Hand Gesture Recognition and Speech Synthesis Data Glove for Children with Non-Verbal Disabilities

Hadeel Ayoub
Department of Computing,
Goldsmiths, University of London
London, United Kingdom
hayou001@gold.ac.uk

Mick Grierson
Creative Computing Institute,
University of the Arts, London
London, United Kingdom
m.grierson@arts.ac.uk

Ed Hill
Software Development,
BrightSign Technology
London, United Kingdom
ed@brightsignglove.com

ABSTRACT

In this study we use a wireless and standalone data glove to track hand shapes, orientation, position and dynamic movement of children with non-verbal disabilities for the purpose of translating custom sign language hand gestures to speech. We apply recorded sensor data to train a personal K-Nearest Neighbours classifier using Dynamic Time Warping (DTW) for each user.

We evaluate the accuracy of personal classifiers when trained with data collected through supervised training sessions conducted at special schools with non-verbal students. We compare that with a general classifier trained by the group data.

This study is part of a programme to develop and produce an affordable hardware technology solution to provide machine translation services from custom hand gestures to written languages, and then by extension to spoken languages in the form of audio output. The primary purpose is to facilitate daily communications between individuals with speech disabilities and the general public.

CCS CONCEPTS

• **Social and professional topics** → **People with disabilities**; **Children**; • **Human-centered computing** → **Gestural input**; • **Computing methodologies** → *Natural language generation*; • **Hardware** → *Sound-based input / output*; Wireless devices.

KEYWORDS

Inclusive child-centered design, enabling technology, gestural input, haptic interaction, auditory and speech interfaces; wearable technology; assistive technology, prototyping; accessible design; healthcare innovation

ACM Reference Format:
Hadeel Ayoub, Mick Grierson, and Ed Hill. 2020. Hand Gesture Recognition and Speech Synthesis Data Glove for Children with Non-Verbal Disabilities. In *IDC 2020: Designing for the Future, June 21–24, 2020, London, UK*. ACM, New York, NY, USA, 7 pages. https://doi.org/https://doi.org/10.1145/3313831.XXXXXXX

1 INTRODUCTION

A data glove is a human-computer interface with certain tactile or other sensory units that are attached to the fingers or joints of the glove, worn by the user. Tactile switches and resistive or capacitive bend/stretch sensors which measure the bending of different joints, offer measurements as to determine if a hand is open or closed and some finger joints are straight or bent. These results are mapped to unique gestures and are interpreted by a computer. The advantage of such a simple device is that there is no requirement for any kind of preprocessing. With very limited computing processing power in the 1990s, such systems showed great promise despite the limited maneuverability due to the need for wired tethers that connected the glove to the computer [5]. While many modern data gloves still use wired connections to a computer, it is now possible to transmit the sensor data wirelessly, negating the need for a cumbersome physical connection.

There are a number of problems with using such devices for signing. One of the major difficulties in accurate recognition of hand gestures is the enormity of many sign language vocabularies. Many feature extraction methods rely on searching for matches within these large vocabulary databases [5].

Even with advancements in computer vision, glove-based sign language recognition offers the widest vocabulary and the best possible recognition accuracy. However, no such recent systems have been reported with sufficiently high accuracy to be considered for commercialisation, possibly

IDC 2020, June 21–24, 2020, London, UK

ACM ISBN 978-1-4503-6708-0/20/04. . . $15.00
https://doi.org/https://doi.org/10.1145/3313831.XXXXXXX

because researchers are more focused on camera-based systems. There are many versions of data gloves that translate sign language to text and/or speech. Most of these gloves rely on a smart device for output and perhaps none have moved beyond prototyping. There is almost no published work showing evidence of sign language data gloves being tested by speech-disabled participants for daily communication, let alone children. This is possibly due to the complex programming and hardware required as well as the ethical considerations and other challenges encountered when working with such vulnerable groups.

