

SENSORY GLOVE SYSTEM AND METHOD FOR SIGN GESTURE SENTENCE RECOGNITION

Technical Field

5

The present invention relates, in general terms, to a sensory glove system for detecting sign gestures, and a method for sign gesture sentence recognition.

Background

10

Sign language recognition, especially sentence recognition, is significant in lowering the communication barrier between the hearing/speech impaired (signers) and those who do not use sign language (non-signers). Specially designed gloves are sometimes used to detect motion of the hands. Existing
15 glove solutions can only recognise discrete single gestures (i.e., numbers, letters, or words). This loses context when those single gestures are part of a sentence, and therefore do not satisfy signers' daily communication needs.

Benefiting from a number of attractive features such as light weight, good
20 compliance, and comfort, wearable sensors hold great promise in various applications spanning from environmental monitoring, personalized healthcare to soft robotics, and human-machine interaction. Typical wearable sensors rely on resistive and capacitive mechanisms for human status tracking and surrounding sensing. Such sensors often use an external power supply to
25 generate excitation signals, inhibiting the self-sustainability and independence of glove solutions.

Being closely associated with sophisticated hand gestures and as an essential part of biomedical care, sign language interpretation is of substantial
30 importance. This is particularly the case since the general public typically cannot sign. Generally, visual images/videos, surface electromyography (sEMG) electrodes, and inertial sensors are conventional means to reconstruct hand gesture information towards sign language recognition applications. These communication solutions are limited by light conditions and privacy concerns,
35 electromagnetic noise and crosstalk with other biopotentials, or a huge amount

of data.

It would be desirable to provide a system that can recognise sign gestures in a more comprehensive and natural way than prior art solutions, that can
5 recognise and translate signed sentences and overcomes the drawbacks of, or at least to provide a useful alternative to, the above systems.

Summary

10 The present disclosure leverages triboelectric nanogenerator (TENG) based wearable sensors. TENGs are simple to fabricate, allow wide material choice, and expeditious dynamic response, are power-compatible and self-sustainable. The present disclosure uses TENGs as the basis for a human-machine interface (HMI). HMIs are often resistive or capacitive, which results in the need for a
15 power supply. However, the proposed TENGs-based solutions provide self-powered triboelectric HMIs. When incorporated into a glove as taught herein, TENGs-based HMIs can seamlessly detect the multiple degrees-of-freedom motions of the dexterous hands, enabling sign language interpretation and system control command recognition.

20

Complex gestures and sentences are interpreted using the present TENGs gloves by applying artificial intelligence (AI) to outputs of TENGs sensors. The use of AI, when combined with the present sensor arrangement, allows extraction of features from which more diversified and complex gesture
25 monitoring is achievable when compared with direct feature extraction – e.g. peak and amplitude detection.

Following from this, presently disclosed is an artificial intelligence enabled sign language recognition and communication system comprising sensing gloves,
30 deep learning block, and virtual reality interface.

Disclosed is a sensory glove system for recognition of sign gestures, comprising:
a first glove for wearing on a first hand of a user, the first glove comprising:
35 a first wrist motion sensor for wrist motion perception of a first

- wrist;
- a finger bending sensor for each finger (including thumb) of the first glove; and
- a palm sensor; and
- 5 a second glove for wearing on a second hand of the user, the second glove comprising:
 - a second wrist motion sensor for wrist motion perception of a second wrist;
 - a finger bending sensor for each finger of the second glove;
 - 10 and
 - two fingertip sensors, one on each of a middle and index finger.

Also disclosed is a method for sign gesture sentence recognition from an input signal comprising a sentence, comprising:

- 15 dividing the input signal into fragments;
 - applying a machine learning model to the fragments to recognise individual said fragments; and
 - inversely reconstructing the sentence to recognise the sentence.
- 20 The fragments may include intact word signals, incomplete word signals and background signals, and wherein the applying the machine learning model to the fragments comprises:
- determining, based on a predetermined threshold number of data points associated with the fragment, if the fragment corresponds to background noise;
 - 25 discarding the fragment if the fragment is determined to correspond to background noise; and
 - recognising each fragment that has not been discarded.

- For each fragment comprising background signals, the method may further
- 30 comprises recognising the fragment as comprising background signals on the basis of the fragment comprising fewer than 100 non-zero data points, and tagging the fragment as "empty". In this context, a non-zero data point is a data point with a value of "0" or sufficiently close to zero that the data point does not constitute part of a word.

Also disclosed herein is a virtual reality conversation system comprising:

a sensory glove system as described above;

a signer terminal for use by a first user wearing the sensory glove system;

5 a non-signer terminal for use by a second user;

at least one processor configured to:

receive an input signal comprising an output from each wrist motion sensor, bending sensor, palm sensor and fingertip sensor;

apply one or more machine learning models to the input signal to

10 recognise a sentence in the input signal; and

project the signal into a virtual reality environment, for display at the non-signer terminal.

Unless context dictates otherwise, the following terms shall be given the
15 meaning set out here: "finger" shall be construed to include the fingers and thumb of the relevant hand; "known sentence", "known word" and similar shall be construed to include words known to a machine learning model due to their appearance in a training data set upon which the machine learning model was trained; "projected into virtual space" shall be construed to include displaying
20 one or more words or a sentence in a virtual reality environment in association with an avatar of the user who inputted the one or more words or sentence; and "inversely reconstruct" means after an AI model recognizes all word elements split from sentence samples (i.e. fragments), it (the model) will reversely to recognize the sentences .

25

Advantageously, embodiments employing machine learning use non-segmentation and segmentation assisted deep learning models to achieve recognition of words and sentences. Significantly, the segmentation approach splits entire sentence signals into word units, referred to herein as "fragments"
30 or "word elements". Then the deep learning model recognizes all word units and inversely reconstructs and recognizes sentences. Furthermore, new sentences – i.e. sentences not yet encountered by the machine learning or AI model – created by a new-order or combination of word elements can be recognized with an average accuracy of 86.67%. Finally, in embodiments providing a VR
35 conversation system the sign language recognition results can be projected into

virtual space and translated into text and audio, allowing the remote and bidirectional communication between signers and non-signers.

