An Experiential Learning Approach to Learning Manual Communication through a Virtual Reality Environment

Edison Rho, Kenney Chan, Elliot Varoy, and Nasser Giacaman

***Abstract*—There is a pressing need for effective pedagogical methods of manual languages, as evident in the decline of manual languages such as the New Zealand Sign Language. Despite being recognized as one of New Zealand's official languages, recent censuses have shown that fluent New Zealand Sign Language signers have been steadily decreasing. There is a cultural responsibility to preserve such languages, yet the combination of barriers to acquisition and the limited availability of effective teaching methods are standing in the way. In light of this, this paper proposes a computer-assisted sign language learning system that incorporates virtual reality and validation-based feedback as tools to implement the experiential learning model. An implementation in the form of VR-NZSL is presented, targeting the set of New Zealand Sign Language alphabet. Results show that a vision-based classification method, using the Leap Motion Controller, is a scalable, accurate, and usable solution for feedback-assisted manual language learning. A formative usability evaluation of n=10 participants showed promising results for engagement, confidence, and memory retention. The results indicate that virtual reality technology is uniquely situated as an innovative medium for the self-directed acquisition of manual languages. It is hoped this work inspires technology researchers to pursue collaborations with the deaf/Deaf community to design and develop pedagogical technology solutions for manual communication.**

***Index Terms*—Educational technology, neural networks, sign language, virtual reality.**

I. Introduction

Manual communication holds deep cultural and social significance across communities; for the deaf or hard of hearing, manual languages are their primary form of communication. However, most hearing individuals lack the ability to speak or understand manual languages, contributing to the communication gap between the deaf and the hearing. For New Zealand, this gap is widening. The 2013 census—New Zealand's most recent census—revealed 20,235 New Zealanders (roughly 0.5% of the population) possessed the ability to use New Zealand Sign Language (NZSL), a 16% decrease compared to 2006 and a 25% decrease compared to 2001 [1].

The 2006 legislation officially recognizing NZSL was a strong symbolic action, making New Zealand the first country to make a sign language an official state language. This action was made possible through the efforts of the Deaf community of New Zealand, whose members possess a strong political identity due to the threat of denied identities [2]. NZSL as a language traces its roots from the British Sign Language but evolved organically over time since the late 19th century, practically affirmed in 1985 with the introduction of professional sign language training courses [3]. Following this, a period of greater acceptance for NZSL took place [3], culminating in the official recognition of NZSL. However, the lack of infrastructure to support NZSL acquisition blunted the effects of the legislation; a survey of 179 Deaf community members revealed that a weak practical form of education in the 2015 curriculum was a critical threat to NZSL [4]. An inquiry by the New Zealand Human Rights Commission aligns with this purveying view, highlighting that technology could provide such practical support in NZSL education [5]. Regardless of the technological sophistication, a manual language learning tool must appreciate the pedagogical barriers inherent in learning a manual language.

In New Zealand, NZSL was added to the official national high school curriculum in 2016 [6]; its relative infancy means the curriculum stands to benefit from exploration into implementations of effective pedagogical theories. Disregarding such theories could lead to failure in matching curricula content to practical scenarios. Such threats can be seen through a survey of tertiary Turkish Sign Language interpretation students, where an emphasis on theory over practice has led to curriculum dissatisfaction [7]. The survey revealed that a lack of computer-assisted tools, visual demonstrations, and contextual information are causing the dissatisfaction.

Teaching manual languages as a second language follows a degree of similarity with second language learning theories [8], [9]. For example, sometimes, learners prefer a self-directed or distance approach (over face-to-face learning) to reduce language anxiety [10]. Self-direction in optimizing one's learning process is a crucial strategy for dealing with anxiety and enhancing motivation [10]. The learner thus seeks out tools and processes to test out their hypotheses and monitor their progress, personalizing their learning. Effective self-directed learning should be coupled with pedagogies that promote autonomous learning, such as task-based pedagogy and constructivist-oriented frameworks [11], [12].

Constructivism has proven to be promising when applied to second language learning, fostering autonomy in students as well as addressing social and interactive skills in learners [13]. Kolb's Experiential Learning model lays much of its philosophical underpinnings in this constructivist school of thought [14], [15]. Experiential learning manifests when the

The authors are with the Department of Electrical, Computer, and Software Engineering, University of Auckland, New Zealand (e-mail: grho390@aucklanduni.ac.nz; kcha582@aucklanduni.ac.nz; evar872@aucklanduni.ac.nz; n.giacaman@auckland.ac.nz).

learner completes a cycle of four stages: a concrete experience is practiced, which is then critically reflected upon, through which abstract hypotheses are conceptualized, which is tested through active experimentation.

This work proposes that pedagogical tools which promote self-direction and experiential learning have a strong potential to lead the forefront of technology-assisted NZSL revitalization. In this regard, Computer-Assisted Language Learning (CALL) [16] serves to be a strong candidate. As the learner need not be physically co-located with people, these CALL environments have shown to decrease anxiety and increase motivation [17], [18]. Despite this, the current climate of CALL for manual languages have shown an inability to provide concrete experiences and reflective observation—key phases of the experiential learning cycle. Research shows virtual reality (VR) has the potential to sufficiently deliver these phases [19], [20], [21].

VR's appeal as an educative medium for concrete experience comes from its environmental vividness, interactivity, and complete immersion [19], which can be used to create realistic and safe learning environments. Furthermore, VR enables real-time interactivity, indispensable for creating authentic tasks based on realistic scenarios [20]. All such factors contribute to VR as a learning tool that adequately satisfies each phase of the experiential learning cycle [21]. However, with regards to manual languages, further support is necessary for a comprehensive module in reflection—such as validation-based feedback. Effective second language learning requires corrective feedback from teachers or peers [22]. This paper postulates that such feedback can be simulated through computer-assisted validation, using the Leap Motion depth-camera as the input apparatus. As a formative study, the project will limit the vocabulary scope to the static gestures of NZSL alphabet. With VR as a medium for facilitating concrete experiences in addition to feedback as an engine for reflective observation, this project seeks to establish an innovative and self-directed experiential learning tool for manual languages.

Driven by the Human Rights Commission's call to expand the technological landscape of NZSL education, this research aims to answer the following questions:

- *RQ1:* Can a reliable and scalable vision-based validation system be built for the NZSL alphabet?
- *RQ2:* What hand-gesture features are necessary for accurate validation of the NZSL alphabet?
- *RQ3:* How effective is a VR-based approach for the self-directed acquisition of the NZSL alphabet?

The contributions of this paper include:

- The Digital Experiential Learning for Manual Communication (DEL-MaC) framework, a digitally-assisted framework for manual language acquisition.
- The VR-NZSL application that implements the above framework for teaching the static gestures of the NZSL alphabet.
- Identifying hand features necessary for accurate classification of the above gestures.

The paper will first delve into related research surrounding CALL-based manual language education in Section II. A brief background of the conceptual and technological foundations underlying this research is covered in Section III. Section IV presents the steps taken to implement the VR-NZSL application. In Section V, the paper will evaluate the work from two perspectives: classification accuracy and usability. Section VI will conclude the paper, detailing the practical impact of the conceptual framework and the limitations of the current iteration of VR-NZSL.

II. Related Work

Technology has assisted in the acquisition of languages in a plethora of different forms, such as through readily accessible online web applications or online courses. A more recent development has been the integration of virtual reality as a medium for language learning. Such environments can be used in conjunction with auxiliary devices, such as depth cameras, which can promote the interface into the virtual world to be more seamless.

CALL for manual languages, such as web applications and online courses, were found to be more relaxing than the traditional second-language classroom, fostering a sense of security that is conducive to learning [23], [24]. Participants of the popular NZSL web application, Learn NZSL, attributed learning effectiveness towards video-based learning and self-assessment tests [25]. On a higher level, the participants noted accessibility, ease of use, and the flexibility in learning at their own pace as success factors. However, the study highlights that users were demotivated to return to the site due to the lack of a validation mechanism. One participant noted that they were unaware of incorrect signing until notified by an NZSL signer.

Virtual Reality Learning Environments (VRLE) are a subset of CALL, specialized in facilitating concrete experiences. Nicoletta et al.'s study [26] explored VR-assisted mathematics education for hearing-impaired primary school students. The VRLE involved interacting with a virtual avatar to purchase candies, applying arithmetic skills in a realistic scenario. Ying et al.'s VR platform [27] was used in a variety of STEM contexts, such as in mathematics and engineering (welding). Ying outlines benefits in safety, convenience, operating and maintenance costs, teachers required, and self-guidance as the contributing factors for the efficacy of VRLEs. These VRLE serves to develop confidence through competence and practice.

With regards to learning languages, a VRLE for English as a Foreign Language (EFL) was explored in Chen's study [28], which consisted of the users learning through a realistic shopping scenario. The findings showed that listening skills and realistic scenarios enabled effective language acquisition. The ability of VRLEs to simulate a culturally-rich environment allows for learning through cultural immersion [29]. Physical or motivational limitations of being immersed in the target language's community can be inundated through simulating the target language community in a VRLE. A study by Cheng et al. [30] developed a VRLE for Japanese foreign language acquisition, hypothesizing VR could improve language acquisition and stimulate interest in the language's culture. Participants stated that the culturally-rich environment and the ability to look eye-level and talk with the non-playable

characters contributed to an immersive learning experience. Such responses show the effectiveness of VR in creating an environmentthat reflects realistic socio-cultural dynamics. A study on the effectiveness of augmented reality (AR) showed an improvement in accuracy of signing single words by 35% when using AR to learn compared to using video-based material [31]. Although this accuracy was 9% worse than when sign language interpreters were used for learning the words, the AR-based accuracy is an impressive feat compared to using the video-based material.

