GLOVE-TALK: AN ADAPTIVE INTERFACE THAT USES NEURAL NETWORKS. *

S. Sidney Fels

Department of Computer Science, University of Toronto,
10 Kings College Road, Toronto, Canada, M5S 1A4

ABSTRACT

An interface which converts hand gestures to speech (Glove-Talk) was created to investigate the use of neural networks for adaptive interfaces. Glove-Talk is implemented with 5 neural networks and has a vocabulary of 203 words with variable speaking rate and stress. Glove-Talk is trained for a single user using automatic learning procedures It says the wrong word less than 1% of the time and says no words about 7% of the time. Current work is underway which looks at a finer grained mapping of hand gestures to speech and thus requires a more complex interface.

INTRODUCTION

Presently, many complex devices exist which require from the user large amount of information in real-time. Examples include a data-glove, a car and a remote control robot arm. To allow the user to provide the necessary high bandwidth control, a compatible mapping between the person's movements and the device is necessary. One solution to ensuring the mapping is good is to allow the user to specify the mapping they want and then automatically adapt the mapping to their specification. Using this extensional programming approach, the user supplies training data to teach the adaptive interface their current model of the device. This data can be used to train the forward system to be the inverse of the inverse model since this is usually a many-to-one mapping which is easy to learn automatically.

Given training examples for a many-to-one mapping, neural networks are an effective technique for automatically approximating the mapping function. Neural networks are composed of simple, neuron-like processing elements which are richly interconnected. A common type of network contains some input units, one or more intermediate layers of "hidden units", and a layer of output units. The number of layers, the connections between layers is typically chosen by hand, but the way in which the network maps vectors of activity over its input units into vectors of activity over its output units is determined by the weights on the connections. These weights are typically learned from a set of training examples that specify the desired output vector for each input vector. Once the network has learned, it can generalize to new cases. Neural networks have been successfully applied to important real-world problems such as speech recognition [1], hand drawn character recognition [2] and adaptive control systems [3]. Several important features of neural networks in the context of adaptive interfaces are:

- Neural networks can learn to approximate any input/output function.
- Neural networks learn input/output functions from examples provided by the user who demonstrates the input that should lead to a specified output. This "extensional" programming requires no computer expertise.
- Adapting the interface to the peculiarities of a new user is simple and quick. The new user has only to create example data to retrain the network.
- Once trained, the networks run very quickly, even on a serial machine. Also, neural networks are inherently suitable for parallel computation.

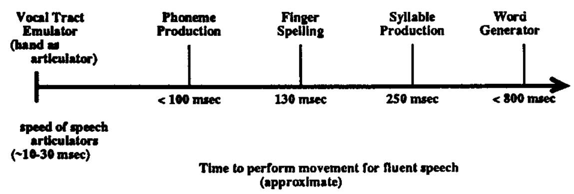

Figure 1: Mapping Spectrum based on Division of Speech

SPECTRUM OF HAND TO SPEECH MAPPINGS

There are many different mapping schemes for converting hand gestures to speech. The choice of scheme depends on the granularity of the speech that you want to produce. Figure 1 identifies a spectrum defined by possible divisions of speech based on the duration of the sound for each division. What is interesting is that in general, the coarser the division of speech, the smaller the bandwidth necessary for the user. In contrast, where the granularity of speech is on the order of articulatory muscle movements high bandwidth control is necessary for good speech. For the Glove-Talk project we chose to implement a system that maps hand gestures to words. Current research efforts are focussed at the other end of the spectrum where we are implementing an artificial vocal tract. For good quality speech with this model it is critical that the mapping is well adapted to the user so that they can produce the signals at a fast enough pace. In addition, the mapping should map all the possible sounds of a vocal tract in some natural way. The mapping must also be fast to ensure little delay between the hand gesture and the sounds produced. Other research with hand-to-speech devices includes a system which is designed for finger spelling [4].