2 RELATED WORK

One of the earliest systems to convert gestures to speech was demonstrated by Fels and Hinton [2], who produced a data glove-based system called Glove-Talk. They used a VPL Data-Glove in 1992 to convert hand gestures to speech via the DECtalk speech synthesizer. Their Glove-Talk vocabulary consisted of 66 root words, each with up to six different endings. The total size of the vocabulary was 203 words. Most of these hand shapes represent the ASL alphabet. They also utilized orientation differences and the varied hand shapes for semantically opposite words such as 'come' and 'go' which have a 180 degree orientation difference. Various endings for words were formed through different hand movements.

In 2011, Oz and Leu developed an American Sign Language (ASL) recognition system based on the Cyberglove™ sensor glove and artificial neural network (ANN) classifiers to translate ASL words into English [4]. The system consisted of a sensory glove and an electromagnetic motion tracker. They trained the ANN model for 50 ASL words with a different number of samples for every word. The final output of the system consisted of audio of recognized ASL words generated by a speech synthesizer. The classification results achieved 90% accuracy which demonstrated that the system could be used successfully for isolated word recognition.

A more recent study by a group of researchers in 2016 [3] demonstrated the additional usage of a sliding window to translate a pre-trained list of sign language hand gestures. Their system achieved continuous recognition of ASL signs using a glove, in real time, with an accuracy level of 98%.

Although these models were technically very successful, none progressed beyond the research phase and no plans were made to go into production. Furthermore, no personalisation of signs was offered on a per-user level to accommodate differences in sign language libraries or motor abilities. This motivated us to introduce a customisable, wireless data glove to the assistive technology space and give a chance to speech disabled individuals, specifically children, to try it and use it both in a controlled environment such as an educational setting as well as at home. This allowed us to better understand the challenges of producing a more refined wearable system that could be used for daily communication.

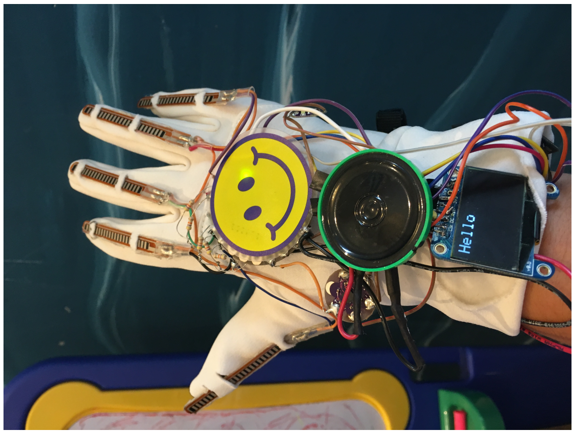

Figure 1: Early glove prototype, produced in 2016

In 2016, we conducted a brief usability study to test an early prototype (as in Figure 1) of our proposed system to translate hand gestures to speech. At the time our system was preprogrammed with hard-coded sensor data to identify a limited vocabulary of ten signs based on static, dominant-hand gestures using an accelerometer data glove. We tested the glove with 2 boys who had non-verbal autism and used the Makaton [8] sign language to communicate. The testing demonstrated that the system was capable of translating sign language to text and speech with an accuracy rate of 80-85% and with about 15% of attempts resulting in no words being spoken, due to a failure to detect the gesture performed. Reasons for failed attempts included the fact that signs varied in timing and speed, even with the same user, particularly where slight changes of hand position occurred [5]. We applied the feedback from the children who participated in the study to redesign the hardware and software. In this paper we present a new approach to solving those issues by using a more child-friendly wearable device which implements dynamic time warping in order to build an on-body device for custom hand signal translation.

3 SYSTEM

Hardware

The hardware for the glove consisted of three primary units; a micro-controller (a Raspberry Pi Zero W [9]), a custom circuit board featuring the speaker and display, and a hand-shaped, flexible PCB that contained the various sensors used. The board and Pi were soldered together on top of one another, with the flexible PCB connected to that main assembly via a short ribbon cable (see Figure 2).

