Project Summary

People who are deaf or hearing impaired have limited communication channels when there is no interpreter available. They must rely on either typing or using hand gestures that a hearing individual can understand. This at times may be frustrating and challenging to the hearing impaired individual because they are not using their natural language to communicate. Communicating with such language barrier also takes extra time than if the conversation was seamless. The proposed project tries to bridge the communication barrier by implementing an ASL to voice video chatting interface system. This project plans to use machine learning, computer vision, and video modification software to allow the targeted individuals the opportunity to have a conversation with someone who does not understand ASL, in real time. The equipment required for this project is the Microsoft Kinect sensor, video modification libraries, and ASL video libraries.

Intellectual Merit

This Small Business Innovation Research Phase I project will lie within the the recognition of gestures and machine learning. The recognition of gestures, specifically finger tracking has become increasingly popular and more researched. Many companies are integrating finger recognition into their products, such as video games and phones. As technology advances, devices will move away from the touch screen and begin to rely more on finger and voice recognition. This project will dive into tracking finger gesture libraries that already exist and will improve them by extend finger gesture technologies to recognize ASL. The translation of ASL hand gestures to text is a current problem without a published and accurate solution. Additionally, this project will involve machine learning, which is a current problem that will always have room for advancement as technology improves. Machine learning will be used as different users will have different sign styles and the software will need to learn when to accept a word and when to reject. The largest anticipated challenge will be distinguishing ASL from common hand gestures. Examples of common hand gestures is scratching a nose or moving a piece of hair from the cheek. Another anticipated problem the project may face will be the processing time and having the translations occur in real time. The more software implemented, the slower the translations. Given that voice detection software is already published, this project will integrate that technology in a way that the transitions are undetected.

Broader Impact

The broader impact of this project is to improve social norms when it comes to communicating with individuals of the deaf or hearing impaired community. Currently, the targeted audience work harder to be understood by their hearing counterparts. What is proposed has the potential to be used on many platforms besides the Microsoft Kinect. The primary demographic of this project can use this software on their mobile devices and personal computers. The outcome of this project is the first step towards bridging the gap between people with a language barrier. This program can also be expanded to work with other sign language dialects across the world and be translated into many foreign languages.

Elevator Pitch

Everyone has participated in a conversation with someone who speaks a different language than them. Many also know the feeling of being misunderstood and the frustration of not being able to clarify their thoughts. It is likely that deaf and hearing impaired individuals feel this way whenever they are communicating with a hearing person because most people do not understand American Sign Language. In today's society and business practices, the most convenient way to virtually interact with others is through videoconferencing. Without the assistance of an interpreter, telecommunication between the deaf/hearing impaired and someone who does not know sign language are limited to text messages and other text chatting systems. What if there was a video chatting interface that could break that language barrier? What if there was a way to allow the people who sign to have a conversation with someone with no signing experience in an effortless manner?

This is the idea behind EC-Chat: The ExClusive Deaf-Inclusive Video Chat Interface. With this proposed interface, the deaf/hearing impaired community will have one more outlet for communicating with individuals who do not understand sign language. This technology will have the ability to translate ASL into text and speech to text in real time. The current technology available for closed captioning is only useful for editing videos after they have been created. No other video chatting interface has the capability to translate a conversation during a live stream. Incorporating these two aspects into EC-Chat will provide a tool that will be useful for business interactions, large conferencing events, and for personal everyday use.

This project will incorporate machine learning and data mining in order to accurately translate the ASL signals. ASL is a complex language and has many classifications: facial expression, palm placement, location, stationary and non-stationary. This project will focus on stationary and basic non-stationary signs due to time constraints.

EC-Chat has three components: the ASL to text translation, the speech to text translation, and the video chatting interface. Transcribing ASL into text will involve using machine learning techniques to classify and interpret the data to convert it to American English. Translating speech to English has already been implemented on various platforms. This software plans to integrate the three components into a usable application that can be accessed online. The outcome for this project is to connect individuals who speak different languages without the complexities of having to translate one's thoughts. EC-Chat aims to handle all of the translations. ASL and American English are the main languages for this project, but this software can be expanded to support other languages in the future.