Recent developments in depth-sensing technology, such as through the commercially available Kinect and Leap Motion controllers, have allowed for affordable avenues of intelligent feature extraction. The Leap Motion controller was utilized by Khelil et al. to classify 10 gestures of the Arabic Sign Language [32] using manually calculated features from the Leap Motion's absolute position data, such as the angles between adjacent fingers and the distance between the fingertips and the palm. An accuracy of 91.3% was achieved using a Support Vector Machine classifier, trained with 1000 labeled frames. Similar accuracies were achieved with Mohandes et al.'s study [33], which used a different feature set and a Feed-forward Neural Network classifier. Misclassified signs were attributed to occlusion of the fingers by the palm or other fingers in the camera's field of view. Kumar et al.'s study [34] solves this problem through the multi-modal use of both Kinect and Leap Motion cameras, where the Kinect is placed in front of the user and the Leap Motion is placed beneath the hand. In this way, occlusion through one camera is compensated by the other camera. Compared to using single devices, combining both input features improved the accuracy rate by 5.91%.

III. Background

A. Experiential Learning and Constructivist Pedagogy

Experiential learning commits the learner to directly experience knowledge through the stimulation of their senses in a contextually-related environment [14], [15]. Kolb formalizes this definition into the experiential learning model, which follows a cyclical sequence starting with concrete experience of the new knowledge, the questioning of existing preconceptions based on the knowledge, using critical reflection to instill emotion and ownership into the new knowledge, then extracting value from the consequences of implementing the knowledge through action. Experiential learning is founded on *constructivism*—the theory that knowledge is a result of a personal reality actively constructed from one's own authentic experiences [35], [14]. The learner contextualizes the constructed knowledge by relating it to existing worldviews or deconstructing existing ones to accommodate the new knowledge [36]. The constructivist learning approach requires a context in which learners are able to effectively construct new knowledge.

B. Virtual Reality

VR is a hardware-independent state of experience within which the individual feels present through interacting with a stimulating yet artificial environment. Presence is defined as the perception of a physical environment, regardless of how that environment is actually mediated [37]. Presence is mediated by the 'Three I's of VR' [38], [39], [40]: *imagination*, *interaction*, and *immersion*. Immersion captures the vividness of the simulation—the richness of the environment's formal features to the senses. Vividness can be broken down into breadth (the number of sensory dimensions that are simultaneously stimulated) and depth (the resolution from those sensory channels). Interactivity is defined as the extent of realism to which the user can map their human actions to those in the mediated environment. Finally, imagination leverages the creativity and problem-solving aptitude of the user. By providing an immersive and interactive environment, the user can role-play in the artificial scenario [19].

C. Leap Motion Controller

The Leap Motion Controller is an infra-red video camera that extracts various information from the hands and recreates them as 3D models. The camera is able to detect hand positions with an accuracy of 0.7mm under ideal conditions [41]. The Leap Motion's API is able to provide various features, such as the rotation of the hands as a quaternion, the grab and pinch strengths made by the hands, and, crucially, the XYZ-coordinates of the individual joints of the hands relative to the position of the camera. The Leap Motion Controller can be used in conjunction with virtual reality, where the camera is attached to the front of the headset.

IV. VRSL Learning System

A. The DEL-MaC Conceptual Framework

The Digital Experiential Learning for Manual Communication (DEL-MaC) framework serves as the project's theoretical backbone for technologically-aided manual language acquisition. Utilizing established pedagogical principles into an implementable framework, DEL-MaC seeks to address the barriers that manual language learners typically experience. DEL-MaC is built upon the Experiential Learning model, incorporating the fundamental phases of concrete experience, reflective observation, abstract conceptualization, and active experimentation. This foundation is then used to build a framework whose components can be delivered through digital means. The DEL-MaC framework is illustrated in Fig. 1 and further described below.

1) Abstract Conceptualization: Abstract conceptualization and concrete experience are modes of grasping experience. Abstract conceptualization achieves this through subconscious input processing—personalizing observed input into understandable concepts. It can be seen as a passive form of experience, which is a crucial component for language acquisition through the *Input Hypothesis* [42]. The hypothesis asserts that the learner must be given ample opportunity to observe communication between individuals of the target language, particularly communication that is comprehensible to the learner. This could be done in indirect ways, such as visual cues and body expressions of the signers. Within a digital medium, input can be provided through videos;

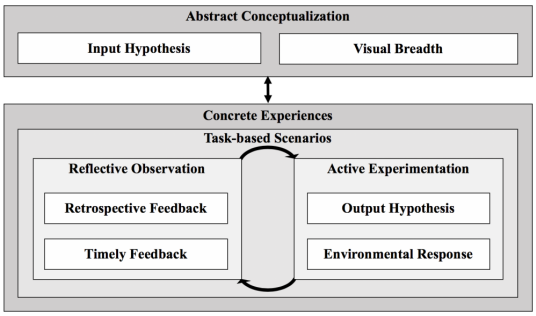


Fig. 1. The core components underlying the Digital Experiential Learning for Manual Communication (DEL-MaC) framework.

combined with replay controls, the user can gain input on demand.

Input is primarily visual with manual communication, consisting of two input modes—a first-person perspective and a second-person perspective; the visual input of one's own practice of signing (first-person) is different from the visual input of observing another person signing (second-person). The perspectives play different roles in acquisition: the first-person perspective is required for the learner to easily follow the gestures, whereas the second-person perspective is instrumental in recognizing gestures made by others [43]. Hence, *Visual Breadth* of input perspectives becomes necessary. By becoming familiar with both perspectives, the learner is able to effectively conceptualize the gestures, providing them with the ability to both recognize and reproduce the gestures.

Crucially, this period of visual comprehension must occur before any active participation of the new knowledge, according to the *Input Hypothesis*. Delaying output allows for the complex and subconscious processing of the target language so that the learner's recognition knowledge of gestures develops before their gesture retrieval knowledge [44]. Once such concepts are formulated, the learner becomes prepared to test those concepts within an experiential context.

2) Concrete Experience: Within the phase of concrete experience, learners actively participate in an immersive knowledge-building scenario. Rather than seeing concrete experience as a phase, the framework sees it as an environment—a context within which the learner is able to actively experiment with new knowledge and reflectively observe those outcomes. With language acquisition, a medium of concrete experience is an opportunity through which culturally-rich stimuli can be conveyed to the learner. To implement DEL-MaC, such an environment must be constructed to be immersive and interactive so that the learner can take the role as an active player in a realistic scenario.

Realistic scenarios can be constructed through task-based design. Tasks provide the means through which a general concept can be used in a variety of contexts; the higher the frequency and diversity of tasks, the greater the necessity for the learner to generalize the skill. Furthermore, tasks are fundamentally goal-oriented, which can be used as a marker for progress—evidence that the learner has met the criteria to competently apply the skills in the real world [45]. A task-based approach is thus a suitable framework for the self-directed learner, whose engagement and continual participation in the acquisition journey is predicated upon indications in improvement [11]. When tasks are designed to reflect real use cases, it is able to provide a context that encourages constructive learning [20].

3) Active Experimentation: Active experimentation is a mode of actively transforming experience into learning [46]. Whereas active conceptualization places an emphasis on the development of theory through input, active experimentation seeks to test the concepts through the *Output Hypothesis*. To produce gestures, the learner must transform the amorphous and abstract concepts of the gestures into accurate output, thereby pushing the learner to process the language at a deeper level [47]. Moreover, prompting output could cause the learner to notice 'holes' in their knowledge, such as identifying which letters they currently struggle with.

Taking risks and making mistakes are strategies to cope with the anxiety of language acquisition, provided those mistakes are inconsequential [10]. Hence, an active experimentation scenario must minimize the consequences from, or even encourage, making mistakes. Public speaking in a foreign language is a major source of anxiety for second language learners, primarily due to the fear of making mistakes [48]; creating an isolated context to output the target language without signing to a real person can create an environment for developing competency without the anxiety from judgment. Furthermore, for a task to enable active experimentation, the environment must respond to the user output to show that their interactions have significance [15]. With a digital environment, this environmental response can be easily designed, with the degree of power and frequency of the response easily calibrated.

4) Reflective Observation: Whereas active experimentation transforms experience into learning through action, reflective observation seeks to perform the transformation through reflection. It plays a cooperative role with active experimentation [49]—errors made within the experimentation will be identified and rectified through feedback, filling the 'holes' in their knowledge. The learner is able to then reflect on erroneous past hypotheses, compare them with the *Retrospective Feedback* received, and restructure their abstract concepts of the gestures to satisfy the feedback. Hence, within the concrete experiential context, the learner constantly transitions between active experimentation and reflective observation.

Feedback is also a crucial mechanism to satisfy the self-directed learner; the ability to ascertain one's weaknesses and strengths in language is a key part of managing self-directed learning [10]. Indeed, self-direction is predicated on monitoring of progress and identifying objectives [50], which can be achieved by feedback and task completion, respectively. *Timely Feedback* after task completion increases learning effectiveness compared to delayed feedback [51]. This is particularly important in learning applications with game-like characteristics, where the retrospective (yet timely) feedback needs to be unobtrusive [52].