The flexible PCB featured 5 flex sensors, one per finger, 1 accelerometer and 1 gyroscope. The accelerometer and

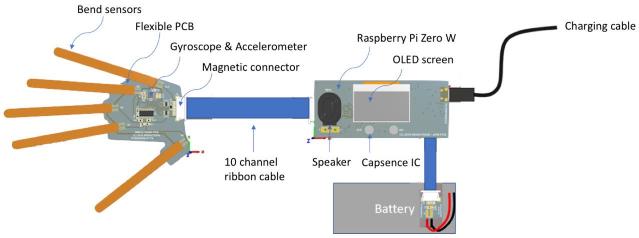

Figure 2: Glove hardware schematic

gyroscope were on the same physical integrated circuit in the centre of the back of the hand. The values from the flex sensors were transmitted to the Pi in raw form, with the accelerometer and gyroscope unit connected via an I2C bus.

Software

The training and classification system was largely written in Python, due to its convenience, and was based on a K-Nearest Neighbours classifier using Dynamic Time Warping (DTW) to calculate a distance metric between time series recordings.

Dynamic Time Warping is an algorithm that enables the calculation of similarity between temporal sequences while allowing for variations in speed and position in time. In our case, this means that differences in the speed of signs being performed, and small variations in the delay when recording, prior to the user performing each sign, will be ignored.

An example of this can be seen in Figure 3. Recordings of the value of one axis of the accelerometer on the glove were taken as a user performed the sign for 'Please' multiple times at different speeds. This value is graphed over time (with the raw value of the sensor on the vertical axis, and time on the horizontal axis) in the below chart in the form of two dark blue line plots. The orange lines represent the mapping between the points in the two series found by the DTW algorithm. In this example, the algorithm has largely correctly identified the mapping between the macroscopic features of each series (namely the two consecutive spikes towards the start of the time series, followed by a slow decrease in value).

Implentations of the DTW algorithm that guarantee the optimal match have at best quadratic ($O(n^2)$) time complexity of computation. As such, alternative algorithms that will find good approximations of the optimum, such as FastDTW [6] and SparseDTW [1], can be used instead, some of which run in linear ($O(n)$) time. The FastDTW algorithm was selected to be used in our system.

When DTW is used in situations where time-series data is multi-dimensional for each frame, two different variants are possible [7] - the dependent variant (DTW_D) and the independent variant (DTW_I). DTW_I calculates a separate match for each dimension of the data, whereas DTW_D finds a single shared optimal match between points for all dimensions simultaneously.

When the children trained the glove, recordings for all sensors were stored for each sign, with multiple examples of each sign being recorded. These recordings were labelled with the name of their corresponding sign for example 'Please' or 'Thank You'.

Children selected the label for the sign for which they wished to record a new sample, while wearing the glove, using the glove's on-screen display and pressed a button to start the recording of sensor data. They then performed the sign, before pressing the button again to cease recording.

Our system, after receiving the raw data from the sensors embedded in the glove, normalised them between 0 and 1, based on values we recorded from each sensor as the maximum and minimum possible readings during normal use. This prevented any sensor from overly weighting the classification result. The sensor values were stored after normalisation.

After pressing another button to begin recording a sign for classification, each child again performed the sign, and pressed the same button to cease recording. As during training, sensor values were immediately normalised to provide values between 0 and 1.

The DTW algorithm was then applied to each of the pre-recorded training samples in turn, with the new recording for classification. This provided the distance between each sample (and therefore its label) and the new recorded gesture.

A K-nearest-neighbours algorithm was then used to select the output audio and text, based on the distances calculated in the previous step. K was set to different values to test its impact on classification accuracy.

The label for the classified sign was ultimately displayed on the screen, with corresponding audio being output from the on-board speaker.