Commercial Opportunity

This product will become a feature that can be integrated into any software on the market. Video chatting interfaces will be able to use this software if the equipment requirements are met and upgraded. This would lead to the innovation of new video chatting software that will help bridge the language barrier gaps between not only American Sign Language and English, but other sign languages and spoken languages. As communication channels within businesses escalate to a new global level, the difficulties of finding bilingual translators has become a large problem in many companies. The time and money that is spent on bridging communication gaps is costly and the future of EC-Chat would solve many of those problems and cut costs dramatically.

Economically, this software could get pricy for installation, given the requirement for a finger-tracking sensor built in. However, the development of future versions of this project could lead to new software and hardware that minimizes costs. Companies that select to purchase and install this software will not need to provide each employee with the equipment, just enough to meet the company's needs. Overall, the one time cost of the hardware and software could end up saving companies a lot of money due to the increase in productivity it could provide.

Customers could be anyone from around the world. While it targets companies at the industry level, individuals may greatly benefit from using this product as it will simplify and expedite communication between the hearing impaired and hearing people. EC-Chat’s capabilities can unify communities within social groups and workplaces. Even though its original functionality is to provide a medium of teleconferencing between American Sign Language and English, it will be extended to other languages and can be used as a standalone tool to translate sign language to another spoken language between two users, even if they are both on the same side. It could be applied heavily in the customer service field. Customer service agents or retail associates globally could utilize this software to help serve the hearing impaired community. EC-Chat’s basic business model is that EC-Chat will help global communication between the hearing impaired and hearing.

Currently, there is no competition on the market for our product. There have been prototypes made by different groups utilizing the Kinect that has a similar outcome to our product, however there is no product on the market, or rumors of one emerging. Similar to when any new product arrives on the market, when EC-Chat enters the market, pending its success rate, other companies will likely develop their own version and deploy. Due to EC-Chat’s dependency on the Kinect, there could be legal issues and patenting rights that may arise, or Microsoft could work to quickly deploy their own version.

The largest risk falls within the reliance on existing software. EC-Chat could be an application similar to many companies that publicly states its integration of open source libraries such as the open Kinect, speech to text and finger-tracking. Pending the success of the EC-Chat software, EC-Chat could expand to develop its own hardware, which it would sell alongside of its software. Additional risks with the innovation of EC-Chat could be the evolvement and expansion of the software to brand new products that lie out of EC-Chat's original goals.

The commercialization approach to this product will be difficult. It will require checking to make sure EC-Chat's documentation references all of the open source libraries it utilizes correctly, especially ones geared towards the Microsoft Kinect. The first step after having a polished product would be to brand it and get its name out. EC-Chat is targeted for everyone, and therefore it could be publicized in many places, especially in the deaf community. After it is publicized, a prototype could be rolled-out in order to gage the industry's reaction. The product and company could quickly expand as more people begin to realize the benefits from EC-Chat.

Aside from the aforementioned economic benefits, the benefits for communities and individuals who are, or who interact with hearing impaired people, will be priceless with the effect of this software. Monetary benefits to communities and individuals could be the increased participation or productivity which could raise or save money for them. EC-Chat itself will not be expensive, it will be the hardware that ends up being the most expensive part. EC-Chat likely will have minimal revenue at the beginning, but once the product is being has gained traction, more employees could be hired, and the company could expand capabilities of this product and grow, increasing in size and stock value. Assuming that the EC-Chat software is sold separately from the Microsoft Kinect, each EC-Chat installation would be fifteen dollars a month for premium features. After EC-Chat gains success, it could expand and provide more services, offering more opportunities for revenue. The only estimate that can be made now is that EC-Chat wants to break even financially. To break even, EC-Chat will have to bring in enough revenue to cover the costs for maintenance and equipment updates that may be needed for future versions.