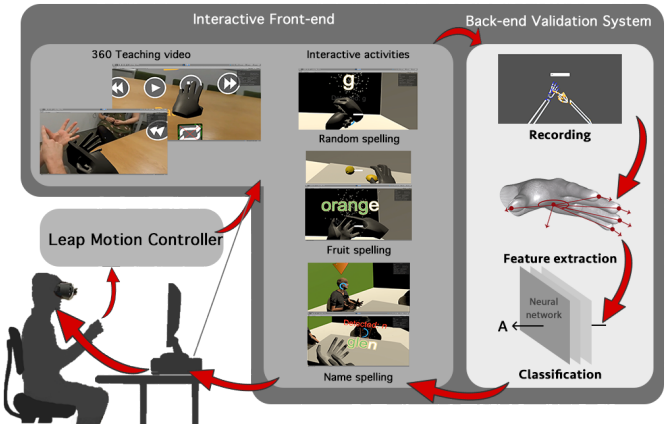


Fig. 2. Modular overview of VR-NZSL. The setup includes a Leap Motion controller connected to the VR headset. The front-end consists of 360 teaching videos and a range of interactive activities, while the back-end largely deals with validation.

B. VR-NZSL

The VR-NZSL application seeks to implement the DEL-MaC conceptual framework through three major components: VR-based 360-video, interactive tasks, and validation-based feedback. Together, these components form a self-directed and self-contained tool for learning the static gestures of the NZSL alphabet. The Leap Motion controller models the user's hands inside the VRLE, allowing the user to make gestures within the VRLE as well as interact with virtual objects or control the VRLE. Fig. 2 illustrates the overall setup; the front-end components are described below, while the back-end validation system is presented in Section IV-C.

1) Setup: The VR-NZSL application uses a mobile VR setup in conjunction with a computer. While the application is running within the computer, the VRLE is streamed to the user through a VR-ready smartphone, attached to the user through a mobile VR headset, such as the likes of Google Cardboard. The accelerometer within the mobile phone captures the head movements of the user, providing the gravity and coordinate data to the computer. Additionally, the Leap Motion controller is attached to the front of the headset, where it captures the hand data. The computer uses these data to render the stereoscopic view of the VRLE, which is streamed back to the mobile phone. With the wide availability of VR-ready smartphones and computers, as well as the affordability of the Leap Motion controller, it is expected that this hardware setup is affordable to potential learners of NZSL. The application is prevented from being a fully-packaged mobile application due to the lack of a Leap Motion mobile SDK. However, the Android SDK for Leap Motion is currently in a closed beta period [53], thus it is expected that a fully mobile VR-NZSL can be implemented in the near future.

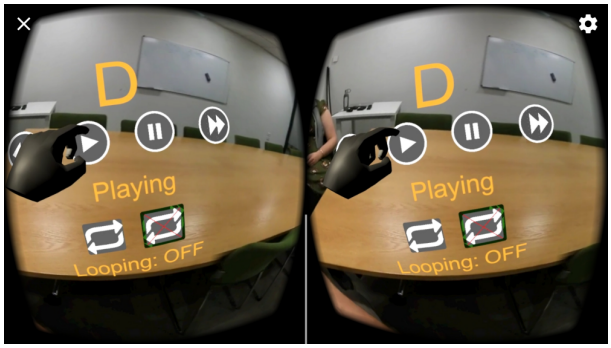
2) 360-video: The 360-video features two fluent NZSL signers role-playing as part of a private lesson on the NZSL alphabet, with one actor playing the Teacher and the other a Student. The video consists of the signers cycling through the letters, repeating each letter twice with the second try being a slower rendition. To prevent visual overloading, the signers sign the letters sequentially, with the Student following the Teacher. By utilizing two actors, the 360-video can achieve *Visual Breadth*. The video camera was positioned by the shoulder of the Student, hence when he signs the gestures, the learner views the gesture from a first-person perspective (Fig. 3(a)), ensuring that the learner can easily follow along. Conversely, when the Teacher signs the gesture, the learner is able to see the gesture made from a second-person mirrored perspective, developing their ability to recognize gestures. With the aid of video controls (Fig. 3(b)), the user can pause, skip to the next or previous letter, and 'loop' the current letter so that the video segment of the current letter continuously repeats.

The 360-video is an implementation of the *Abstract Conceptualization* phase of DEL-MaC. Through a mostly passive input processing phase, the learner is given the opportunity to subconsciously process the observed gestures. They are given the time to recognize patterns between gestures, aiding in recall. Hence, a medium for the *Input Hypothesis* is achieved. The use of VR as a medium for the 360-video is two-fold. Firstly, it becomes more cohesive for the learner when coupled with the interactive VR tasks. More importantly, watching the 360-video in VR offers a greater sense of spatial awareness and immersion. By placing the Teacher-Student actors in a virtual world, the learner is immersed into thinking they are right next to the actors, allowing them to more easily mimic with the gestures being made as well as making them feel accustomed to real signers communicating with each other. In this sense, *Visual Breadth* is more effectively incorporated into the application.

3) Interactive Tasks: The interactive task module is composed of three activities: random letter spelling, fruit spelling, and name spelling. Together, these activities provide *Concrete Experiences* within which the user can *actively experiment* with the concepts learned in the 360-video. A key aspect of *Concrete Experience* is immersion. To maintain immersion, a hint system was introduced that allows the user to remember the gesture and continue with the activity unimpeded; when a user performs the hint gesture, a diagram of the gesture appears within the VRLE, providing visual aid. Without a hint system, this immersion is broken when the learner cannot remember a gesture, as they would have to temporarily terminate the task to go back to the 360-video to relearn the gesture.

The random spelling activity involves the user being prompted to sign random letters that appear in the VRLE. The activity emphasizes repetition, focusing entirely on implementing the *Output Hypothesis* by maximizing the efficiency of outputting gestures, whilst allowing them to familiarize themselves with the game mechanics of the interactive tasks. In the fruit spelling activity, the user is tasked with spelling the names of a set of fruits that are laid out within the VRLE. The user is instructed to pick up the fruits and place them on a counter, upon which the name of the fruit appears in the VRLE. This activity emphasizes interaction and *Environmental Response*—within the VRLE, the user is able to appreciate that their actions in the environment elicit consequences, providing a sense of meaning into their learning. Additionally, the task promotes comprehensible and visual association with the gestures being learned, aiding in recall [42]. The name

(a) Following a first-person demonstration



(b) Interacting with gesture video controls

Fig. 3. Stereoscopic views of the 360-video in progress, with the user (a) following the gesture of the first-person demonstration while the second-person perspective waits, and (b) interacting with hand gesture video controls.

spelling activity places the user in a setting surrounded by 3D avatars, where they role-play in befriending the avatars through finger-spelling their names. This activity seeks to replicate a real use-case of the NZSL alphabet through the medium of a realistic scenario and environment. As such, the task follows constructivist instructional principles of simulating realistic scenarios [20], developing the learner's confidence in applying the alphabet in real life.

All tasks are supported by a validating mechanism through which the user is encouraged to make *Reflective Observations*. When the learner attempts a gesture to spell a letter, the application begins to classify the gestures being made, providing *Timely Feedback*. Due to the fact that the Leap Motion controller captures frames at up to 1000 frames per second, the classification would commence immediately upon the task being prompted. Hence, to enhance usability and responsiveness, the feedback mechanic uses a loading meter, as shown in Fig. 4. A frame is captured every 0.1 seconds, upon which it is classified by the validation system. Once 10 such frames of the correct letter have been classified, the loading meter is completed, upon which the user must sign the next letter. Incorrect letters do no contribute to this meter. Meanwhile, the letter that the validation system thinks the user is making is constantly visible to the user. Consequently, the feedback system provides to the user a means to correct their gesture if their gesture is wrong, or validating the gesture if it is correct, providing *Retrospective Feedback*. With this validation system, there is zero penalty for wrong classifications, aligning with established anxiety-coping strategies and research that reports fear of failure and embarrassments are primary causes of anxiety [10], [48].

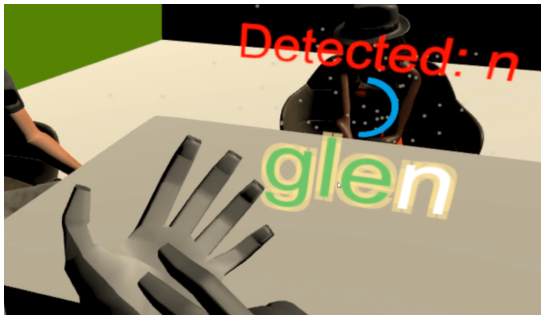


Fig. 4. The name spelling activity: the blue circle represents the loading bar for validation in progress.

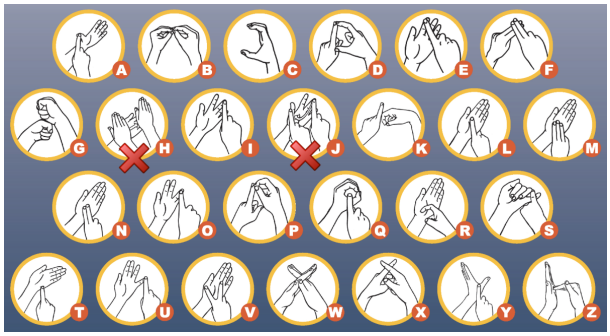


Fig. 5. The NZSL alphabet [54] used in the evaluation. Letters 'H' and 'J' were excluded as they involve dynamic (moving) gestures.

C. Validation

VR-NZSL uses the feed-forward neural network as its classification engine for the validation system. Neural networks have shown promising accuracy in previous research, particularly in computer vision classification [33].

