TranslatAble: Giving Individuals with Complex Communication Needs a Voice through Speech and Gesture Recognition

Meredith Moore
Center for Cognitive Ubiquitous Computing
Arizona State University
Tempe, AZ, USA
Mkmoore7@asu.edu

Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing
Arizona State University
Tempe, AZ, USA
panch@asu.edu

ABSTRACT
This paper presents the design of TranslatAble, a new Augmentative and Alternative Communication system. Taking advantage of the convenience and mobility of modern smartphones, TranslatAble allows individuals with complex communication needs—specifically individuals with severe dysarthria or individuals who primarily use non-sign language gestures—to build their own system of communication through speech or gesture input. This system uses a machine learning model—specifically dynamic time warping—to match the user's input to their unique dictionary to help them be understood by their communication partner. TranslatAble is highly flexible and person-centered, which allows it to adapt to fit the unique needs of each user's unique communication needs. We expect this novel device to enable individuals with complex communication needs to interact with a larger social circle. This may help decrease feelings of loneliness and increase self-confidence, self-esteem, and independence, as well as help maintain strong relationships outside of the user's support network.

Keywords
Augmentative and Alternative Communication; AAC Device; Gesture Recognition; Speech Recognition; Complex Communication Needs

1. INTRODUCTION
Augmentative and Alternative Communication (AAC) professionals generally define communication as the ability for a person to establish meaning with another person or group[1]. This joint establishment of meaning is fundamental to most aspects of life. Individuals who have complex communication requirements may not have the communication skills necessary to meet all of their needs. Namely, an individual's ability to communicate effectively affects their ability to build and maintain relationships, make choices, and participate in everyday life. These components are all incredibly important for quality of life.

There are two main use cases that we will focus on for the purposes of this paper. In the first use case, the user is an individual who does not have speech, and communicates primarily through gestures. There have been a number of studies that attempt to solve the problem described in case 1, but this body of work is generally focused solely on sign language translation. These solutions are divided by the type of data they use to classify the signs, but some of the common methods are to use sensor gloves or computer vision to classify hand gestures[2,4,5]. While data gloves have been used to achieve promising gesture recognition results, they have also shown to be too invasive and expensive for the everyday user. The main drawback of computer vision methods is that they prevent the user from being mobile. While both computer vision and sensor gloves may eventually work well to translate sign language, there also exists a group of people who may not have the cognitive ability to learn sign language. However, this group still uses concise and repeatable gestures to communicate. The only translation option available to these individuals currently is to draw a line down the middle of a notebook and describe the gesture on one side, and the translation on the other. While this rudimentary solution has helped many people communicate, the technology exists today to create a system that will recognize these gestures in real time to translate these individual's unique communication systems to their interaction partners.

In the second use case, the user is an individual who has speech, but their speech may be difficult for communication partners to understand. This area is a very underdeveloped field of research, but there is one app, Talkitt, that has yet to be released, which has a similar goal of giving individuals with speech difficulties a voice of their own by translating their speech to a more easily understood computer generated voice.

ASSETS'16, October 24–26, 2016, Reno, NV, United States.

DOI: http://dx.doi.org/10.1145/12345.67890

2. PROPOSED SOLUTION
To fill this gap in user needs and provide a pervasive and ubiquitous computing experience for individuals with complex communication needs, we developed a novel smartphone-based AAC device that translates both gestures and speech in real time.

2.1 Goals
We have three goals for the design of the application:

Pervasive/Non-Invasive Interface: The solution should not interfere with the user's daily activities.

Real-Time Accurate Translation: Communication is a real-time activity which necessitates fast response times. A timing delay of 0.1 s is considered 'instantaneous,' while a delay of 0.2 s is noticeable by the user, but still acceptable. Any response time greater than 1 s needs to be acknowledged as processing[3].

Facilitate Communication: The device should enable the user to easily facilitate communications with people around them.

2.2 System Architecture

Input for TranslatAble comes from two main sources: the built in iPhone sensors and the Myo Armband (Fig 1C). Audio input is recorded from the built in microphone in the iPhone (Fig 1D), while both inertial data (acceleration and rotational data) and electromyography (EMG) data come from the commercial grade Myo Armband. Myo Armband has 8 EMG pods that record at a rate of 200 Hz, as well as an accelerometer and gyroscope that record at 50 Hz.

The Myo Armband is connected to the iPhone via a Bluetooth LE connection and is worn on the forearm. These features keep it from interfering with the user's daily activities.