Social and Global Impact

The goal of this project is to allow two individuals who speak different languages the opportunity to converse with one another. More specifically, this project is intended to bridge together the deaf and hearing-impaired community with the hearing population. People avoid interactions with others who speak different languages because it is hard to communicate one's thoughts, it is time consuming, and both parties do not want to make the other person feel frustrated or judged. In the past, people who were hard of hearing were ridiculed, degraded, and insulted because they were viewed as being different. As society is becoming more inclusive, the population has become more open-minded and conscientious about dealing with language barriers; however, there is still work that can be done. EC-Chat can be advance the solutions to this problem.

This commercial opportunity can be used in a wide range of areas, from businesses, conferences, and for personal use. In an effort to improve work-life balance, many companies allow for their employers to work from home. Telecommunication is key in this particular business model. Having a tool that translates sign language into text can allow for hearing impaired people to work from home. Companies could save a lot of time and money by using this product. Instead of hiring multiple interpreters to take turns translating, EC-Chat can do the translations for long periods of time. This will be especially beneficial in large conferences or workshops. EC-Chat can give the person who is hard of hearing the courage to speak out and become more involved in the conversation. This will hold true in their personal lives as well. Interacting with people who are not hard of hearing, in a less frustrating manner, allows for relationships to grow between the deaf and hearing communities.

While EC-Chat aims to create another means of communication for people who sign, there are some environmental issues to consider. First, the current implementation requires the use of a Microsoft Kinect. For businesses, this would be feasible for a company to provide reasonable accommodations for their employers. It would not be worthwhile for people to purchase this equipment for personal use. For workshops, EC-Chat would best be used on mobile devices. Second, the design of this product has to take into account that people who sign have distinct mannerisms just as hearing individuals can have accents and slang that is specific to a region. EC-Chat will be designed for American Sign Language. It will not work for signs that small groups create amongst each other.

Another thing to consider are the external factors that affect the accuracy of the translations. If EC-Chat were to be compatible on mobile devices, the camera quality, the size of the mobile device, and the location can influence the results. There are certain camera qualities to consider such as the resolution, the filters on the camera, and the focus/positioning of the camera. Currently, EC-Chat is not responsive, which means that the graphical user interface will not adapt to the varying sizes of mobile devices. Lastly, the location of where the user is can affect the video stream. The room could be poorly lit, or the natural light from the windows could interfere with the camera.

As with any product on the market, EC-Chat is vulnerable to being used in unethical situations. Since some of the software incorporated into this program will be extensions of open source implementations, the application can skew the way this program interprets signs. For example, if someone signed a gesture that the software deems inappropriate, the software may not interpret or relay that information. Mishandling the data, could be a potential issue on the other end of EC-Chat as well, where the speech recognition software may not display profane language. This is an issue that could arise in any situation. Interpreters ultimately decide what is or is not relevant to translate. The priority of EC-Chat is to make sure that the software is accurately interpreting the signs.

EC-Chat will also be susceptible to hacks. For instance, a hacker could program the Kinect to turn on, at any moment, to record and store photos. They could also reprogram this product to misinterpret the data. At this stage in development, there is no security measure to prevent these attacks from happening. Providing security measures will have to require collaboration with security companies, so that they can provide the necessary tools for this application.

The global impact that can be achieved through EC-Chat is immeasurable. There are over 100 different dialects of sign languages, such as British Sign Language and Mexican Sign Language. Developing countries may have a greater divide between the hearing and the hearing-impaired groups. EC-Chat could be extended to support these different dialects. This would allow for more resources to be brought to hard of hearing individuals in developing countries. Furthermore, EC-Chat could potentially support video conversations between two individuals who speak different sign languages.

Technical Discussion

As mentioned earlier, the largest technical challenge will be selling a product that has such strict hardware requirements. EC-Chat will require a Microsoft Kinect sensor and specific software installations. This could lead to licensing disagreements with Microsoft and other patent issues when EC-Chat is trademarked and marketed. EC-Chat could grow and have the funds later to build their own hardware to sell themselves, making the created hardware more compatible with the EC-Chat software and desired functions, especially as EC-Chat could evolve to become a basic translation device.