1) Recording: The dataset used to train the neural network was recorded manually, with a total of 24,000 frames being recorded. To record the data, a Unity application was created that allows for frames to be labeled with the letter that it represents. The application provides visual feedback of the signer's hands so that the signer can confirm the Leap Motion controller is correctly recording the data. Once recorded, the data is serialized for later use. The recording process involves informing the application of the letter that is currently being recorded, followed by the recorder signing the gesture. A frame is stored every second for 10 seconds—the time intervals between captures allow the recorder to reposition their hand, increasing variation in training data. This 10-second procedure is repeated 100 times, generating 1000 frames for each letter in total. It is then applied for the 24 static NZSL gestures (all letters in Fig. 5 except the dynamic letters 'H' and 'J'), amounting to 10 hours of recording.

2) Feature Extraction: Feeding the absolute positions of the joint positions into the neural network will cause the neural

Table I
FEATURES EXTRACTED FOR THE NEURAL NETWORK AND THEIR ASSOCIATED NUMBER OF VARIABLES NEEDED TO REPRESENT THEM.

Feature	Symbol	Variable count
Quaternion: Hand Rotation	HR	8
Vector: Hand Direction	D	6
Vector: Hand Normal	N	6
Grab Strength	G	4
Pinch Strength	P	4
Distance: Fingertip–Palm Center	FPD	10
Angle: Adjacent Fingertips	AA	8
Angle: Fingertip–Palm Normal	FPA	10
Distance: Fingertip–Opposite Hand's Palm Center	FOPD	10
Distance: Fingertip–Opposite Hand's Fingertips	FOFD	25

network to be dependent on the position of the hands relative to the Leap Motion controller. Two options were available to make the neural network robust to the hand positions. The first option would involve having a larger training size with a deliberate effort to record data at different positions. Another option was to selectively extract features that are position-independent. The latter was chosen to minimize effort in acquiring training data. As there was initially little information on how much training data would be needed, it was crucial to make design decisions that could minimize the training data required by the classifier. The features extracted are shown in Table I.

Features {D, N, FPD, AA, FPA} were chosen based on a similar study for Arabic sign language [32], [33]. However, the studies were based on one-handed gestures, thus features FOPD and FOFD were added to account for the two-handed nature of the NZSL alphabet. The remaining features were added iteratively based on the models' in-sample recall score; if a particular letter had very low recall scores, then a feature was added that could rectify the current model's weaknesses.

Features {HR, D, N, G, P} are features that are directly provided by the Leap Motion API. The remaining features must be manually calculated based on vectors of the hand-joints' XYZ-Cartesian coordinates. FPD is calculated for each finger on each hand as the absolute difference between the fingertip position and the palm center of the hand in Euclidean space. AA is calculated for all fingers for each hand as the angle between adjacent fingers. FPA is calculated for each finger as the angle between the finger and the palm of the hand. FOPD is calculated for all fingers on each hand as the difference in position between the fingertip and the opposite hand's palm center. Finally, FOFD is calculated as the difference in position between fingertips of opposite hands. The feature values were normalized using a Min-Max scaler before being used as training data.

3) Architecture Design: The architecture of the neural network, as illustrated in Fig. 6, consists of the following layers:

- Input layer with 91 neurons (the same value as the total number of feature variables in Table I)
- Three hidden layers, each with 400 neurons and using the rectified linear unit as activation functions.
- Output layer with 24 neurons using the softmax activation function, with each neuron representing the static letters of the NZSL alphabet.

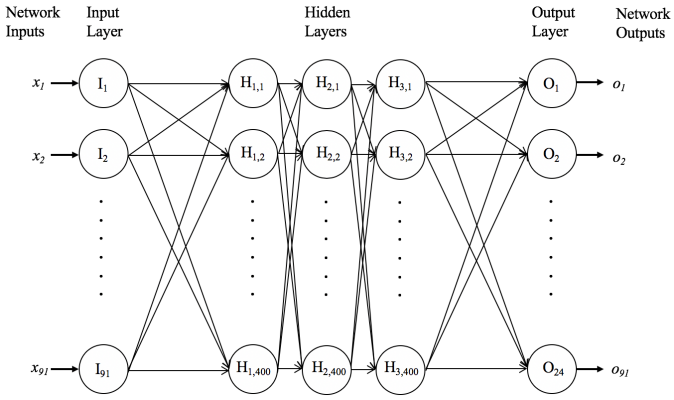


Fig. 6. Architecture of the final neural network model used.

Table II
THREE MEASURES USED TO INDICATE ACCURACY OF THE NEURAL NETWORK.

Accuracy Measure	Symbol	Accuracy (%)
Cross-Validated Accuracy of Architecture	CVA	98.7
Test Set Recorded by Researchers	TRA	95.3
Test Set Recorded by NZSL Professionals	TPA	85.6

The neural network was trained with 30% validation split, and 20 epochs under a batch size of 512 samples, with a mean epoch training time of 0.154ms. For the test sets, the final model took 0.935ms to classify each image, on average. The network was trained using an nVidia 1080 GTX GPU.

V. RESULTS

A. Classification Evaluation

To analyze the classification performance of the neural network, three different indicators were used (Table II). CVA was calculated using a 10-times 10-fold stratified cross-validation process, where a new neural network with identical architectures described in Fig. 6 is trained for each fold iteration. The accuracies for each fold iteration is recorded and the mean is used to calculate CVA. CVA is an analysis of the general performance of the architecture, rather than one specific model. For the final model used in the application, separate test sets were used, each with 150 frames per letter. The TRA test set was recorded by the two researchers that recorded the training set. TPA was independently recorded by two fluent NZSL signers independent from the study, with minimal supervision provided during the recording.

All three indicators exhibited high accuracies, giving promising indications for the reliability aspect of ***RQ1***. Due to TPA being recorded by signers not involved during the training of the neural network, it exhibited the lowest accuracy at 85.6%, indicating the model had overfitted to the researchers' training data. Section V-A5 gives insight into why this over-fitting occurred. Despite the lower accuracy, the professionals commented on the high usability of VR-NZSL, barely noticing the underperformance of the model with the exception of the worst classified letter, V.

1) Performance Measures of Predictive Models: The performance measures of *accuracy* and *recall* are established performance measures for a model that tries to predict a factor that can take one of many possible classes (i.e. letters in the NZSL alphabet) [55]. Accuracy for the *i*-th class is defined as:

$$Accuracy_i = \frac{TruePositive_i + TrueNegative_i}{Positive_i + Negative_i}$$

A *positive* sample is a sample with class *i* and a *negative* sample is a sample that is not class *i*. A *true positive* is a positive sample correctly classified as class *i*, and a *true negative* is a negative sample correctly classified as not class *i*. As the equation shows, the numerator is the sum of both true positives and true negatives for the *i*-th class. For a given class *i*, there are *n*–1 other classes, each contributing to high true negative classifications for the *i*-th class. Consequently, due to the very high true negative and negative values for a given class, the accuracy for each class will be very similar thus it can be a deceptive measure when comparing between accuracy across different classes.

To account for this, the recall performance measure is used. For a given class *i*, recall has the following property:

$$Recall_i = \frac{TruePositive_i}{Positive_i}$$

Recall is sensitive to false negatives (when class *i* samples are incorrectly classified as not class *i*) hence it is a great measure for identifying letters that are not being correctly recognized. This measure is a more realistic depiction of the performances of an individual class thus this measure will be used when comparing between classes.

Both measures can be reliably calculated using the 10-times 10-fold stratified cross-validation technique [56]. This technique is a standard procedure for minimizing bias in estimates in predictive models.

2) Architecture Development: The architecture was designed using a grid-search cross-validation across different numbers of layers and neurons. Specifically, up to six layers were explored with 50, 100, 200, 400, and 1200 neurons explored for each layer. Finer granularities of 25 and 800 neurons were used for the first layer for exploratory purposes. A 10-times 10-fold cross-validation was performed for each combination and the accuracy was recorded. For networks with five or six layers, the larger numbers of neurons could not be tested due to memory limitations of the hardware.

Across the combinations explored, three hidden layers with 400 neurons gave the highest cross-validated accuracy at 99.0%, therefore this architecture was chosen to create the final model. Combinations with a high number of parameters showed decreasing accuracies; neural networks with a large number of parameters have a greater tendency to overfit the data. Although this could be mitigated with a larger training size, the accuracy for the best combination was already exceptional, hence the effort required to record the data to support deeper neural networks was not justified.

3) Training Size: A performance analysis of models trained across different training sizes was used to validate whether the recorded training size was sufficient, as well as to evaluate the scalability of the manual recording approach. Training sizes of $support = \{5, 10, 25, 50, 100, 200, 400, 800\}$ frames per letter were used in a stratified 10-times 10-fold cross-validation process. In addition to the method described at the beginning of Section V-A, each training run of the neural network involved randomly selecting examples of size $support$, without replacement, out of the training folds. For each fold, a neural network with three layers and 400 neurons was trained and the accuracy against the test fold was recorded. Consequently, for each training size tested, 100 accuracies were recorded, from which a 95% confidence interval can be calculated based on the Central Limit Theorem.