Design & Enclosure

To ensure the safety of the children wearing the technology and to comply with the ethical approval, all sensors were embedded within an inner lining of each glove. Insulating the sensors was a necessary design and safety solution to prevent direct contact with the children's skin and to make the glove appearance discreet for participants who did not wish to wear an obvious assistive technology device. The children were invited to communicate their desired glove designs. Each child received a right or left hand glove (corresponding to their dominant hand when signing) that was personalised to their preference, in terms of size, colour, and design. We show a plain glove (Figure 4) to preserve anonymity, as gloves also had the children's initials and/or names embroidered. We found that involving the children in designing their own

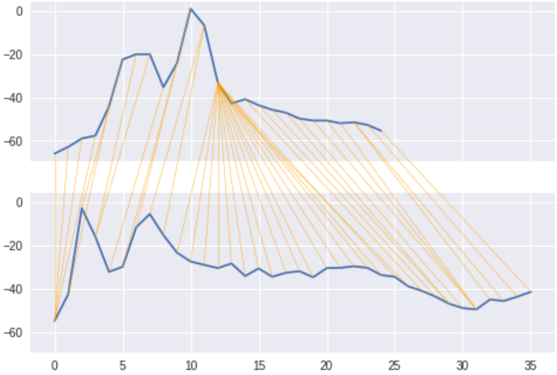

Figure 3: Example of Dynamic Time Warping to measure distance between two time series for the hand sign for 'Please'.

gloves helped in overcoming their initial intimidation by the technology that we had faced in the previous study, where the gloves were more obtrusive and the design was unified across participants.

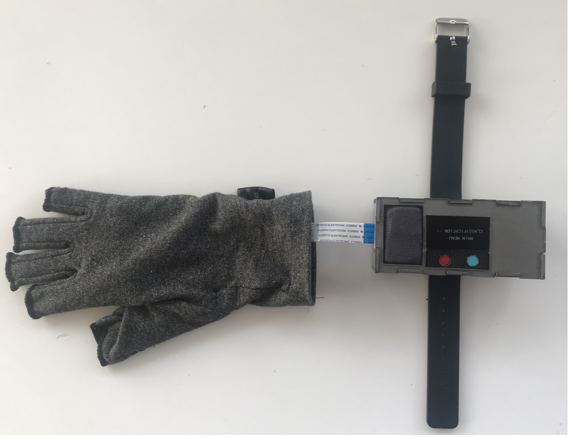

Figure 4: The assembled glove, flexible PCB and hard enclosure

A hard-case wrist band was designed to house the microcontroller, a custom circuit board featuring the speaker, screen and two buttons, and a battery (Figure 4). The case was sealed to ensure children were not able to access any of the electronic components. The two buttons were added to allow children to interact with the glove and use it for sign language training and translation, a red button for training (record a new gesture) and a blue button for translation (recognise a gesture and output the corresponding audio). The battery was charged using a USB port without opening the case. An automatic fail-safe switch was added to disable the operation of the glove while charging as an additional safety feature.

4 METHOD

Our system was programmed to record dynamic gestures' sensor data received from a right or left handed glove. In signs using both hands, only the participant's dominant hand was used for training. This was effective because in the majority of signs using both hands, either both hands are the same or one hand stays motionless in holding one position, while the other hand makes the sign. The data glove screen showed a list of words with a user interface menu for the user to scroll through them. Ten words were selected by the children from a list of 50 most used words in school, provided to us by attending teachers. Each word (label) corresponded to a notional gesture which initially had no data recorded for it. Gesture data was recorded by the children during training sessions (described below). We chose this method of a predefined dictionary of words although our software supports the definition of personalised labels. This was due to school regulations and the granted ethics clearance which did not allow our technology to connect to the internet, in order to protect children's data and to ensure none of the testing data was sent to the cloud or stored on any external servers.