Brief description of the drawings

5

Embodiments of the present invention will now be described, by way of non-limiting example, with reference to the drawings in which:

Figure 1 schematically illustrates a communication system for facilitating
10 communication between signers (those who use sign language) and non-signers (those who do not use sign language), whether or not those communicating are at a distance from each other;

Figure 2 shows the detailed location, area information and channel label of
15 sensors on gloves;

Figure 3 shows the data analysis for signals of words and sentences;

Figure 4 demonstrates word and sentence recognition based on the non-
20 segmentation method;

Figure 5 shows the train and validation accuracy increase with epochs;

Figure 6 demonstrates word and sentence recognition based on segmentation
25 method;

Figure 7 illustrates division of input signal, comprising a sentence, into fragments, identification of fragments and reconstruction (i.e. inverse reconstruction) of the sentence;
30

Figure 8 shows the recognition result of three new sentences;

Figure 9 is a radar comparison map of non-segmentation and segmentation recognition methods based on their pros and cons; and
35

Figure 10 comprises a demonstration of communication between the speech impaired and the non-signer.

Detailed description

5

Described below are sensory glove systems with sensor configurations that are optimised to minimise the number, and in some cases the area, of sensors required to determine sign language signals. When combined with AI models for interpreting the outputs of the TENGs sensors, the systems described herein
10 can interpret a large number of sign language words and sentences based on known words and sentences, but may also construct new sentences.

TENG gloves have been demonstrated for monitoring finger motions using magnitude analysis or pulse counting. However, their data analytics are mostly
15 based on direct feature extraction (e.g., amplitude, peak number). This leads to a limited variety of recognizable hand motions/gestures with substantial feature loss. Sophisticated hand gesture discrimination remains challenging. Such limitations are not present in some sensory glove systems of the present disclosure due to the placement of sensors and the manner in which sensor
20 outputs are processed.

In addition, previous attempts at using AI technology for interpreting sign signals have been limited in the number of sign language gestures they can understand. Previous approaches lack an effective and practical approach for
25 real-time sentence recognition of sign language which is more significant for the practical communication of signers and non-signers. Besides, the interfaces (e.g., mobile phone or PC) that sign language recognition results are projected into or displayed on usually do not allow the signer's interaction with non-signers – i.e. they are one-way communication systems. Some systems
30 disclosed herein enable two-way communication through a VR inspired communication system. Here, we show a sign language recognition and communication system comprising triboelectric sensor integrated gloves, AI block, and the VR interaction interface. The system successfully realizes the recognition of 50 words and 20 sentences and can be expanded to a greater
35 number of word sand sentences, even learning from new sentences – i.e.

sentences that did not form part of the training dataset but that are recognised by the segmentation approach described below, by identifying all the known words in the new sentence. The recognition results are projected into virtual space in the forms of comprehensible voice and text to facilitate barrier-free
5 communication between signers and non-signers

Figure 1 shows a virtual reality conversation system 100. The system 100 comprises a sensory glove system 102 as described with reference to Figure 2. A user wearing the sensory glove system 102 interacts –by direct typing and/or
10 through commands or input signals issued by the sensory glove system 102 – with a signer terminal 104. Input signals received from the sensors of the sensory glove system 102 are processed by a processor in at least one of (e.g. it may be a distributed processor) the sensory glove system 102 itself, the signer terminal 104 (processor 106), a non-signer terminal 108 that is used by
15 a second user (processor 116), and a central server 110 (processor 118). Instructions (i.e. program code) 112 are sent to the processor from memory 114, which may be in the signer terminal 104 as shown, or elsewhere in the system 100.

20 The processor, of which there may be more than one, is configured to implement a method 120. The method 120 involves receiving an input signal comprising an output from each sensor of the sensory glove system 102 (step 122), applying a machine learning model, or multiple models as needed, to the input signal to recognise a sentence in the input signal (step 124), and projecting the
25 signal into a virtual reality environment, for display at the non-signer terminal (step 126). The VR environment of VR space may be hosted by one of terminals 104 and 108, or may central server 110. Notably, particularly where the VR environment involves a large number of parties in addition to those communication through terminals 104 and 108 – as represented by nodes 128,
30 130, which are two of potentially very many – it can be useful for the VR space to be hosted by a central server 110 to facilitate flexible resourcing. All terminals may communicate over network 132, which can be any suitable network.

Glove configuration and sensor characterization

To measure as many gestures as possible with a small number of sensors – i.e. parsimonious sensor usage – the sensor positions should be optimized. By referring to the frequently used sign language in the American Sign Language guide book, an analysis was conducted for involved motions in daily sign expressions of the speech/hearing impaired. It was determined that sign language includes three major motion dynamics, including elbow/shoulder motions, face muscle activities, and hand movements. Hand motion dominated the other motion dynamics, accounting for 43% of communication. Analysis of individual finger, wrist, palm and fingertip motions were then analysed, with communication relying on finger bending (56%), wrist motion (18%), touch with fingertips (16%), and interaction with palm (10%). With reference to Figure 2, image (b), sensors were therefore placed to optimise communications while minimising the number of sensors. Correspondingly, in the glove system 200, a sensor 202 (e.g. a triboelectric sensor) is mounted on each finger for finger bending measurement, and two sensors 204 are put on wrists for wrist motion perception. The finger motion sensors and/or wrist motion sensors are positioned on the outside of the respective finger and wrist. The index and middle fingertips of the right hand are in frequent use, and the left hand is often used to interact with the right hand and other parts of their body. Thus, two sensors 206 are placed on the fingertips of right index and middle of the glove, and one sensor 208 is incorporated on the palm of the left hand. Sensor placement on the gloves will be described with reference to left and right gloves. However, the sensors of the left glove may instead be provide on the right glove and vice versa, depending on user preference and hand dominance.