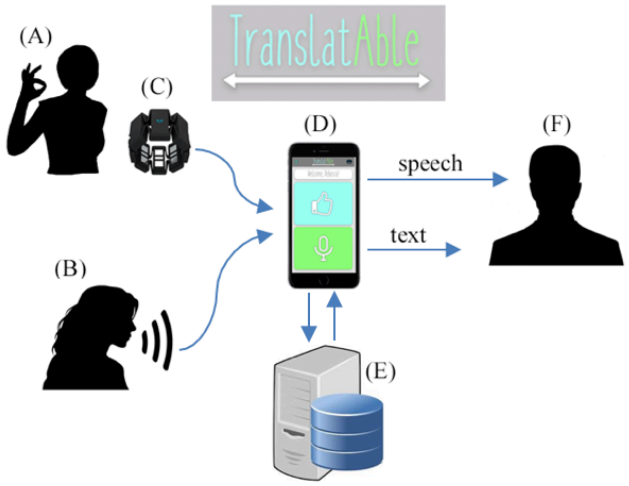

Figure 1: Overview of the TranslatAble System

2.3 Usage

There are two main ways to use TranslatAble, through gesture (Fig 1A) and through speech (Fig 1B). For both methods of communication, the process is very similar. For gestural communication systems, the user begins by building their unique library of gestures. The user individually adds the gestures into their library by performing the gesture 3-4 times. This repetition allows the system to obtain a template of the gesture. They then enter into translation mode, and perform the gesture. The data is sent to the phone (Fig 1D), which then processes the data and returns the recognized output to the user's screen. The user has the option to configure the output as either text or speech. For speech based communication systems, the process is very similar, but instead of performing gestures, the user speaks into the device for both the training portion and the translation portion.

2.4 Data Processing

The primary task of the system is to match gestures or audio input to the user-created templates. Once the input is collected, the data is sent to a remote database to be processed and matched against the other gestures or audio clips that are part of the user's profile. The input data is preprocessed, and the task is to recognize the gesture or audio based on the nearest matches between the input samples (received during translation mode), and stored samples (collected during training to build a user's library). To find these nearest matches, we are using Dynamic Time Warping (DTW), a well-established method for finding the optimal alignment between two given time-dependent sequences. DTW is performed on each modality of the input and a list of potential matches is created. From this list, the nearest match is chosen.

2.5 Output

There are two main output modalities that are being considered at this time: text and speech. The output modality will be an option that the user can control. Myo is also equipped with a vibration motor, and haptic feedback will be available to cue the user when Myo is ready to translate, as well as provide feedback when the recognition is successful.

3. FUTURE STUDIES

Thus far, we have built a prototype of the user interface, and have received some informal feedback on the usability of the interface. We are in the process of implementing the gesture and speech recognition functionalities of the application using Dynamic Time Warping. We are also in the process of creating a large dataset of gestures using the Myo armband in order to test more robust machine learning algorithms such as deep learning. While the prototype is being developed, we are setting up focus groups with speech and hearing clinicians, as well as individuals with complex communication needs and their support networks. The feedback we receive from these focus groups will further shape the design and functionality of the application. We are also planning user studies that will focus on discovering how well TranslatAble facilitates communication for individuals with complex communication needs.

4. ACKNOWLEDGMENTS

The authors would like to thank the National Science Foundation and Arizona State University for their funding support. This material is partially based upon work supported by the National Science Foundation under Grant No. 1069125.

5. REFERENCES

[1] Blackstone, S.W., Williams, M.B., and Wilkins, D.P. Key principles underlying research and practice in AAC. *Augmentative and Alternative Communication* 23, 3, (2007) 191-203.

[2] Kim, K.W., Lee, M.S., Soon, B.R., Ryu, M.H., and Kim, J.N., Recognition of sign language with an inertial sensor-based data glove. *Technology and Health Care* 24, s1 (2015), S223-S230.

[3] Miller, R.B. Response Time in Man-computer Conversational Transactions. *Proceedings of the December 9-11, 1968, Fall Joint computer conference, Part I* ACM (1968), 267-277.

[4] Praveen, N., Karanth, N., and Megha, M.X. Sign language interpreter using a smart glove. *2014 International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, (2014), 1-5.

[5] Wu, Y. and Huang, T.S. Vision-Based Gesture Recognition: A Review. *Gesture-Based Communication in Human-Computer Interaction.* Springer Berlin Heidelberg, 1999, 103-115