GLOVE-TALK

Glove-Talk was created to demonstrate the usefulness of neural networks for adaptive interfaces. A complete description of it can be found in [5]. Figure 2 shows a block diagram of the Glove-Talk system. The current Glove-Talk vocabulary consists of 66 root words, each with up to six different endings. The total size of the vocabulary is 203 words. Three examples of the initial mapping of hand shapes (and orientations) to words are shown in table 1. The hand shapes are specified by the user during training as are the gestures controlling word endings, rate and stress. Orientation differences in the hand shapes are usually used for semantically opposite words; for example, the hand shapes for "come" and "go" have the same finger angles but are 180° of roll apart (i.e. "come" is made with the palm up and "go" is made with the palm down). The various endings of the words are formed by different directions of the hand movement. The 66 root words and six endings were extracted from the 850 word vocabulary of Basic English [6]. The hand gesture data is sensed by a VPL DataGlove [7] that has two sensors for each finger. The sensors are fibre optic transducers which measure

*This research was supported by the Canadian Institute for Robotics and Intelligent Systems and the Canadian National Science and Engineering Research Council. The DECtalk speech synthesizer was provided by Terry Sejnowski and the Xerion neural network simulator we used was written by Tony Plate.

CH3068-4/91/0000-1264 $01.00 © 1991 IEEE

root word	hand shape
come	
go	
I	

Table 1: Examples of Glove-Talk Language

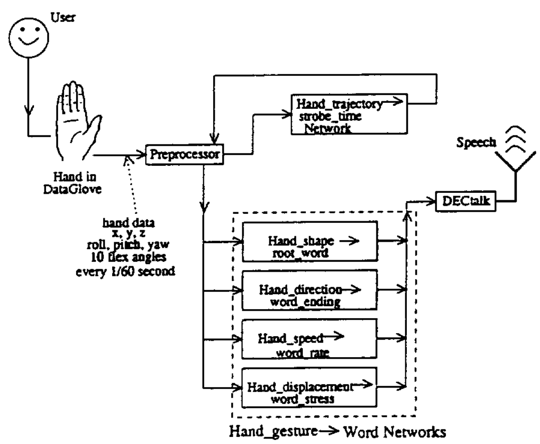

Figure 2: Glove-Talk System

the finger flex angles. There is also a "polhemus" sensor attached to the back of the glove which measures the x, y, z, roll, pitch, and yaw position of the hand relative to a fixed source. All 16 parameters are measured every $1/60^{th}$ second. The speech synthesiser is a DECtalk model DTC01 from Digital Equipment Corporation. This synthesizer can perform text-to-speech synthesis and allows user control of speaking rate and word stress.

Once trained, the Glove-Talk system works as follows: the user forms a hand-shape which represents a root word. Then s/he makes a movement forward and back in one of six directions. The direction of the movement determines the word ending, and the duration and magnitude of the gesture determine the speech rate and stress. The precise time at which the word is spoken is more complex. The user imagines that the end of the forward movement is like a button press that causes the word to be produced immediately, much like a conductor producing music. To avoid delays in producing the word, the shape of the hand is actually "read" near the beginning of the deceleration phase of the forward movement. This moment is called the "strobe-time" and is detected by a hand_trajectory→strobe_time network ("strobe network") that continually monitors the directionless speed, x, y and z component speeds and acceleration of the whole hand using preprocessed information from the polhemus device. The strobe network has 50 input units, 10 hidden units and 1 output unit (50-10-1 network). It was very difficult to train. The network requires the user to generate many instances of strobe-times which are then hand-labeled as example data for training the strobe network. For Glove-Talk, 638 strobe-times were hand labeled out of 30,356 time steps of data. Once, trained the network had a 5% error rate. When the strobe network detects a strobe-time, it sends a signal back to the preprocessor. The preprocessor then sends the appropriate buffered data to each of four neural networks to do the hand gesture to word mapping. Once the strobe network is working it is easy to create the training data for the other four networks since the target output for the networks can be displayed for the user to generate. When the strobe network detects a strobe-time the buffered data is stored and automatically labeled for training data.