While a palm sensor could be placed on each glove, only one palm sensor is allocated for present purposes. The reduction in sensors assists with minimising the number of sensors and system complexity, while maintaining gesture detection accuracy. The left palm is often open and facing up at the end of a gesture. Therefore, the left palm usage, and thus left palm sensor measurement, is likely more influential in gesture detection than the right palm.

With reference to Figure 2, image (a), the gloves of system 200 are thus configured with 15 triboelectric sensors in total. The area of sensors is optimised based on key parameters that largely influence the triboelectric voltage output

performance, such as sensor area, force, bending degree, and bending speed. The voltage output increases when sensor area, force, bending degree and bending speed increase. With reference to Figure 2, image (c), where the sensors are triboelectric sensors (noting that capacitive, resistive or other sensors may be used despite being generally less desirable due to power consumption requirements), the triboelectrification layers are composed of ecoflex and wrinkled nitrile. With the conductive textile as electrodes, a flexible and thin triboelectric sensor is fabricated.

More particularly, TENG sensors may comprise 00-30 Ecoflex with a 1:1 weight ratio of part A and part B coated on a conductive textile (for electrode) for consolidation. Then the positive wrinkled nitrile is attached on the conductive textile contacts with the negative Ecoflex layer to assemble the triboelectric sensor. Finally, the conductive electrodes are encapsulated by non-conductive textiles to shield the ambient electrostatic effect. Different sensors with different areas are fabrication for customized glove configuration.

The glove itself may comprise the abovementioned two-layered triboelectric sensors sewed into a small cotton textile pocket. The finalized encapsulated triboelectric sensors for each hand position attached to the glove by E7000 textile glue for a seamless fit with fingers.

As the user makes gestures with the sensory glove system 200, input signals are generated by the sensors – an input signal comprising the output signals of all sensors. This can include sensors of parts of the hand that are not involved in the gesture. The input signal is then received (per step 122) and processed according to step 124. Processing can occur in various ways. Discussed with reference to Figures 3 and 4 is a non-segmentation method, whereby the input signal is processed by a machine learning (i.e. AI) model trained using data comprising various known words and sentences. The non-segmentation method effectively compares the input signal with the training data to determine which known sentence or word is most similar, or correlates best, to the input signal. The non-segmentation non-segmentation AI framework described below achieves high accuracy for 50 words (91.3%) and 20 sentences (95%) by independently identifying word and sentence signals. Naturally, this can be

extended to cover a greater number of words and sentences. To overcome the issue of unable to recognizing new sentences, the developed segmentation AI frame splits the entire sentence signal into word units (fragments) and then recognizes all the signal fragments. Inversely, the sentence can be reconstructed and recognized with an accuracy of 85.58% upon established correlation between word units and sentences. In some embodiments, this requires the fragments to represent known words. Thus, the segmentation approach may comprise dividing the input signal into fragments, applying a machine learning or AI model to each fragment to recognise one or more words in the fragment, and reconstructing the sentence from the words reflected in the fragments. The segmentation approach can be employed to recognise new sentences (average correct rate: 86.67%) that are created by word elements or fragments combined in a new order. The new sentence can be stored and, if repeatedly encountered, used as training data for the machine learning model employed in the non-segmentation method.

To train the sensory glove system 102, system 100 or terminal 104, to perform the non-segmentation approach and segmentation approach, the sensory glove system or terminal to which it is connected may comprise a signal acquisition module (e.g. as part of processor 106) that acquires the generated triboelectric signals generated by different gestures – e.g. using Arduino MEGA 2560 microcontroller. For the non-segmentation method, during training 100 samples for each word and sentence are collected where 80 samples (80%) for training and 20 samples (20%) for testing. Thus, there are 5000 samples for total 50 words and 2000 samples for total 20 sentences. For training the segmentation approach, since all the data form sentences, there are 50 samples with number series labels for each sentence (17 sentences and 850 samples in total). The division window in the Matlab workspace with the length of 200 data points slides at a step of 50 data points to extract the signal fragments – in some embodiments each data point may be zero or non-zero, and thus constitute background signal or potentially part of a gesture or word. Thus, each sentence sample with 800 data points is segmented into 13 elements. The extracted data elements are labelled with numbers of corresponding words or empty. Each number represents a signal fragment that may be an intact gesture signal, background noise, or an incomplete gesture signal (e.g. part of a word in sign

language). 60% of samples for each number (i.e., 0-19) are used for training, 20% for validation, and 20% for testing. Finally, 5 samples for 3 new sentences are employed to verify the capability of new sentence recognition of the segmentation-assisted CNN model without prior training.

5

Data analysis of signals of 50 words and 20 sentences

The tentative data analysis is beneficial for the preliminary understanding of the raw data. Firstly, gestures were selected based on their daily use in a signer's life. These gestures are reflected in Figure 3, image (a). Using the sensory glove
10 system 200, corresponding triboelectric signals of these gestures are given in Figure 3, image (b). These triboelectric signals are the input signals for each respective word.