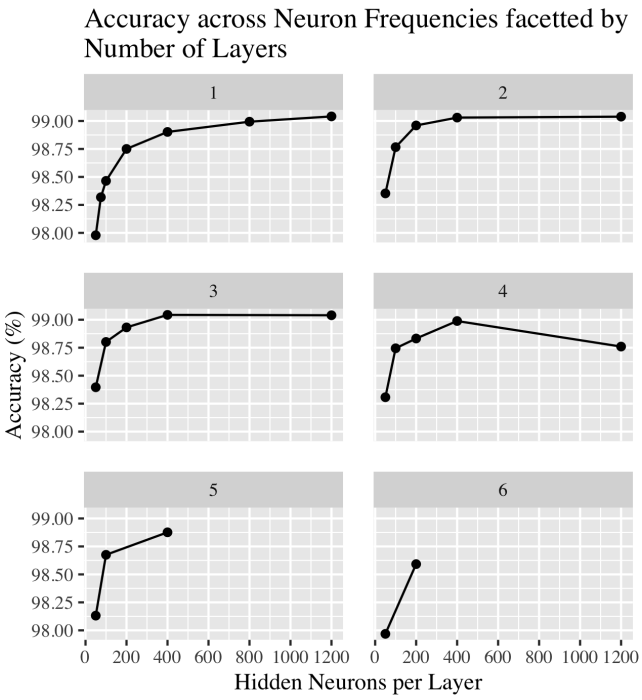
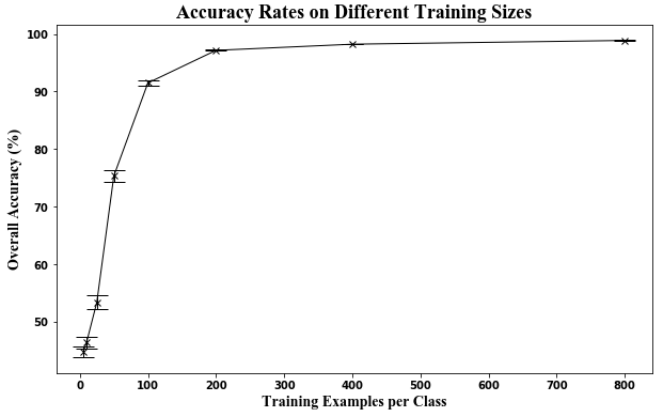


Fig. 7. Facetted line graph showing the iterative exploration of cross-validated accuracies, facetted across different numbers of layers.

Fig. 8 shows a rapid increase in cross-validated accuracy for *support*$\leq$100, after which the rate of improvement gradually plateaus. The original *support*=1000 is well across the plateauing area, showing that the training size was more than sufficient for training the final model. Furthermore, the graph shows significant diminishing returns after 200 samples per letter. Proportionally, this would indicate that similar accuracies could have been achieved within a two-hour recording session, as opposed to the original 10 hours; optimizing the recording application could have further reduced the recording time. Ultimately, the training size analysis serves to show the scalability of the recording procedure, which is promising for further research seeking to expand the vocabulary. Consequently, the scalability aspect of ***RQ1*** has been answered.

4) Feature Importance: An analysis of how each feature contributed to the final model's performance was crucial in answering ***RQ2***. The analysis involved applying a 10-times 10-fold stratified cross-validation on each of the 1024 differ-

Fig. 8. Cross-validated accuracies across different training sizes, with error bars representing 95% confidence intervals.



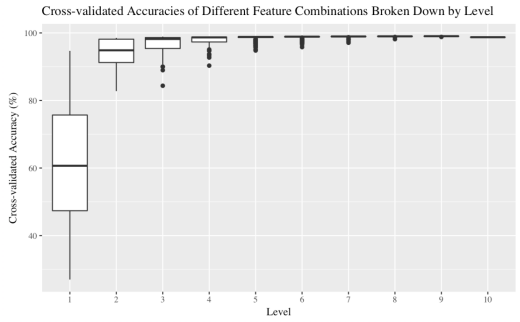


Fig. 9. Cross-validated accuracies of the 1024 feature combinations, broken down by level.

ent combinations of features to extract the accuracies. Once recorded, the top ten combinations for each combination sizes, or levels, were identified. The ten combinations were then tested on using the combined {TRA,TPA} test set to find the best combination for each level. For this final test stage, the combined test set was used to train 50 neural networks of identical architectures and the average accuracies across those neural networks were taken. This was necessary to mitigate the randomness in neural network training, as well as to make the evaluation agnostic to any one specific model for each combination.

In general, Fig. 9 shows that the more features there are in a combination, the higher the accuracies and smaller the spread. This can be explained by the fact that the weaknesses of one feature can be compensated by the strength of another; the more variables there are, the greater the effect of this phenomenon. Table III shows the addition of new feature variables causing previously misclassified letters to be correctly classified. For example, feature combination #5 struggled with the letter 'R', but the addition of pinch and grab strengths in feature combination #6 captures the anatomy of the pinching and grabbing gestures of the right hand that are necessary to make an 'R' gesture.

FOFD was consistently in the best combination for each level except in $level = 2$, suggesting a high feature importance due to its ubiquity across all levels' best combination. This could be attributable to the fact that FOPD requires a large

Table III
THE BEST MODELS FOR EACH LEVEL BASED ON ACCURACY TOWARDS THE COMBINED {TRA,TPA} TEST SET, AS WELL AS THEIR WORST CLASSIFIED LETTERS IN TERMS OF RECALL PERFORMANCE.

	Feature Combination	Test Accuracy (%)	Top 3 Worst Letters
1:	{FOFD}	85.6	F, L, E
2:	{AA, FOPD}	89.5	E, R, X
3:	{N, FPA, FOFD}	88.5	V, O, T
4:	{HR, AA, FPA, FOFD}	90.1	E, X, V
5:	{HR, AA, FPA, FOPD, FOFD}	90.3	E, X, R
6:	{HR, G, P, AA, FPA, FOFD}	91.0	K, V, E
7:	{HR, G, FPD, AA, FPA, FOPD, FOFD}	91.4	V, K, R
8:	{HR, D, G, P, AA, FPA, FOPD, FOFD}	90.4	L, K, O
9:	{HR, D, G, P, FPD, AA, FPA, FOPD, FOFD}	90.8	L, V, K
10:	{HR, N, D, G, P, FPD, AA, FPA, FOPD, FOFD}	90.4	V, O, L

number of variables to describe it, as shown in Table I, as well as FOFD. Feature combination #7 ({HR, G, FPD, AA, FPA, FOPD, FOFD}) provided the highest accuracy, answering ***RQ2***. Despite this, most combinations had a similar accuracy (with the exception of level one combinations), thus it is entirely possible that the highest accuracy of feature combination #7 was coincidental due to specific test sets that were used. Nevertheless, the results show that a small number of the correct features can be used to achieve similar accuracy, which is a promising result for future development of manual gesture hardware with limited feature collection power.

5) Recall of Letters: Analysis of the neural network's performance across the different NZSL letters is crucial to understand the strengths and weaknesses of the Leap Motion classification system, and hence to answer the reliability dimension of ***RQ1***. As discussed in Section V-A1, recall is a suitable performance measure against individual letters. For the following analysis, test set accuracies will be used in place of cross-validation, as Section V-A3 indicates the test set is of sufficient size to generalize to NZSL gestures. The analysis first focuses on the performance of the final model on the combined {TRA, TPA} test set in an attempt to make the analysis independent of the signer. As a means to explore the relationship between different letters, the analysis also explores 2nd order and 3rd order recall performances. The *n*-th order recall for a given letter is calculated by considering the top *n* frequently classified letters as true positives. For example, if 56% of 'V' samples were classified as 'V' and 42% were classified as 'N', then the 2nd order recall for 'V' is the combined 98%. Recalls at different orders can be used to measure the breadth of classification, indicating whether a gesture is confused with very few or many other letters. Furthermore, recalls of the individual letters were compared between TPA and TRA to analyze the model's classification behavior across different individuals.

6) Combined Recall: With the exception of letters such as L, O and V, Table IV shows that there is generally a high first-order recall rate across the letters. The second-order recalls show significantly high improvements from the first-

Table IV
RECALLS ACROSS THE STATIC NZSL ALPHABET AT $order = \{1, 2, 3\}$ FOR THE COMBINED TEST DATASET {TRA, TPA}. LETTERS WITH *order* = 1 RECALL ABOVE 99% HAVE BEEN OMITTED FROM THE TABLE.

Letter	Recall (%)		
	Order=1	Order=2	Order=3
G	95.0	97.3	98.3
I	91.3	96.1	98.7
K	78.7	93.3	95.7
L	61.6	87.1	93.2
M	91.7	95.3	97.0
O	59.0	99.7	100
P	97.4	99.4	100
R	90.3	96.8	98.4
T	93.0	96.0	97.7
U	79.7	96.8	99.4
V	56.7	95.3	99.0
X	84.3	94.3	99.7

Table V
RECALL COMPARISONS BETWEEN TRA AND TPA. LETTERS WITH RECALL ABOVE 99% FOR BOTH TEST SETS HAVE BEEN OMITTED.

Letter	TRA Recall (%)	TPA Recall (%)
G	92.0	98.0
I	83.1	100
K	86.0	71.3
L	94.0	31.3
M	90.0	93.3
O	98.7	19.3
P	95.0	100
R	92.7	88.1
T	94.0	92.0
U	88.8	70.0
V	90.0	23.3
X	88.7	80.0

order recall, showing that most poorly performing letters were primarily confused with one other letter. Generally, letters had perfect or near-perfect classifications by the third-order recall, with all third-order recalls being above 93%. This supports the hypothesis that the final model generally confuses a letter with very few other letters. However, there is a correlation between poorly performing letters across the three different orders, such as letters such as K, L, and M. These letters consist of similar features with many other letters—particularly L, which shares the common structural features with M, N, T and the vowels. These results can serve as a basis for further feature engineering, where new features could be introduce to target the worst-performing letters based on recall.