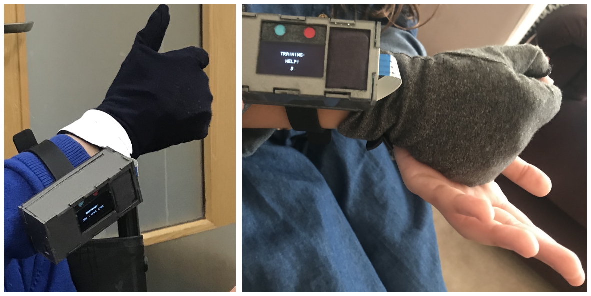

Figure 5: Two examples of participants performing signs

We recruited ten non-verbal participants, between the ages of 5 and 15 years old. Teachers identified the students who would be good candidates for the study. Selection criteria was based on familiarity with a form of sign language, consistency in signing and those who could benefit from using this technology to overcome communication challenges in school. A preliminary meeting was held at participating schools with children's parents and teachers to introduce the technology and describe all features. A usability guide was distributed to ensure adults who supervised children using the gloves, in school and at home, were aware of the safety regulations. Training sessions consisted of a two hour long task (described below) and were broken into four, fifteen minute segments. Training sessions were done with the researcher and the participant's speech therapist in attendance. The first segment was reserved for getting the children familiar with the glove and asking them if they wanted to wear it. Once they gave consent we helped them with putting the glove on and demonstrated how to use it. The child was always the one who pressed the buttons while wearing the glove.

The glove has two modes: *Training* and *Translating*. Participants were first shown how to use the glove to record signs (Figure 5). Each participant recorded up to 10 sign samples for each word by pressing the record button before and after making the hand gesture. This was necessary to train a personal classifier for each user. Gesture data was captured at 20 frames per second. To classify signs (translate), the participant then switched to *Translating* mode on the glove and made a sign. If the sign had a match it was displayed as text on the screen and spoken as speech though the speaker. Children had a list of male or female voices to choose from. If no match was found, the screen displayed a ‘failed’ message and returned to *Translating* mode, waiting for new signs. If a sign continued to give a failed message it would be re-trained. The newly recorded samples would then replace the old ones for that sign.

5 RESULTS

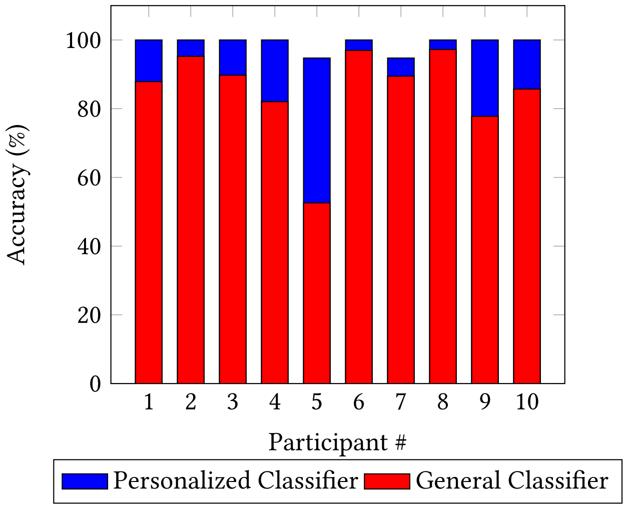

Figure 6: Comparison of personalized and general classifier accuracy

The results of our sessions with the children, in terms of the translation accuracy that they experienced have been summarised in the plot in Figure 6.

Between thirty and forty test signs were performed by each child, with the number dependent on when the child wished to stop. 8 out of 10 participants had 100% accuracy of sign classification when their recorded sample data was trained with their own personally trained classifier. The remaining 2 participants had a 95% accuracy level.

For comparison, we also trained an identical classifier with the aggregate of all of the participants’ data, excluding the user whose data was being classified in each instance. This acted as a general, non-personalised classifier, representing a pre-trained solution that would require and allow no individual customisation or training.

All participants achieved lower accuracy results using this generalised classifier, although some had more success than others. Those who had relatively poor accuracy results using their own personally trained classifier, also tended to have worse results when using the generalised one.

There are a number of factors that affected the results, the primary one being the number of training samples recorded for each sign. This was closely followed in impact by the participant’s consistency of signing and unsurprisingly, the age of the child. The more consistent the child was with their signing, the higher the accuracy of the classification. Both of these two factors were significantly affected by the age of the child, as younger participants tended to be less willing to record larger numbers of training samples, and were also substantially less consistent due to their lesser experience using sign language. The children with the lowest accuracy levels tended to also be those of the youngest age. We believe this was due to both a lower-than-average proficiency and consistency in sign language, as well as the fact that some were distracted by the technology during the training session.