The hand_shape→root_word network (16-80-66) determines the correct root word based on the static hand shape and orientation at the strobe-time. It was trained on 8912 examples, giving a false alarm rate less than 1% and a miss rate about 2%. From the direction of hand movement the hand_direction→word_ending network (30-10-6) determines which of the six possible endings (-s, -ed, -ly, -er, -ing, and normal) the user intended. The hand_speed→word_rate (40-15-8) and hand_displacement→word_stress (40-15-1) networks determine the speaking rate and word stress, based on the speed and magnitude of the hand movement respectively. Once the mapping is completed, the appropriate commands are sent to the speech synthesizer which then speaks the word with the specified rate and stress.

CONCLUDING REMARKS

Fairly rapid, intelligible speech is possible with the current Glove-Talk system. About 1% of the words spoken are incorrect, and about 7% of attempts result in no word being spoken due to failure to detect the gesture or failure to confidently identify the root word.

Obvious improvements include user control of pitch and loudness, continued "online" training while the system is in use, and a method for explicitly spelling out words that are not in the fixed vocabulary. One major improvement would be to increase the number of root words to the entire 850 words of Basic English. This may require substantial restructuring of the system, since static hand-shapes and orientations may not be sufficiently reproducible to allow 850 discriminable alternatives. It may therefore be necessary to encode root words by time-varying hand shapes or to use hand position or its temporal derivatives to distinguish between root words. A sequence of two static hand-shapes would only require 30 discriminable alternatives to specify 900 root words, and a neural network should be good at adapting to co-articulation effects between the two consecutive hand-shapes.

We are currently investigating a finer-grained mapping in which finger movements play the role of the normal speech articulators. Such a mapping would be difficult to learn, but would allow an unlimited vocabulary and much greater control of the speech. Instead of initially allowing the user to specify their inverse model as training data, it may be preferable to use a neural network to learn both an inverse mapping and a forward mapping with constraints on the interface between the two. This would "boot-strap" the process of determining an initial mapping.

Finally, it would be illuminating to investigate the performance that can be achieved at this kind of real-time task using other approaches such as standard pattern recognition techniques or expert systems.

REFERENCES

[1] Lippmann, R.P., "Review of Neural Networks for Speech Recognition", *Neural Computation*, 1:1-38, 1989.

[2] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R., Hubbard, W., and Jackel, L.D., "Backpropagation applied to handwritten zip code recognition", *Neural Computation*, vol 1:4, 1990.

[3] Nguyen, D., and Widrow, B., "The Truck Backer-Upper: An example of Self Learning in Neural Networks", *Proceedings of International Joint Conference on Neural Networks*, 2:357–363, Washington D.C., USA, 1989.

[4] Kramer, J., and Leifer, L., "The 'Talking Glove': A Speaking Aid For Nonvocal Deaf and Deaf-Blind Individuals", in *Proceedings of RESNA 12th Annual Conference*, 471–472, Louisiana, USA, 1989.

[5] Fels, S.S., *Building Adaptive Interfaces Using Neural Networks: The Glove-Talk Pilot Study*, Technical Report CRG-TR-90-1, University of Toronto, Toronto, Canada, 1990.

[6] Ogden, C.K., *Basic English: International Second Language*, prepared by Graham, E.C., Harcourt, Brace & World Inc., NY, USA, 1968.

[7] VPL Research Inc, *DataGlove Model 2 Operating Manual*, CA, USA, 1989.

Author's address: Department of Computer Science, University of Toronto, Toronto, ON, CANADA, M6C 2Z6, phone: (416) 978-7453, email: ssfels@ai.toronto.edu

CH3068-4/91/0000-1265 $01.00