7) Recall of Different Test Datasets: Table V shows that TPA performs generally worse than TRA across all letters, particularly for the letters L, O and V, indicating that the final model has overfitted to the researcher's data. Whether this is due to the structural differences of the hands (such as finger length or width) can be ascertained through analyzing performances of individual letters, which can give insights into the types of hand features that the model struggles with. Table V indicates that the overfitting is attributable to the *stylistic* differences of the gestures in combination with limitations in the Leap Motion controller's sensitivity, rather than the structural differences of the recorders' hands. The root cause of such discrepancy is attributable to TPA being signed independently

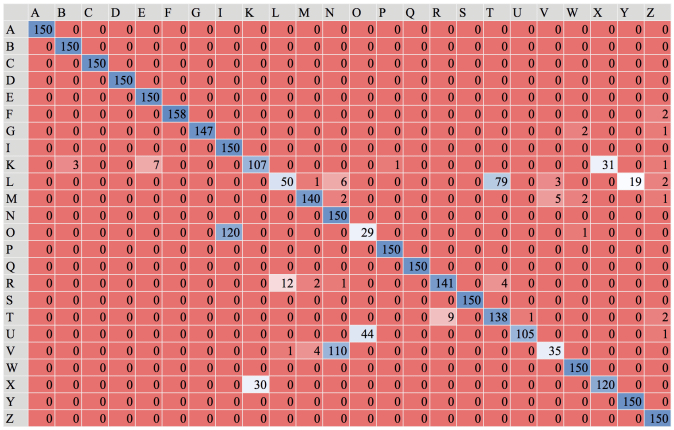


Fig. 10. Confusion matrix for TPA, with columns as the predicted class and the rows as the actual class.

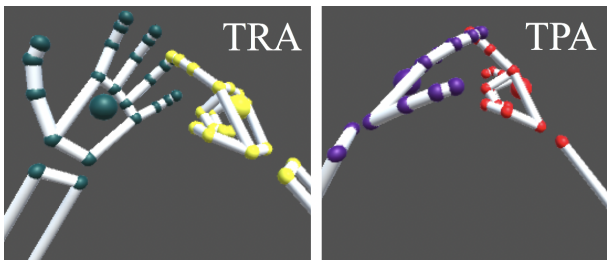


Fig. 11. Comparison of the stylistic differences in signing the letter 'O' between the TRA (left) and TPA (right) test sets.

by the NZSL professionals, with minimal corrective guidance from the researchers, leading to the professionals signing in their own NZSL 'dialect'. From the confusion matrix in Fig. 10, it can be seen that 'O' samples were almost exclusively confused as 'I'. From Fig. 11, the misclassifications can be explained. The gesture 'O' involves the right hand's index fingertip touching the left hand's ring fingertip. TRA's gesture style involved the left hand's palm facing the Leap Motion controller. However, TPA's 'O' gestures were stylistically different, with the left hand's palm facing perpendicular to the Leap Motion controller. This causes the left hand's index fingertip occluding the right hand's index fingertip, causing the Leap Motion controller to predict that the right hand's fingertip is touching the left hand's middle fingertip (which is the gesture for 'I') instead of the ring fingertip.

'V' samples from TPA were the second most misclassified samples, with 73% of 'V' frames being misclassified as 'N'. Fig. 12 illustrates how both gestures have similar features. By inspecting the Leap Motion's remodeling of 'V' and 'N', it is observed that the differences are hardly distinguishable. TRA's recall for 'V' is much higher than TPA, primarily due to the researchers making a conscious effort during recording to separate the right hand's middle and index fingers to an unnatural degree so that the Leap Motion controller could detect the separation. During TPA recording, the recordings were much more natural, preventing the detection of the separation and causing the final model to misclassify the frames as 'N'. Ultimately, these stylistic differences failed to be caught by the final model, suggesting that future training

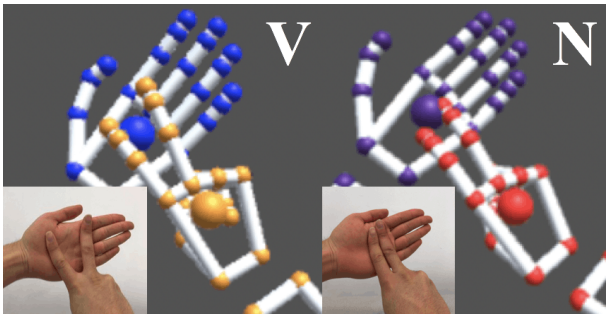


Fig. 12. Comparison of the Leap Motion Controller's remodeling of V (left) and N (right) gestures.

samples should aim for broader stylistic variations so that classifiers can capture such differences. For completeness, the confusion matrix for TRA is also presented in Fig. 13.

	A	B	C	D	E	F	G	I	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	149	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
B	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	149	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1	0	0	0	0	0	138	0	0	0	0	0	0	1	0	0	3	1	0	0	5	0	0	1
I	0	0	0	1	15	0	0	133	8	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0
K	0	0	0	5	0	0	0	0	129	0	0	0	0	3	0	0	0	0	0	0	0	13	0	0
L	0	0	0	0	0	0	0	1	0	141	0	0	0	0	7	0	0	1	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	135	3	0	0	1	1	0	1	2	6	0	0	0	1
N	0	0	0	0	0	0	0	1	0	0	0	148	1	0	0	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	2	0	0	0	0	148	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	2	0	0	0	0	0	0	0	0	0	152	6	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	8	0	0	2	0	0	139	0	1	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	141	4	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	0	1	0	9	0	0	0	0	0	142	8	0	0	0	0
V	0	0	0	0	0	0	0	0	0	2	7	6	0	0	0	0	0	0	0	135	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0
X	0	0	0	16	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	133	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	149	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	148

Fig. 13. Confusion matrix for TRA, with columns as the predicted class and the rows as the actual class.

Through analysis of the neural network's recall performances across the different letters, as well as observations during development, the following deficiencies of the Leap Motion controller were identified:

1) The Leap Motion's sensitivity falls when the hands are close together, which can cause failure to distinguish minute details (such as in 'V') or even failure to detect the presence of the hands altogether (such as in 'G'). The latter case can be easily rectified through repositioning the hands but this nevertheless has an impact on usability.
2) Occlusion causes the Leap Motion to make intelligent, though less than perfect, estimations on the position coordinates of occluded joints.

Both deficiencies may be limiting factors for the adoption of Leap Motion controllers in manual language acquisition. Despite this, the overall accuracy was considerably high, enabling practical usability as will be described next.

B. Usability Evaluation

1) Experimental Design: In answering ***RQ3***, the goal of the VR-NZSL application was to evaluate the effectiveness of a VR-based DEL-MaC implementation with regards to self-directed learning, memory retention, and engagement. As a benchmark for effectiveness, the Learn NZSL web-based NSZL acquisition platform [57] was used for comparison, which includes both Learn NZSL Alphabet videos [25] and Deaf Aotearoa 2D diagrams [54]. Although not strictly a DEL-MaC implementation, Learn NZSL is nevertheless an effective self-directed learning tool that is widely available for potential learners [25]. Participants in the evaluation included 10 hearing tertiary students and staff with no prior NZSL knowledge. The participants volunteered to take part in the study hence no sampling technique was used. Participants were instructed to learn eight letters (A, G, L, M, N, O, P, R) within the 360-video module. Afterwards, the participants performed the random letter task, the fruit-spelling task, and the name-spelling task (using the same set of eight letters). Once completed, the participants were tested on their memory retention.

After completing the VR-NZSL module, the participants were asked to learn a different eight set of letters (D, I, Q, S, T, U, V, Y) that are anatomically similar to the subset learned in VR-NZSL phase. The participants were instructed to memorize those letters, using both the Learn NZSL application and the alphabet diagram (from Fig. 5) at their discretion. Once completed, the participants were tested through being asked to sign the gestures in front of an observer who scored their performance. For both tests, the score for each letter is either *0:* cannot remember, *0.5:* remembered but with obvious anatomical deviations, and *1:* perfect. Hence, the total score for the test was out of eight points. Both sessions were time-boxed at 10 minutes. The participants were asked to complete a Likert-scale post-questionnaire to evaluate their satisfaction, as well as a set of qualitative questions.

2) Statistical analysis: Table VI indicates promising results, with all questions showing high responses for the various usability and perceived effectiveness aspects of VR-NZSL. The results from Q1 had the lowest mean score with the largest variability in responses, suggesting that the usability of the video controls have room for improvement. Buttons may not be the most intuitive interface with the Leap Motion controller in a VRLE, thus explorations into alternative video navigation methods that do not rely on buttons could be made. Despite this, participants commented on the usefulness of the 'loop' functionality, stating that it was useful for the difficult gestures that were hard to follow. It was also encouraging to find that the participants found the overall VR experience an immersive one; as immersion is a key aspect in having concrete experiences, the results show that VR-NZSL satisfactorily implemented the DEL-MaC framework. The results from Q3 and Q4 directly address *Timely Feedback* and *Retrospective Feedback* components respectively, further showing that VR-NZSL closely follows DEL-MaC. The results show *Timely Feedback* and *Retrospective Feedback* are essential mechanisms for the self-directed manual language learner, as well as the necessity for *Reflective Observation* in an Experiential Learning tool.

Table VII presents the self-reported confidence (Q5) and engagement (Q6) results for the three CALL methods. To see if participants felt any difference between the three CALL

Table VI
QUESTIONNAIRE RESULTS FOR VR-NZSL EXPERIENCE (5: STRONGLY AGREE, 4: AGREE, 3: NEUTRAL, 2: DISAGREE, 1: STRONGLY DISAGREE).