6 DISCUSSION

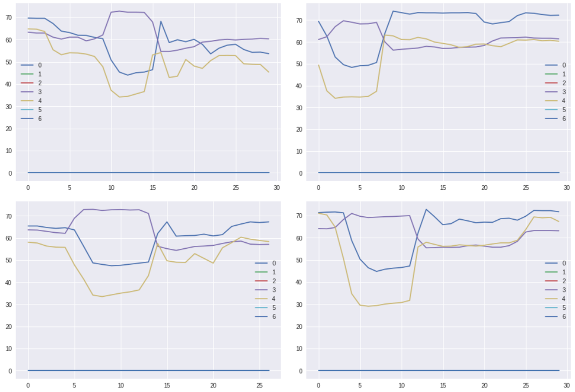

Figure 7: Four different users’ signs for ‘Good Morning’

Gesture data was analyzed to show individual differences and variance across participants. Figure 7 is a draft of hand gesture sensor data for ‘Good Morning’ being signed by four different children. Overall the sign shape and orientation are similar but a closer look reveals variation in speed and duration between participants.

In comparison, multiple sign samples from the same user (Figure 8) reflect relatively minor differences which in turn

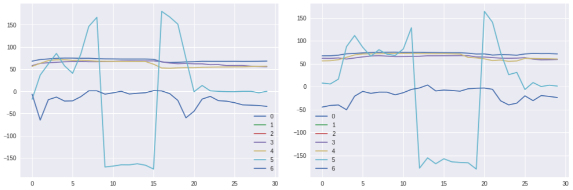

Figure 8: Two examples of the sign for 'Thank You' by the same user

enables the classifier to produce a match with a high level of accuracy.

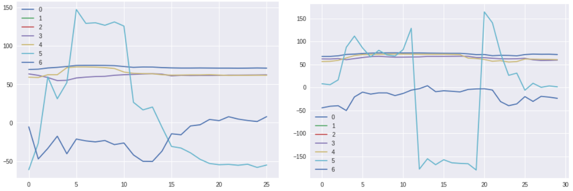

Figure 9: An example of similar gestures with differing duration

When participants trained signs that were different only in orientation or duration, issues with accuracy were observed.

The signs 'Please' and 'Thank you' utilize the same hand shape, and differ largely solely in duration (Figure 9). The classifier was successful in distinguishing the difference between the two signs in all cases, though the confidence provided by the classifier in such cases was lower than the average.

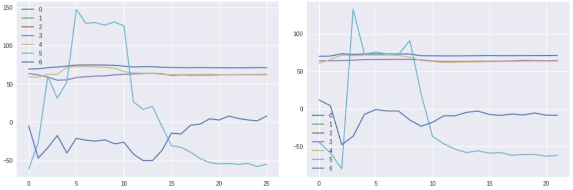

Figure 10: An example of similar gestures, differing only in orientation

The signs 'Please' and 'Stop' utilize the same hand shape and motion, with only a difference in rotation about a vertical axis (Figure 10).

There are sensors readily available in the market that can identify such differences (such as a magnetometer), however, no such sensors were used in our design. The classifier was not able to reliably distinguish the difference between those two signs.

7 CONCLUSION

We have proposed to train personal classifiers to recognise custom dynamic hand gestures using DTW. Our data shows that personal classifiers universally produced more accurate results than general classifiers due to individual differences in hand movements and motor abilities, with a confidence of 99.97% (a z-score of 3.42).

Recognising custom hand gestures widens the application of this technology to extend beyond the sign language community to include individuals who do not use a standard library of sign language due to their personal disabilities and physical limitations of hand movement, such as those seen in stroke victims and in those with other neural disorders.

8 FUTURE WORK

We plan to take this project further in a number of ways, the principle one being the ability to sign continuously with the speech output occurring during, rather than after the sentence has been completed. This would allow users to chain signs together in quick succession to combine individual phrases into full sentences. We also intend to adapt the glove for connection to a smart device in order to provide further customisation.