The system 100 then applies one or more machine learning models to the input
15 signal, to recognise a sentence in the input signal. In the present example, the sentence comprises a single word in each case. The system 100 recognises the words by applying the machine learning model to perform signal similarity and correlation analysis against the original data. In the example, there are 100 data samples for each gesture in the dataset, where 'Get' is shown in the
20 enlarged view in the below middle of Figure 3, image (b). For correlation analysis, the below left of Figure 3, image (b) shows a new input 300 of 'Get' compared with its own database 302, from which a mean correlation coefficient of 0.53 is calculated using any appropriate means – e.g. least squares error between data points (x-axis) of the new input signal over the average values of
25 the database signal. 0.53 is above a predetermined threshold (presently 0.3) that is the nominal cut-off for a high correlation. The same can apply to processing multi-word sentences. To produce the database 'Get' or database word in each case, the training samples may be averaged to enable correlation with the new input signal. Also, the signal of 'Must' is compared with the 'Get'
30 database. The mean correlation coefficient of 0.37, between the input signal 300 and the database signal for 'Must' 304, is also larger than 0.3. This shows confusion by the system, since both 'Get' and 'Must' indicate high signal similarity (the bottom right of Figure 3, image (c)). Therefore, the traditional basic recognition algorithm (e.g., correlation analysis mentioned here) may not
35 be suitable for a discrimination between individual gestures where there is a

large number of gestures between which is must discriminate. This also means the recognising individual gestures using traditional methods can result in nonsensical sentences – e.g. gestures representing "Get the doctor" may be recognised as "Must the doctor". To solve this issue, present methods employ
5 advanced data analytics such as machine learning to get a better recognition accuracy, and perform full sentence recognition against a database of known sentences.

Figure 3, image (c), is a matrix representation of the correlation coefficient
10 between input signals (i.e. gestures corresponding to individual words) of known words and database signals for all other known words. The same information is reflected in Figure 3, image (d), being a distribution curve of correlation coefficients from which it can be observed that signals of several gestures have a strong correlation. This means there is a high similarity among these gesture
15 signals. These gestures are susceptible to incorrect classification. The same is shown for sentences in daily use, in Figure 3, images (e) and (f).

Word and sentence recognition upon non-segmentation method

For sequence modelling of signals, hidden Markov model (HMM), recurrent
20 neural network (RNN), and more recently developed long short-term memory (LSTM) have been widely utilized owing to enabling memorizing the output of the last moment for circulated self-updating and adapting. Convolutional Neural Networks (CNN) have also been designed to process data in the form of multiple arrays. In particular, a 1-dimensional (1D) CNN is a simple and feasible solution
25 for recognizing the time-series signals of human motions from triboelectric sensors. Because the positions of features in the segment are not highly correlated. To optimize the CNN model toward more effective recognition behaviours, adjustment of kernel size, the number of filters, and convolutional layers is implemented. To optimise the CNN for use with the present sensory
30 glove system in detecting and interpreting gestures signals, box plots were used to assess kernel size (Figure 4, image (a)), filter number (Figure 4, image (b)), and number of convolutional layers (Figure 4, image (c)). The box plots indicate median (middle line), 25th, 75th percentile (box) and 5th and 95th percentile (whiskers) as well as outliers (single points). The machine learning model used
35 to assess similarity or correlation between input signal and known words or

sentences may comprise a 1D CNN model with 5 kernels, 64 filters, and 4 convolutional layers. The detailed CNN structure parameters are presented in Table 1.

No	Layer Type	No. of Filters	Kernel/ Pool Size	Stride	Input Size	Output Size	Padding
1	Convolution 2	64	5	1	(None, 100, 32)	(None, 100, 64)	same
2	Max Pooling 2		2	2	(None, 100, 64)	(None, 50, 64)	same
3	Convolution 3	128	5	1	(None, 50, 64)	(None, 50, 128)	same
4	Max Pooling 3		2	2	(None, 50, 128)	(None, 25, 128)	same
5	Convolution 4	256	5	1	(None, 25, 128)	(None, 25, 256)	same
6	Max Pooling 4		2	2	(None, 25, 256)	(None, 13, 256)	same
7	Convolution 5	512	5	1	(None, 13, 256)	(None, 13, 512)	same
8	Max Pooling 5		2	2	(None, 13, 512)	(None, 7, 512)	same
9	Flatten				(None, 7, 512)	(None, 3584)	same
10	Dense (500)				(None, 3584)	(None, 500)	
11	Dense (50)				(None, 500)	(None, 50)	

Table 1: CNN parameters. The detailed parameters for constructing Convolutional Neural Network (CNN).

The CNN is schematically represented in Figure 4, image (d), in which the 15-channel signals are served as inputs without segmentation. In other words, the sentence signals are not pieced into word elements and not linked with the basic word units within a non-segmented deep learning framework. The signals of words and sentences are isolated from each other and independently recognized.

To better understand the clustering performance of the proposed CNN structure, the results of last fully-connected layer is extracted in which each data sample with 3000 points finally has been stretched to 3584 after convolution and pooling to compare with the raw data. The present machine learning model may use any dimension reduction method, such as Principal Component Analysis (PCA), to reduce the dimension of the output data while maintaining information. PCA can be performed using eigenvalues and eigenvectors of an analogue signal matrix of gestures, which can point along the major variation directions of data. The dimension of data could be reduced with adjustable principal component matrix which comprises the non-unique eigenvectors of the data matrix. Eventually, the data of input layer and the last fully-connected layer is reduced to 30 dimensionalities after PCA.

The detailed mathematics behind PCA

PCA is used to reduce the dimensions whilst retaining maximum information. The principle of PCA relies on the correlation between each dimension and provides a minimum number of variables which keeps the maximum amount of variation about the distribution of original data. PCA employs the eigenvalues and eigenvectors of the data-matrix which can point along the major variation directions of data to achieve the purpose of dimension reduction. For the detailed mathematics of PCA, the vector X includes the component that is the input signal x_i for each gesture,

10

$$X = \{x_1, x_2, x_3, \dots, x_n\} \quad (1)$$

Then the mean value X_{mean} of X is calculated as,

15

$$X_{\text{mean}} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2)$$

Determining the difference d_i between the input and mean value,

20

$$d_i = x_i - X_{\text{mean}} \quad (3)$$

Based on Equation (3), we get the covariance matrix S ,

$$S = \frac{1}{n} \sum_{i=1}^n d_i d_i^T \quad (4)$$

25 According to linear algebra, the eigenvalues λ and eigenvectors p of the covariance matrix are defined as,

$$Sp = \lambda p \quad (5)$$

30 Because there are more than one the eigenvalues and eigenvectors, the principal component matrix P is,

$$P = \{p_1, p_2, p_3, \dots, p_k\} \quad (6)$$

Thus, the input signal can be projected into a new output matrix Y with reduced dimensions,

$$Y = \{y_1, y_2, y_3, \dots, y_i, \dots, y_n\} \quad (7)$$

5

Where $y_i = P^T(x_i - x_{\text{mean}})$ and k controls the principal component matrix P and hence controls the dimension of output Y .