Another innovation issues that EC-Chat could have in the arising future is integrability or creation of hardware that does not require an external hardware. Even if EC-Chat creates their own external hardware, all individual users will either need to purchase it to use in conjunction with their computer, or go to a computer set up with the appropriate hardware. The optimal solution would be for EC-Chat to one day be able to utilize general cameras to get the required data needed to interpret. However this would require long term innovation in basic cameras that are included on personal computers and laptops. That is not foreseeable in the near future, so EC-Chat will continue to rely on their customers purchasing hardware alongside of the software.

During the Phase I research, the biggest accomplishment was determining the minimum hardware requirement EC-Chat had, given the existing hardware that is on the market. The Microsoft Kinect was the optimal camera given the required ability get depth, joint and skeletal positioning of the entire body, not just the hand or fingers. This affects the commercial feasibility as it answers the question to what all EC-Chat users must have in order to run EC-Chat's software successfully. The price of the Kinect Sensor could be the limiting factor to potential customers selecting EC-Chat in the future.

The critical milestones that must be met to get the product to the market would be a larger ASL library built with more accurate and fast translation. The more accurate the translation is, the more benefit customers can gain. If the translation only catches basic phrases and is slow, then many customers will not have their desired needs met, and will either write bad reviews or may hesitate on purchasing to begin with.

A long term technical advancement would be, as stated earlier, the creation of EC-Chat's own sensor that would be more compatible, affordable, and portable for users. It would have a faster initial setup, and be licensed to EC-Chat itself.

R&D Plan

Timeline

Research is already being conducted with the different modules of this project: converting ASL to text, creating an ASL library, converting speech to text, and creating a video chat interface that encapsulates the other components. The development team started the initial groundwork in August 2017. This involves obtaining the necessary hardware and software to complete EC-Chat. The projected launch of EC-Chat is April 2018. Each component requires extensive research and experimentation before any implementation can be done. The proposed plan is to spend the first few months gathering requirements and assessing the importance of each module. Since the makeup of EC-Chat will be mostly of open-sourced software, the months following the requirements gathering phase will be the experiment stage where the software engineers will analyze and investigate how the existing code works. This stage includes manipulating and testing the code. The next stage will be to achieve the first level of each module, which will be discussed below. Afterwards, the next phase will be to add code to satisfy the outlined requirements. Once that is complete, the team will do unit testing. After unit testing, the team will integrate the models and test their functionality.

The first few months were for deciding the best software and hardware to use. It was decided that the Microsoft Kinect Camera were to be used to give coordinate data such as x, y, and depth. Using a Windows 10 computer will prove to be the most versatile, as the Kinect was designed to be compatible with Windows. For the speech to text aspect, Google Cloud Speech will provide the most accuracy and compatibility. Scikit was chosen because of its capability to interpret skeleton models.

Understanding how the Kinect functions took about two months. The first part was dedicated to investigating the different operating systems to use with the device. The latter part of the month was spent on learning how to use the Kinect on a Windows 10 machine. Most of the Kinect's software is written in C#, so the team must become familiar with Visual Studio. The main objective for this part of the ASL to Text Component is to obtain an image stream and the data, so that the data can be manipulated and used with SciKit Learn.

The timeline for the machine learning framework is dependent upon the timeline for understanding the Microsoft Kinect. Fake data can be used for the research and investigation phase, but in order to make strides towards EC-Chat, real data needs to be used for classifying the data. This is expected to be take two months to complete. Getting real data will take time because it involves having multiple people perform the same signs. This is so that SciKit Learn can predict the signs more accurately.

While working simultaneously with the Microsoft Kinect, the team is currently researching and understanding the Google Cloud Speech API. It is projected that this phase will take approximately two weeks before the API will be fully installed and working. Afterwards it is projected that it would take a month to have a starter program that satisfies the needs of EC-Chat. Once this is complete, the component can be integrated into the video chat interface, so that when a person speaks, the words will be transposed to the screen. This will be done concurrently with the work being done with the machine learning tool.

