen, Haizhou Li, An overview of text-independent speaker recognition: From features to supervectors, Speech Communication, Volume 52, Issue 1, January 2010, Pages 12-40

[33] Antonio De Lima Fernandes, Andrew Ho, Real Time Gender Identifier based on Voice, May 2015. https://github.com/antoniorohit/CS289A/tree/master/Final%20Project

APPENDIX A: Porter 5 Forces Analysis

As discussed in Section 2 (Transcense's Industry), Transcense has positioned itself at the intersection of three industries: services for the deaf, mobile tech and voice-based productivity services. In these industries, the threat of new entrants is high. Both barriers to entry and barriers to success are low, which makes the market potentially attractive for new entrants. Why are they low? This is because Transcense does not have any patentable technology. For now, it uses Google Speech algorithms to perform the speech-to-text recognition function. These algorithms are available to anyone, which means that it is quite easy for a new player to enter the market using the same or better technology.

Secondly, the bargaining power of suppliers is moderately high. Transcense is using different types of suppliers from Amazon Web Services for the servers to Google for the speech recognition engine. Some of these suppliers have a higher bargaining power than others. For example, Google's speech engine is the best and easiest to use on the market. This makes Transcense highly dependent on Google with no bargaining leverage.

On the other hand, the bargaining power of buyers is relatively low. As explained previously, legacy services for the deaf are expensive or not good enough: there are no good alternatives to Transcense on the market.

When it comes to the substitutes, there are no real substitutes but potential rivals may come. Skype Translator and Google Glass could easily become substitutes; Skype Translator by providing real time english-to-english captioning on a Skype conversation and Google Glass by providing the captioning on the glasses. These products are developed by more powerful players, with an established name, an efficient distribution channel and a high level of advertising. There are thus real threats to Transcense.

To conclude, there exist three critical [10] forces in Transcense's industry: the threat of new entrants, the high bargaining power of some suppliers and the threat of substitutes.

APPENDIX B: Test Protocol Document

Transcense shall create two audio files for evaluating speaker ID performance - one for training and one for testing the Speaker Identification system. This document outlines the format and details of these files.

Note: Each audio file shall be in *.wav format, and have an associated label *.txt file. The audio shall be edited and labeled using Audacity.

Training file

- The training file shall be divided into 6 sections of 4 minutes each; 24 minutes in total.
- Each section shall have 4 people (Leo, Antonio, Helene and Mathilde) speaking for about one minute in a non-specific order.
- The audio file shall not follow a pre-determined script/text.
- The 5 variables to study are:
    - A: Overlapping Speakers: i.e. 2, 3, 4 speakers at the same time; speakers join at 15 second intervals {(1-2-3-4), (2-3-4-1), (3-4-1-2), (4-1-2-3)}
    - B: Background Voices: i.e. people speaking from the other side of the room, whispering (pre-scripted)
    - C: Room acoustics: i.e. Echo vs non-echo
    - D: Microphone: i.e. Android phone (old) vs Macbook Pro (new)
    - E: Background noise: i.e. some non-specific combination of music, doors slamming, sneezing/coughing, rustling (pre-planned)
- The training file shall be constituted with one clean recording followed by 5 other recordings in which each variable shall be simulated independently.
    - Section 1: Clean Recording (echo-free room, Macbook pro microphone, no BG noise)
    - Section 2: Clean Recording + Variable A
    - Section 3: Clean Recording + Variable B

- Section 4: Clean Recording + Variable C
- Section 5: Clean Recording + Variable D
- Section 6: Clean Recording + Variable E

Test file

- The training file shall be divided into 6 sections of 4 minutes each; 24 minutes in total and use the same 4 voices as the ones from the training file.
- The variables shall be compounded in the sections as follows:
  - Section 1: Clean Recording
  - Section 2: Clean Recording + Variable A
  - Section 3: Clean Recording + Variables A, B
  - Section 4: Clean Recording + Variables A, B, C
  - Section 5: Clean Recording + Variables A, B, C, D
  - Section 6: Clean Recording + Variables A, B, C, D, E

APPENDIX C: Test Output File for Android Device

```
************************************************
                    Android
************************************************
Chunk Length: 200000
GMM for Android with 32 gaussians and All condition
GMM accuracy on training : 70.20 %
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
              BGN
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
------------------------------------------------
For speaker leo(BGN), the average score: 50 %
------------------------------------------------
------------------------------------------------
For speaker helene(BGN), the average score: 68 %
------------------------------------------------
------------------------------------------------
For speaker mathilde(BGN), the average score: 64 %
------------------------------------------------
------------------------------------------------
For speaker antonio(BGN), the average score: 88 %
------------------------------------------------
================================================
Avg Score for BGN :67.5%
================================================
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
              BGV
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
------------------------------------------------
For speaker leo(BGV), the average score: 71 %
------------------------------------------------
------------------------------------------------
For speaker helene(BGV), the average score: 70 %
------------------------------------------------
------------------------------------------------
For speaker mathilde(BGV), the average score: 77 %
------------------------------------------------
------------------------------------------------
For speaker antonio(BGV), the average score: 55 %
------------------------------------------------
================================================
Avg Score for BGV :68.25%
================================================
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
              clean
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
------------------------------------------------
For speaker leo(clean), the average score: 57 %
------------------------------------------------
------------------------------------------------
For speaker helene(clean), the average score: 82 %
------------------------------------------------
------------------------------------------------
For speaker mathilde(clean), the average score: 69 %
------------------------------------------------
------------------------------------------------
For speaker antonio(clean), the average score: 81 %
------------------------------------------------
================================================
Avg Score for clean :72.25%
================================================
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
              echo
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
------------------------------------------------
For speaker leo(echo), the average score: 59 %
------------------------------------------------
------------------------------------------------
For speaker helene(echo), the average score: 56 %
```

```
-------------------------------------------------
-------------------------------------------------
For speaker mathilde(echo), the average score: 86 %
-------------------------------------------------
-------------------------------------------------
For speaker antonio(echo), the average score: 41 %
-------------------------------------------------
=================================================
Avg Score for echo :60.5%
=================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Overall Score for Android: 67.125 %
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```