Figure 4, image (e), shows feature clustering result of 50 words for the input layer. Feature clustering can be achieved using any known method. Presently, it was achieved by the t-distributed Stochastic Neighbor Embedding (t-SNE) in the dimension of principal component 1 (PC1) and principal component 2 (PC2). Figure 4, image (e) shows the performance of feature clustering is poor with category overlaps before going through the CNN network. After undergoing the feature extraction and classification of CNN, the visualization result realized by t-SNE in Figure 4, image (f) indicates a desirable classification performance of the developed CNN model with clear boundaries between these 50 classes with less overlap. The same can be said of sentence recognition, as reflected in Figure 4, images (h) and (i) respectively.

20

Further in this regard, as shown in Figure 5 after 50 training epochs, the accuracy reaches almost 90%. As a consequence, upon 80 training samples (80%) and 20 test samples (20%) of each word in the dataset, a high recognition accuracy of these 50 words is achieved with the value of 91.3%, as shown in the confusion matrix in Figure 4, image (g) for words and image (j) for sentences. The long-length signal of sentence provides more distinguishable features than the single word signal. Therefore, a high sentence recognition accuracy of 95% is obtained.

30 Though with high accuracies in discriminating words and sentences, the CNN model upon non-segmentation approach only enables classifying known sentence signals in the dataset. It cannot identify new (i.e. not before seen) sentences even where a sentence consists of known words, if those known words are in an order not previously seen by the system 100. This limitation results from labelling each word or sentence with a single label as a distinct item

35

– i.e. the CNN does not distinguish between word and sentence for analysis purposes. In addition, there is time latency in processing, particularly for large sentences, because the CNN model employing non-segmentation methods needs to leverage the whole long-data-length signal (generally 200 data points for the word signal and 800 data points for the sentence signal) to trigger the recognition procedure. This makes it undesirable for the real-time sign language translation.

Word and sentence recognition upon segmentation method

To explore the feasibility of recognizing new/never-seen sentences, a segmentation method is described herein. The segmentation method involves dividing the input signal into fragments, analysing the fragments to identify words, and inversely reconstructing the sentence based on the identified words. The analysis step may be twofold – firstly, distinguishing between fragments representing background noise and those representing words (i.e. filtering) and, secondly, performing recognition on fragments that represent words.

Sentences may be divided into fragments using a data sliding window. Each window may comprise an intact word signal, incomplete word signal or background signal (i.e. noise). These fragments split from all sentence signals are recognized first, and then the CNN model will inversely reconstruct and recognize the sentences. In this way, the segmentation approach can identify known sentences but also identify or interpret new sentences from known words. Since the input signal can be divided as soon as data comprising the 'next' window is received, the input signal can be processed progressively. Consequently, the recognition latency is significantly reduced due to small-distance sliding window length (50 data points per sliding) as the recognition process will be triggered with each sliding. Any new sentences can be used to expand the sentence database. Therefore, the machine learning models only need to be trained on words, with the segmentation method enabling individual words to be detected and sentences constructed in the database – e.g. memory 132 of Figure 1.

Figure 6, image (a) shows 19 words (W01-W19) labelled from 0 to 18 based on usage frequency. Background noise is considered an 'empty' word (W20)

labelled 19. The label action is a critical step for the supervised learning of the CNN model. Figure 6, image (b) demonstrates the detailed process of sentence signal division and labelling. The system 100 takes the sentence 'The dog scared me'. It analyses the voltage on each channel (i.e. the output of each sensor, as reflected in the input signal) as detected at many points in time, the x-axis reflected 800 data points for the sentence. Using a 200 data point sliding window and a sliding step 50 data points, the sentence becomes 13 fragments. Each fragment is an intact single word, background signal, and mixture signal (i.e., incomplete word signals and background signals). When the sliding window contains the intact word signal, then this fragment signal will be labelled with the number of the corresponding word as shown in Figure 6, image (a). The background noise is labelled as 'empty' (W20) with number 19. For mixed signals, the label is assigned based on the principal component. If word signal or background noise accounts for more than 50% of the sliding window size, the label will be the word number of the word in the fragment, or 'empty' 19, respectively. 'The dog scared me' is represented by the label number series '[5 5 19 19 19 15 15 15 19 19 0 0 0]'. The table in Figure 6, image (c) summarizes the essential information of 20 sentences: sentence category codes (Y01-Y17 and New1-New3), and word/gesture components, an example label series, and unique number orders derived from the label series minus background signals. New1, New2 and New3 are unknown sentences used to verify the potential of recognizing new sentences. There is no corresponding label series for these three sentences.

There are two ways in which the segmentation method was employed. The first involved detecting words from fragments in a single pass. The second involved a hierarchical filter for separating background signals from word fragments, and then performing recognition on the word fragments. The single classifier was built to identify all the fragments from 17 sentences without preliminarily filtering empty signals (Figure 6, image (d)). The dataset contains 50 samples for each sentence and hence 850 (50*17) sentence samples in total. The word fragments are from 510 samples for training (60%), 170 samples for validation (20%), and 170 samples for testing (20%). The confusion matrix in Figure 6, image (e) shows an accuracy of 81.9% for the single classifier recognizing word units W01-W20. In the segmentation method, the sentence is inversely

reconstructed – i.e. constructed from recognised word fragments to form a text or audio output. As indicated in Figure 6, image (f) the result mapping of sentence recognition illustrates an accuracy of 79.41% in sentence reconstruction for known sentences. This is too poor an accuracy, particularly
5 for Y01 and Y04. This results from the system endeavouring to classify empty words (i.e. W20). A hierarchical classifier is shown in Figure 6, image (g). The hierarchical classifier 600 has a first-level classifier 602 to separate background signals from intact word signals (i.e. fragments comprising an intact word or at least 50% data points relate to a word), and a second-level classifier 604 for
10 precise identification of word signals. The recognition accuracy of the hierarchical classifier is 82.81% for word fragments – the related confusion map is shown in Figure 6, image (h). Sentence recognition was also higher, at an average accuracy of 85.58% as shown in Figure 6, image (i).