Question	SD	D	N	A	SA	$\bar{x}$	s
Q1. I found the 360-video easy to navigate and control	0	1	1	6	2	3.9	0.88
Q2. I found the VR environment immersive	0	0	0	6	4	4.4	0.52
Q3. I can learn NZSL more easily because I can see my hands in the VR environment	0	0	1	5	4	4.3	0.67
Q4. I found the application telling me when I am wrong or right beneficial to my learning	0	0	0	3	7	4.7	0.48

Table VII
COMPARING CONFIDENCE AND ENGAGEMENT FOR THE DIFFERENT CALL METHODS (5: STRONGLY AGREE, ..., 1: STRONGLY DISAGREE).

CALL method	Median	Range	
		Min	Max
Q5. I feel confident that I can learn NZSL alphabet using {CALL method} by itself			
VR-NZSL	5.0	4	5
Learn NZSL 2D videos	2.5	2	4
Deaf Aotearoa alphabet 2D diagrams	3.0	1	4
Q6. I felt engaged learning NZSL alphabet using {CALL method}			
VR-NZSL	5.0	4	5
Learn NZSL 2D videos	3.0	2	3
Deaf Aotearoa alphabet 2D diagrams	3.0	1	4

Table VIII
ENCODING OF POST-QUESTIONNAIRE QUALITATIVE QUESTIONS.

Question	Qualitative Encoding
Q7. What was your favorite activity? Why?	Fruit: Engagement (6), Interaction (4)
Q8. Did you prefer the 360-videos or the 2D videos? Why?	360-video: Engagement (6), Visual Breadth (4)
Q9. What were the pros and cons of the VR approach?	Pro: Immersion (4), Interaction (3), Usability (3)
	Con: Leap Motion Sensitivity (3), Video Navigation (3), Slow (2), Motion Sickness (2)
Q10. What were the pros and cons of learning NZSL using a 2D video?	Pro: Speed (6), Video Controls (4)
	Con: Hard to follow (5), No validation (3), Uninteresting (2)
Q11. What were the pros and cons of learning NZSL using 2D diagrams?	Pro: Speed (10)
	Con: Uninteresting (6), No validation (3), Hard to follow (1)

methods, a Friedman ANOVA repeated-measures test is used. The difference of the self-reported confidence (Q5) is significant across the three groups, $X_r^2 = 10.85$, $p = 0.00441$. The difference of the self-reported engagement (Q6) is even more significant across the three groups, $X_r^2 = 15.2$, $p = 0.0005$. Next, Wilcoxon signed-rank tests are performed between each pair of CALL methods to understand where the significant differences are arising. In the case of both confidence and engagement, it is not possible to calculate an accurate p-value for the differences between the 2D videos and 2D diagrams (i.e. insignificant differences). Only in the case of comparing VR-NZSL to 2D videos, or comparing VR-NZSl to 2D diagrams, does the Wilcoxon signed-rank test report significant differences: VR-NZSL vs 2D video confidence ($Z = -2.5205$, $p < 0.01$), VR-NZSL vs 2D diagram confidence ($Z = -2.6656$, $p < 0.01$), VR-NZSL vs 2D video engagement ($Z = -2.8031$, $p = 0.00512$), and VR-NZSL vs 2D diagram engagement ($Z = -2.8031$, $p = 0.00512$).

Based on Q5, VR-NZSL is perceived as an effective learning tool that can be used alone to learn NZSL. In comparison, participants commented that the Learn NZSL video was too quick and the perspective was second-person only, making it difficult to follow along. Additionally, the 2D diagrams were static images of the final positions of the gesture, hence participants commented that it was sometimes difficult to produce the gestures. The participants thus frequently switched their mode of instruction between the diagram and the video during the web-based session. The difficulty to follow gestures directly hinders the actively experimenting learner. Furthermore, answers for Q6 showed very strong evidence that the VR-NZSL was a more engaging method of learning, with participants scoring the application around 2 Likert points higher than the web-based approaches.

All participants found that the fruit activity was the most engaging task (Q7), commenting that having the ability to directly 'hold' objects within the environment helped them feel more engaged. This is not too surprising, as the task incorporated the greatest degree of interaction through carrying the fruit onto the bench to initiate the spelling game, thereby manifesting *Environmental Response*. The result supports the hypothesis that by maximizing interactivity, the learner feels a sense of power over their learning process, embodying a constructivist form of learning. For Q9, One participant noted:

I was having so much fun in the VR that I was motivated to keep learning and complete each task.

Many of the answers in Table VIII followed a similar theme—that engagement is a precursor to motivation. This could open up further research regarding gamification for manual language acquisition, which could exploit engagement as a means to motivate the self-directed learner. For the memory test, the scores attained by the participants for VR-NZSL versus the web-based session were not normally distributed, so a paired t-test could not be carried out. A Wilcoxon signed-rank test could also not be performed due to the small N size. As a result, it cannot be determined whether there is any difference in memory performance between the VR or web-based sessions. Despite inconclusive statistical significance, participants commented that visual recall through connections with fruit objects seemed to contribute to the ease of recalling gestures. Ultimately, promising evaluations in the self-reported usability, confidence, and engagement of participants in the VR-NZSL session serve to reliably answer ***RQ3***, despite no evidence of significant memory retention difference.

Threats to Validity: External validity regarding memory effectiveness is jeopardized by the experimental design of the memory test. Specifically, it is possible that the letters selected for testing in the VR setting are easier to remember than those learned in the web-based setting. Furthermore, participants generally took longer in the VR session than in the web-based session as they often voluntarily terminated the latter much earlier than the designated time-box of 10 minutes, possibly due to a loss in engagement. Time may have

affected the memory score of the participant. Furthermore, the limited sample size of 10 impacts the generalizability of the experiment. The internal validity of the experiment is also threatened, particularly with regards to the Likert-scale post-questionnaire. Participants could have rated the VR-NZSL higher than the web-based approaches by virtue of the Rosenthal effect. By being aware of the background of VR-NZSL, participants may have been biased towards the novelty of the project or influenced by the observer-expectancy effect. Due to the human ethics approval requirements, participation in the study was voluntary and prohibited collection of demographic information. Finally, results in terms of perceived effectiveness and incoming familiarity with NZSL were both self-reported.

VI. Conclusion

The lack of digitally-assisted pedagogical tools has contributed to a climate of manual language education that lacks availability and practical exercises [7]. The paper aims to tackle this issue by introducing an innovative digitally-assisted pedagogical framework as a means to acquire manual languages. A VR-based interpretation of this framework is manifested in VR-NZSL—a VRLE for the NZSL alphabet that uses the Leap Motion controller and machine learning as engines to drive corrective feedback. Despite limitations with the Leap Motion controller, the validation system has shown to be reliable, reaching accuracies of up to 99.0% for the static gestures of the NZSL alphabet. Through the evaluation of the classifier, reasonable accuracy could be reached within a two-hour recording time, indicating scalability towards expanding the classifiable vocabulary. The evaluation also identified the set of hand features required for accurate classification using the Leap Motion controller.

While previous researchers have investigated using the Leap Motion controller for other sign languages (predominantly the single-handed Arabic sign language), these studies were done without an emphasis on the pedagogical implications of the controller. This study focuses on the two-handed New Zealand sign language with detailed classification evaluations as well as preliminary evaluations on participants with no background in sign language. It is hoped this will assist future sign language recognition systems in pin-pointing the hand features necessary for accurate classifications. Through the implementation of the DEL-MaC framework, the VR application has been shown to motivate the self-directed learner by providing engagement and usability, suggesting that the framework could be a reliable foundation for VR-based education. The first implication of this study's findings reveals that there is still much-needed advancement of suitable hardware in order to more accurately capture intricate sign language gestures. Second, the accuracy differences between TRA and TPA suggests the existence of possible 'accents' when signing that must be accounted for during model training.

This research leaves room for future expansion in the following areas. As VR-NZSL is currently limited to the NZSL alphabet, expanding its recognizable vocabulary is a potential avenue for future expansion. This will involve time-dependent models so as to capture the sequence-based motions of the NZSL vocabulary. Another future expansion is the use of multiple Leap Motion cameras placed at different angles in order to remove the effects of occlusion—a key factor in decreasing the accuracy of the current NZSL iteration. While this research has shown promising results for simple single-alphabet gestures, this eventually needs to be extended to supporting vocabulary with increasing complexity, and gradually to recognizing the grammatical structure of manual communication. Through the approach proposed, it is believed that a valuable contribution has been made towards the digital revitalization of manual languages.

Acknowledgment

The authors would like to extend a sincere thanks to the NZSL professionals, Nichola Buisman and James Pole, for providing valuable feedback as well as providing test data to evaluate the classification system. Special thanks must also be given to the 10 usability participants.

References

[1] Statistics NZ, "Census QuickStats about culture and identity." http://archive.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-culture-identity/languages.aspx, 2013.

[2] A. Hynds, S. C. Faircloth, C. Green, H. Jacob, *et al.*, "Researching identity with indigenous d/deaf youth," *New Zealand Journal of Educational Studies*, vol. 49, no. 2, p. 176, 2014.

[3] R. L. McKee *et al.*, "The eyes have it! Our third official language: New Zealand Sign Language," *Journal of New Zealand Studies*, no. 4/5, 2005.