The remaining time before deployment will be used to integrate the three components into one working product. The initial step is to have the video chat interface running on its own. The allotted time for this is estimated at two weeks. Then, the team will integrate the speech to text aspect into the interface. That part is expected to take about two weeks which will include testing. The final module will be added, and this part is expected to take longer because of all the smaller components it entails. Once all of the parts are integrated, the final step is testing and maintaining the software. After deployment, the team will make improvements to the codebase as necessary.

Levels

ASL-to-Text

The first level of the ASL to text component is to get the Kinect to run successfully and begin programming it to collect data. The data it collects must be representative of hand gestures, placement and follow a series of movements. It will need to consider placement and depth and will be represented by three points, an x coordinate, a y coordinate and a depth measurement. As the Kinect is described more in depth later, understanding and fetching usable data from the Kinect is the first step to the ASL to text component of this project. The second step is to train the data gathered from the kinect via machine learning. The data that is gathered will be utilized to recognize signs in real time to estimating what sign the user is representing, based on the current library and what the sign most likely is. This will be done by using a support vector machine that will gather more data and interpret signs based on best fit.

The second level to the ASL to text component is the machine learning library. The machine learning library that this project will interpret the incoming data from the Kinect, and translate to the sign that is best fit. The machine learning component will allow the data to be continuously trained as more data is inputted into the library. The software will take each sign, interpret, and add it to the existing library to add more plots to the support vector machine. As more data is added, the accuracy and precision will naturally improve. There will be an initial ASL library set for the beta product, that is composed by the product makers. However, this ASL library set will be very basic and EC-Chat software will need more data to get smarter in order to do better translation.

The EC-Chat initial library will be composed by gathering volunteers to sign a specific set of signs and the sign’s two dimensional plotting data (x coordinate, y coordinate, and depth) and english translation will be recorded. The data will be trained as each English translation will have a set of known ASL plotting data that corresponds to that word or phrase will be collected. As more ASL plotting data is collected for each English phrase, the ranges of two dimensional plotting data will be slowly established and able to distinguish between different signs and English words. The data will be trained to look at an incoming set of data and see which category it fits best in, and grabbing the english translation in turn to output.

The third level of the ASL to text component is testing the initial pipeline and translation, without all of the library built. The initial ASL to text component and trained data library will be tested with simple sign recognition of the hand being extended to the right or left. The software must be able to decipher if the user is signing right or left. The output must be successfully displayed via text to the other user on the opposite side of the video interface. This will be the pipeline experiment for the EC-Chat beta software because it will show that all of the parts have the basic functionality working and the integration has been successful.

There will be initial experimentation through the process as the ASL to text is being constructed. The machine learning component will be able to estimate ASL signs and find the best fit, and output the corresponding English translation. It will be built up from scratch, with an initial ten people per ASL sign to catch discrepancies. An initial inventory of thirty signs, each with ten people signing will be built for the beta version of the EC-Chat software. The signs will be tested each individually by testing whether another individual can successfully sign one of the thirty known words and have it recognized successfully.

Speech-To-Text

Level one of this module is to have the Google Cloud Speech API installed and working functionally on the computer. This API is very extensive and it requires creating an account and having an authentication key. The next level is to use the codebase within Visual Studio. Having the text-to-speech working within this software will allow for seamless integration into the video chatting Interface. For it to work, some dependencies will need to be installed so that Visual Studio can import the necessary namespaces. After level two, the next phase is to test and manipulate existing code. This involves making sure that the code compiles. For EC-Chat to work, the code needs to be manipulated so that the program will start and stop recording. This would be a challenge as the code as is listens for an allotted amount of time.

Video Interface and Integration

For this module, the first step is to create the video interface using Visual Studio. When this is complete, networking needs to be incorporated into the interface so that the user can interface with someone on another computer. The next level is to add the speech module into the video interface. The speech to text module will write the translated speech to a text file. The video interface will read from that text file and transpose it to the screen. The final level is to add the ASL-to-text into the interface. This part will work similarly as the speech to text where it would write to the same text file, and the interface will read from it to display the words on the screen.