15 The method employed by system 100 may therefore comprise dividing the input signal into fragments that include intact word signals, incomplete word signals and background signals. The machine learning is then applied to the fragments to determine, based on a predetermined threshold number of data points associated with the fragment, if the fragment corresponds to background noise
20 (e.g. 50% of data points relate to word or background noise), discard each fragment if the fragment is determined to correspond to background noise, perform recognition on each fragment that has not been discarded.

Recognizing new sentences upon segmentation approach

25 Taking the advantages of segmentation, the never-seen sentences New1-New3 are successfully recognized where these sentences are created by new-order word recombination. The process in Figure 7 illustrates recognition of a new sentence by segmentation, real-time sequential fragment identification, and sentence recognition. The CNN classifier is used, with 5 samples for each new
30 sentence engaged validating the CNN classifier performance in recognizing new sentences. As provided in Table 2, both classifiers render the number series prediction for the total 15 inputs of these 3 new sentences although the CNN model has not had the opportunity to be trained on the true label series. The hierarchy classifier shows a reliable performance for new sentence recognition,
35 indicated by fewer incorrect predictions – bold and underlined figures. Figure 8

shows the results after translation or inverse reconstruction of the new sentence, with Figure 8, image (a) being the accuracy for the single classifier – 60% - and Figure 8, image (b) being the accuracy for the hierarchical classifier – 86.67%.

Sentence	Sample	True label	Single classifier predicted label	Hierarchy classifier predicted label
New1	1	[0 0 19 19 19 7 7 19 19 19 5 5 5]	[00 00 19 19 08 07 07 19 19 19 05 05 05]	[00 00 19 19 19 07 07 19 19 19 19 05 05]
	2	[0 0 19 19 7 7 7 19 19 19 5 5 5]	[00 00 19 11 07 07 07 19 19 19 05 05 05]	[00 00 19 19 07 07 07 19 19 19 05 05 05]
	3	[0 0 19 19 19 7 7 19 19 19 19 5 5]	[00 00 19 19 06 07 07 19 19 19 19 05 05]	[00 00 19 19 19 07 07 19 19 19 19 05 05]
	4	[0 0 19 19 07 7 7 19 19 19 5 5 5]	[00 00 19 19 07 07 19 19 19 19 05 05 05]	[00 00 19 19 07 07 19 19 19 19 05 05 05]
	5	[0 0 19 19 19 7 7 19 19 19 19 5 5]	[00 00 19 19 19 07 07 07 19 19 19 05 05]	[00 00 19 19 19 07 07 19 19 19 19 05 05]
New2	1	[0 0 0 19 19 19 4 4 4 19 19 19]	[19 00 00 00 19 19 04 04 04 19 19 19]	[00 00 00 00 19 19 04 04 04 19 19 19]
	2	[0 0 0 19 19 19 19 4 4 19 19 19]	[0 0 00 00 19 19 19 04 04 04 19 19 19]	[00 00 00 19 19 19 04 04 04 19 19 19]
	3	[0 0 0 19 19 19 19 4 4 19 19 19]	[00 00 00 19 19 19 04 04 04 04 19 19]	[00 00 00 19 19 19 04 04 04 19 19]
	4	[19 0 0 0 19 19 19 4 4 4 19 19]	[19 00 00 18 19 19 19 04 04 04 19 19]	[01 00 00 00 19 19 19 04 04 04 19 19]
	5	[19 0 0 0 19 19 19 19 4 4 19]	[19 03 13 13 19 19 19 19 04 04 04 19]	[19 00 18 19 19 19 19 19 04 04 04 19]
New3	1	[17 17 19 19 19 1 1 1 19 19 2 2 2]	[17 17 17 19 19 01 01 00 19 19 02 02 02]	[17 17 19 19 19 01 01 01 19 19 02 02 02]
	2	[17 17 17 19 19 1 1 1 19 19 2 2 2]	[17 17 17 07 01 01 01 01 19 19 02 02 02]	[17 17 17 19 19 01 01 01 19 19 02 02 02]
	3	[17 17 17 19 1 1 1 19 2 2 2 19]	[17 19 17 19 01 01 01 19 02 02 02 19]	[17 17 17 19 01 01 01 19 02 02 02 19]
	4	[17 17 17 19 19 1 1 1 19 2 2 2 19]	[17 17 17 19 01 01 01 01 19 02 02 02 19]	[17 17 17 19 01 01 01 01 19 02 02 02]
	5	[17 17 17 19 19 1 1 1 19 2 2 2 19]	[17 19 17 07 19 01 01 01 19 02 02 02 19]	[17 19 19 19 01 01 01 01 19 02 02 19 19]

Table 2: Detailed prediction of three new sentences. The predicted and true labels of single classifier and hierarchy classifier for three new/never-seen sentence recognition with numbers in red standing for wrong prediction.