[4] R. McKee and V. Manning, "Evaluating effects of language recognition on language rights and the vitality of New Zealand Sign Language," vol. 15, pp. 473–497, 2015.

[5] H. R. Commission, "A new era in the right to sign: He houhanga rongo te tika ki te reo turi," *Report of the NZ Sign Language Inquiry*, 2013.

[6] J. Garretsen, "Big step forward for NZ Sign Language." https://www.radionz.co.nz/national/programmes/morningreport/audio/201796520/big-step-forward-for-nz-sign-language, 2016.

[7] P. P. Akmese, "Examination of sign language education according to the opinions of members from a basic sign language certification program," *Educational Sciences: Theory and Practice*, vol. 16, no. 4, 2016.

[8] M. Marschark, *Raising and educating a deaf child: A comprehensive guide to the choices, controversies, and decisions faced by parents and educators*. Oxford University Press, 2007.

[9] B. R. Schirmer, *Psychological, social, and educational dimensions of deafness*. Allyn & Bacon, 2001.

[10] M. Hauck and S. Hurd, "Exploring the link between language anxiety and learner self-management in open language learning contexts," *European Journal of Open, Distance and E-learning*, vol. 2005, no. 2, 2005.

[11] C. Lai, "A framework for developing self-directed technology use for language learning," *Language Learning & Technology*, vol. 17, no. 2, pp. 100–122, 2013.

[12] T. Rashid and H. M. Asghar, "Technology use, self-directed learning, student engagement and academic performance: Examining the interrelations," *Computers in Human Behavior*, vol. 63, pp. 604–612, 2016.

[13] T. Can, "Learning and teaching languages online: A constructivist approach," *Novitas-Royal*, vol. 3, no. 1, 2009.

[14] A. Y. Kolb and D. A. Kolb, *Experiential Learning Theory*, pp. 1215–1219. Boston, MA: Springer US, 2012.

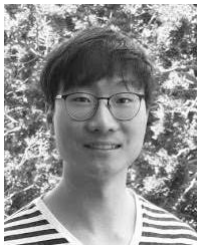
[15] D. A. Kolb, *Experiential learning: Experience as the source of learning and development*. FT press, 2014.

[16] C. Doughty, "Relating second-language acquisition theory to call research and application," 1982.

[17] P. Munday, "The case for using duolingo as part of the language classroom experience," *RIED. Revista iberoamericana de educación a distancia*, vol. 19, no. 1, 2016.

[18] A. K. Wehner, A. W. Gump, and S. Downey, "The effects of Second Life on the motivation of undergraduate students learning a foreign language," *Computer Assisted Language Learning*, vol. 24, no. 3, pp. 277–289, 2011.

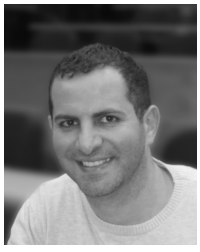
[19] H.-M. Huang, U. Rauch, and S.-S. Liaw, "Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach," *Computers & Education*, vol. 55, no. 3, 2010.
[20] J. R. Savery and T. M. Duffy, "Problem based learning: An instructional model and its constructivist framework," *Educational technology*, vol. 35, no. 5, pp. 31–38, 1995.
[21] C. J. Chen, S. C. Toh, and W. M. F. W. Ismail, "Are learning styles relevant to virtual reality?," *Journal of research on technology in education*, vol. 38, no. 2, pp. 123–141, 2005.
[22] V. Kohonen, "Experiential language learning: second language learning as cooperative learner education," *Collaborative language learning and teaching*, vol. 1439, 1992.
[23] W. Vicars, *ASL Online: The design and implementation of a web-based American Sign Language course*. Lamar University-Beaumont, 2003.
[24] C. L. Radford, *Exploring the Efficacy of Online American Sign Language Instruction*. ERIC, 2012.
[25] S. Pivac Alexander, M. Vale, and R. McKee, "E-learning of New Zealand Sign Language: Evaluating learners' perceptions and practical achievements," vol. 23, pp. 60–79, 11 2017.
[26] N. Adamo-Villani, E. Carpenter, and L. Arns, "3D sign language mathematics in immersive environment," *Proc. of ASM*, 2006.
[27] L. Ying, Z. Jiong, S. Wei, W. Jingchun, and G. Xiaopeng, "VREX: Virtual reality education expansion could help to improve the class experience (VREX platform and community for VR based education)," in *Frontiers in Education Conference (FIE)*, pp. 1–5, IEEE, 2017.
[28] Y.-L. Chen, "Virtual reality software usage in an EFL scenario: An empirical study," *JSW*, vol. 9, no. 2, pp. 374–381, 2014.
[29] K. Schwienhorst, "Why virtual, why environments? implementing virtual reality concepts in computer-assisted language learning," *Simulation & Gaming*, vol. 33, no. 2, pp. 196–209, 2002.
[30] A. Cheng, L. Yang, and E. Andersen, "Teaching language and culture with a virtual reality game," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 541–549, ACM, 2017.
[31] I. Kožuh, S. Hauptman, P. Kosec, and M. Debevc, "Assessing the efficiency of using augmented reality for learning sign language," in *International Conference on Universal Access in Human-Computer Interaction*, pp. 404–415, Springer, 2015.
[32] B. Khelil and H. Amiri, "Hand gesture recognition using leap motion controller for recognition of Arabic sign language," in *3rd International conference ACECS'16*, 2016.
[33] M. Mohandes, S. Aliyu, and M. Deriche, "Arabic sign language recognition using the leap motion controller," in *Industrial Electronics (ISIE), 2014 IEEE 23rd International Symposium on*, pp. 960–965, IEEE, 2014.
[34] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, 2017.
[35] E. von Glasersfeld, "The reluctance to change a way of thinking," *The Irish Journal of Psychology*, vol. 9, no. 1, pp. 83–90, 1988.
[36] D. H. Jonassen, "Designing constructivist learning environments," *Instructional design theories and models: A new paradigm of instructional theory*, vol. 2, pp. 215–239, 1999.
[37] J. Steuer, "Defining virtual reality: Dimensions determining telepresence," *Journal of communication*, vol. 42, no. 4, pp. 73–93, 1992.
[38] T. B. Sheridan, "Interaction, imagination and immersion some research needs," in *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 1–7, 2000.
[39] G. C. Burdea and P. Coiffet, *Virtual reality technology*. John Wiley & Sons, 2003.
[40] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence*, vol. 7, no. 3, 1998.
[41] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.
[42] S. Krashen, "We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis," *The modern language journal*, vol. 73, no. 4, pp. 440–464, 1989.
[43] K. Emmorey, R. Bosworth, and T. Kraljic, "Visual feedback and self-monitoring of sign language," *Journal of Memory and Language*, vol. 61, no. 3, pp. 398–411, 2009.
[44] V. A. Postovsky, "Effects of delay in oral practice at the beginning of second language learning," *The Modern Language Journal*, vol. 58, no. 5-6, pp. 229–239, 1974.
[45] J. G. Nicholls, "Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance," *Psychological review*, vol. 91, no. 3, p. 328, 1984.
[46] A. Y. Kolb and D. A. Kolb, "Learning styles and learning spaces: Enhancing experiential learning in higher education," *Academy of management learning & education*, vol. 4, no. 2, pp. 193–212, 2005.
[47] M. Swain, "The output hypothesis and beyond: Mediating acquisition through collaborative dialogue," *Sociocultural theory and second language learning*, vol. 97, p. 114, 2000.
[48] D. J. Young, "An investigation of students' perspectives on anxiety and speaking," *Foreign Language Annals*, vol. 23, no. 6, pp. 539–553, 1990.
[49] J. Hattie and H. Timperley, "The power of feedback," *Review of educational research*, vol. 77, no. 1, pp. 81–112, 2007.
[50] H. Holec, *Autonomy and foreign language learning*. ERIC, 1979.
[51] B. Opitz, N. K. Ferdinand, and A. Mecklinger, "Timing matters: the impact of immediate and delayed feedback on artificial language learning," *Frontiers in human neuroscience*, vol. 5, p. 8, 2011.
[52] R. J. Nadolski and H. G. Hummel, "Retrospective cognitive feedback for progress monitoring in serious games," *British Journal of Educational Technology*, vol. 48, no. 6, pp. 1368–1379, 2017.
[53] Leap Motion, Inc, "Leap Motion goes mobile." https://developer.leapmotion.com/android/#107, 2019.
[54] Deaf Aotearoa, "NZSL Alphabet." http://deaf.org.nz/nzslw-resources, 2019.
[55] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
[56] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.
[57] NZSL Board, "Learn NZSL." http://www.learnnzsl.nz, 2019.



Edison Rho completed a Bachelor of Engineering Honours degree (conjoint with a Bachelor of Science) from the University of Auckland, New Zealand. He is currently working as a software engineer at Atlassian. His development interests include exploring the applications of machine learning in online software services.

Kenney Chan is a Software Engineer, having completed his Bachelor of Engineering Honours degree with the University of Auckland, New Zealand. He is currently working as a software developer at Datacom. His interests include applications of machine learning in educational methodologies and graphic design tools.

Elliot John Varoy completed his Bachelor of Engineering Honours and Master of Engineering (both Software Engineering) with the University of Auckland, New Zealand. He is currently a doctoral student focusing on the application of virtual reality within educational settings. His interests include STEM and computational thinking education.



Nasser Giacaman is a Senior Lecturer in the Department of Electrical, Computer, and Software Engineering at the University of Auckland, New Zealand. His disciplinary research includes parallel programming, with current research focusing on digital solutions across a number of different educational domains.