Hardware Design

Based on the hours of training carried out with the children, it was very clear to us that a less bulky, but still wearable solution would make usage substantially easier. We therefore are looking to minimise the on-body embedded device. We also plan to add a Bluetooth Low Energy (BLE) chip to the flexible glove hand PCB in order to send gesture data to smart devices for wider applications, and provide further personalisation options as detailed below.

Software

In order to give users more control, an app would need to be developed for use on such smart devices. The app would enable users to save their gesture data under a label of their choice, allowing them to build and edit a personal library of signs and could support the ability for the user to modify the language and age of the voice produced by the speech synthesizer.

As described above, the classifier currently waits until the end of the sign before being applied to the entire duration of it. We instead plan to carry out classification continuously on a sliding window over the incoming data. We believe that previous researchers results [3] show great promise and could be applied to our users' custom libraries of hand gestures, while still allowing them to train personalized classifiers. To improve accuracy, this could be combined with a Bayesian model to predict future words based on those already signed.

Signs could also be separated into sub-libraries to reduce the number of possible matches.

9 SELECTION AND PARTICIPATION OF CHILDREN

Speech therapists at participating schools selected children who are non-verbal and used sign language to communicate. Children gave consent by nodding or making the hand sign for ‘Yes’ when asked if they wanted to wear the glove. All testing sessions were supervised by a member of school staff and with a guardian present, both of whom also provided consent. Ethical approval for this study was issued by the University of London ethics committee. Testing data was stored locally on the glove and was not accessible remotely. Measures were implemented to ensure the safety of our wearable device by encapsulating electronic components and adding a fail-safe switch, which turns the battery off while charging. Public liability insurance was in place to protect all participating children.

ACKNOWLEDGMENTS

We would like to thank Essex County Council for sponsoring this study, the participants who took part, their parents and especially their teachers. We also thank Goldsmiths Digital Studios for providing us with a working space and resources to build our prototypes. Finally, we would like to thank BrightSign Technology Ltd for providing hardware and software support on-location throughout the study.

REFERENCES

[1] Ghazi Al-Naymat, Sanjay Chawla, and Javid Taheri. 2009. SparseDTW: A Novel Approach to Speed up Dynamic Time Warping. In *Proceedings of the Eighth Australasian Data Mining Conference - Volume 101 (AusDM ’09)*. Australian Computer Society, Inc., AUS, 117–127.

[2] S Sidney Fels and Geoffrey E Hinton. 1993. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE transactions on Neural Networks* 4, 1 (1993), 2–8.

[3] Granit Luzhnica, Jorg Simon, Elisabeth Lex, and Viktoria Pammer. 2016. A sliding window approach to natural hand gesture recognition using a custom data glove. In *2016 IEEE Symposium on 3D User Interfaces (3DUI)*. 81–90.

[4] Cemil Oz and Ming C Leu. 2011. American Sign Language word recognition with a sensory glove using artificial neural networks. *Engineering Applications of Artificial Intelligence* 24, 7 (2011), 1204–1213. DOI:http://dx.doi.org/10.1016/j.engappai.2011.06.015

[5] Prashan Premaratne. 2014. *Human Computer Interaction Using Hand Gestures*. Springer Science & Business Media.

[6] Stan Salvador and Philip Chan. 2007. Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intell. Data Anal.* 11, 5 (10 2007), 561–580.

[7] Mohammad Shokoohi-Yekta, Jun Wang, and Eamonn Keogh. On the Non-Trivial Generalization of Dynamic Time Warping to the Multi-Dimensional Case. In *Proceedings of the 2015 SIAM International Conference on Data Mining*. 289–297. DOI:http://dx.doi.org/10.1137/1.9781611974010.33

[8] The Makaton Charity. 2019. Makaton. (2019). https://www.makaton.org/

[9] The Raspberry Pi Foundation. 2019. Raspberry Pi Zero W. (2019). https://www.raspberrypi.org/products/raspberry-pi-zero-w/