- Figure 9 illustrates the pros and cons of non-segmentation and segmentation methods. The non-segmentation classifier (i.e. non-segmentation approach) has better recognition accuracy but with higher latency and inability to recognise new sentences. Each word or sentence is recognised individually and treated as a separate class. The segmentation method (i.e. segmentation-assisted CNN) model can recognise new sentences and can perform real-time or near real-time recognition (i.e. translation) of sentences. The method may employ both techniques in sequence – e.g. the input signal may be first processed using the non-segmentation approach. If this results in a correlation between the input signal and known words/sentences below a predetermined threshold (e.g. 0.3), then the segmentation method is employed using a hierarchical classifier. New sentences identified by the segmentation classifier (i.e. segmentation method) can be incorporated into the training dataset.
- The sensory glove system and the methods described above can overcome the drawbacks of image recognition systems used in sign language recognition. On the one hand, the sensory glove system is not affected by varying luminance and can work well even under entirely dark conditions with higher environmental tolerance. On the other hand, sensory glove systems do not capture facial and other data that is otherwise inadvertently captured by image recognition systems.

VR communication interface for the signer and non-signer

- As mentioned with reference to Figure 1, the system 100 may implement a virtual reality conversation system. The signer and non-signer terminals in VR space allow display sign recognition results and the direct typing of non-signers. The VR interface facilitates two-way remote communication, linked with the AI front end for sign language recognition. The VR conversation system may implement a broader VR environment in which more than two parties may interact, regardless of the number of signers and non-signers among them. The deep learning integrated VR interface realizes bidirectional communication.

- To meet the requirement of high recognition accuracy in practical applications, the non-segmented deep learning model is used to achieve communication in

the virtual space. In some embodiments, however, it will be appreciated that the segmentation approach or sequence of non-segmentation and segmentation approaches could be used to improve flexibility of communication in the VR space. The system 100, used for recognition and communication, therefore
5 comprises five major blocks including triboelectric gloves (sensory glove system 102) for hand motion capture, the printed circuit board (PCB) for signal pre-processing (this may be in the processor 104 or 118), IoT module Arduino connected PC for data collection (this may be part of the signer terminal 104), deep-learning-based analytics for signal recognition (stored in central server
10 110, though may be elsewhere in some embodiments), and VR interface in Unity for interaction. Terminals 104 and 108 and nodes 128 and 130 may each be a VR interface, and may comprise VR headset, speaker for audio output of recognised input signals from terminals operated by signers and of text input from terminals operated by non-signers, and a display.

15

The sensory glove system (1000) and thus the method performed thereby (1002, though steps of the method are performed by other components of a system such as system 100), will output a signal corresponding to the recognised words or sentences. In the VR context, this recognised word or
20 sentence will be projected into cyberspace, in which AI will send corresponding commands based on recognition results and deal with inputs from the non-signer, to control the communication in VR interface based on Transmission Control Protocol/Internet Protocol (TCP/IP).

25 The deep learning block recognizes the non-mainstream sign language, translates it into the prevalent conversational medium such as text and audio, and transmits such information into the next virtual space. As shown in the sequence in Figure 10, image (b), sequence (i) to (v) the VR interface is designed for communication between the speech impaired and those without
30 speech impairment, with switched host-guest views in cyberspace. The client and server are built in the VR interface based on TCP/IP and are accessible for the signer and non-signer, respectively. Owing to the assistance of deep learning enabled sentence recognition and translation, the client of VR interface (akin to terminal 104) allows the speech impaired to use the sign language that
35 they are familiar with to engage in the communication. More precisely, the sign

language delivered by the speech impaired is recognized and translated into speech and text by deep learning. Then the speech and text are captured and sent to the non-signer-controlled server (akin to terminal 108, or server 110). Next, the healthy user employs the server to type directly responding to the
5 speech disordered user.

The sequence (i) to (v) in Figure 10, image (b) demonstrates, a greeting scenario is created to demonstrate the feasible communication between the speech/hearing impaired and healthy user under the identical local area network
10 (LAN), making interaction more realistic with AI integrated VR interface. Two virtual characters in the VR space, Lily and Mary, represent the created avatars for the signer and non-signer, with controllable and programmable multiple degrees-of-freedom motions. In the first step Figure 10, image (b), sequence number (i), the speech-impaired user Lily performs sign language 'How are you?' which is recognized and converted to text and audio by the deep learning
15 model. By means of TCP/IP, the client connected with deep learning component receives the text/speech message 'How are you?' and transmit it to the non-signer Mary controlled server. Projecting to the virtual space, the signer avatar Lily slightly lifts her hand to greet her friend (i.e., the non-signer Mary).
20 Correspondingly, the non-signer types 'Not good. I have a stomach ache.' to respond to the signer. The virtual girl Mary represents the non-signer shakes her head and covers her stomach with hands to show her illness, as shown in Figure 10, image (b), sequence number (ii). Then the speech-impaired user replies to the non-signer Mary with the sign language 'You need a doctor' Figure
25 10, image (b), sequence number (iii). Figure 10, image (b), sequence numbers (iv) and (v), similarly show exchange of greetings between the speech impaired (Lily) and the non-signer (Mary). The conversation of Figure 10, image (b), is summarized as shown in Figure 10, image (c). The VR communication interface linked with AI allows interaction of the speech/hearing-disordered and healthy
30 people closely and even remotely, providing a promising platform for the immense interactions between two population groups.

Demonstration of pure words and sentences recognition

While various systems can be employed for word and sentence recognition, Figure 10, image (a) shows the triboelectric sensory glove system 1000 with 15 sensor channels connected to an Arduino MEGA 2560 (1004). By serial communication with Python (1006 – implementing CNN) in PC, the raw data of different gestures is acquired and recognized by the trained CNN model with Tensorflow and Keras frames, then displaying recognition results on the Python interface.

VR communication interface building and demonstration

The words or sentences recognized by the Python script (1006) are sent into virtual Unity space via TCP/IP communication and displayed on the VR interface. The server (for non-signers) and client (for signers) terminal building in Unity also rely on TCP/IP. Within the identical LAN, the speech/hearing impaired and healthy people can remotely communicate via the VR interface where the speech/hearing disordered uses their familiar sign language, and the non-signer types directly.

Described herein are a sign language recognition and communication system comprising a smart triboelectric sensory glove system, AI block (implemented by a processor on one or more of the terminal and nodes shown in Figure 1), and the back-end VR interface (server 110 of Figure 1). The system enables the separate and independent recognition of words (i.e., single gestures) and sentences (i.e., continuous multiple gestures) with high accuracies of 91.3% and 95% within non-segmentation frame. Furthermore, to overcome the limitation of incapability of recognizing new sentences, the segmentation method is proposed. It divides all the sentence signals into word fragments while AI learns and memorizes all split elements. Then the deep learning architectures, single and hierarchy classifiers, inversely infer, reconstruct and recognize the whole sentence (recognition accuracy: 85.58%) benefitting from the established correlation of basic word units and sentences. Finally, the embedded VR interface can act as the bridge of user terminals, transmitting messages back and forth, configuring a closed-loop communication system with gloves and AI component. On the VR platform, the speech/hearing impaired can directly perform sign language to interact with the non-signer (i.e., human-to-

human interaction), while the non-signer directly involves in the communication process via direct typing.

It will be appreciated that many further modifications and permutations of various aspects of the described embodiments are possible. Accordingly, the described aspects are intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims or statements.

Throughout this specification and the claims or statements that follow, unless the context requires otherwise, the word "comprise", and variations such as "comprises" and "comprising", will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

15

The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

25

Claims

1. Sensory glove system for recognition of sign gestures, comprising:
a first glove for wearing on a first hand of a user, the first glove
5 comprising:
a first wrist motion sensor for wrist motion perception of a first
wrist;
a finger bending sensor for each finger of the first glove; and
a palm sensor; and
10 a second glove for wearing on a second hand of the user, the second
glove comprising:
a second wrist motion sensor for second motion perception of a
second wrist;
a finger bending sensor for each finger of the second glove; and
15 two fingertip sensors, one on each of a middle and index finger.
2. The sensory glove system claim 1, further comprising a processor, the
processor being configured to:
receive an input signal comprising an output from each wrist
20 motion sensor, bending sensor, palm sensor and fingertip sensor;
apply one or more machine learning models to the input signal to
recognise a sentence in the input signal.
3. The sensory glove system of claim 2, wherein at least one said machine
25 learning model is configured to:
recognise the sentence based on a similarity between the sentence
and one or more known sentences.
4. The sensory glove system of claim 2, wherein the processor is further
30 configured to:
divide the input signal into fragments and apply the machine
learning model to the input signal by applying at least one said machine
learning model to the fragments to recognise one or more words in the
fragments; and
35 inversely reconstruct the sentence from the one or more words.

5. The sensory glove system of claim 2, wherein the processor is further configured to:
- 5 calculate a similarity score between the sentence recognised from the input signal and one or more known sentences; and
- if the similarity score is at or above a predetermined threshold, output the sentence recognised from the input signal; or
- if the similarity score is below a predetermined threshold:
- 10 divide the input signal into fragments and apply at least one said machine learning model to the fragments to recognise one or more words in the fragments; and
- inversely reconstruct the sentence in the input signal from the one or more words.
- 15 6. The sensory glove system of any one of claims 1 to 5, wherein each sensor is a triboelectric sensor.
7. The sensory glove system of claim 6, wherein triboelectrification layers of the triboelectric sensors comprise an additional cured silicon layer.
- 20 8. The sensory glove system of claim 6 or 7, wherein triboelectrification layers of the triboelectric sensors comprise a wrinkled nitrile layer.
9. The sensory glove system of any one of claims 1 to 8, comprising 15 said sensors in total.
- 25 10. The sensory glove system of any one of claims 1 to 9, wherein the finger bending sensors are disposed on a front side of the respective finger.
- 30 11. The sensory glove system of any one of claims 1 to 10, wherein the first wrist motion sensor and second wrist motion sensor are position on a back side of the respective first wrist and second wrist.

12. The sensory glove system of claim 11, wherein the processor is configured to divide the input signal into fragments as the input signal is received.
- 5 13. A method for sign gesture sentence recognition from an input signal comprising a sentence, comprising:
dividing the input signal into fragments;
applying a machine learning model to the fragments to recognise individual said fragments; and
10 inversely reconstructing the sentence to recognise the sentence.
14. The method of claim 13, wherein the input signal is received from a sensory glove system according to any one of claims 1 to 12.
- 15 15. The method of claim 13 or 14, wherein the fragments include intact word signals, incomplete word signals and background signals, and wherein the applying the machine learning model to the fragments comprises:
determining, based on a predetermined threshold number of data points associated with the fragment, if the fragment corresponds to
20 background noise;
discarding the fragment if the fragment is determined to correspond to background noise; and
recognising each fragment that has not been discarded.
- 25 16. The method of claim 15, wherein, for each fragment comprising an intact word, the method further comprises labelling the fragment with a reference to the intact word.
- 30 17. The method of claim 15 or 16, wherein, for each fragment comprising background signals, the method further comprises recognising the fragment as comprising background signals on the basis of the fragment comprising fewer than 100 non-zero data points, and tagging the fragment as "empty".

18. The method of any one of claims 13 to 17, wherein dividing the input signal into fragments comprises receiving the input signal and dividing the input signal into fragments as the input signal is received.

- 5 19. A virtual reality conversation system comprising:
- a sensory glove system according to claim 1;
 - a signer terminal for use by a first user wearing the sensory glove system;
 - a non-signer terminal for use by a second user;
 - 10 at least one processor configured to:
 - receive an input signal comprising an output from each wrist motion sensor, bending sensor, palm sensor and fingertip sensor;
 - apply one or more machine learning models to the input signal to recognise a sentence in the input signal; and
 - 15 project the signal into a virtual reality environment, for display at the non-signer terminal.

20. The virtual reality conversation system of claim 19, wherein the processor is further configured to:

- 20
- receive an input signal from the non-signer terminal, the input signal comprising one or more words; and
 - project the input signal into the virtual reality environment for display at the signer terminal.

25