

Using Normalized RBF Networks to Map Hand Gestures to Speech

S. Sidney Fels

Department of Electrical and Computer Engineering
The University of British Columbia, Vancouver
Canada

Glove-TalkII is a system that translates hand gestures to speech through an adaptive interface. Hand gestures are mapped continuously to 10 control parameters of a parallel formant speech synthesizer. The mapping allows the hand to act as an artificial vocal tract that produces speech in real time. This gives an unlimited vocabulary in addition to direct control of fundamental frequency and volume. Currently, the best version of Glove-TalkII uses several input devices (including a Cyberglove, a 3-space tracker, a keyboard and a foot-pedal), a parallel formant speech synthesizer and 3 neural networks. The gesture-to-speech task is divided into vowel and consonant production by using a mixture of experts architecture where the gating network weights the outputs of a vowel and a consonant neural network. The gating network and the consonant network are trained with examples from the user. The vowel network implements a fixed, user-defined relationship between hand-position and vowel sound and does not require any training examples from the user. Volume, fundamental frequency and stop consonants are produced with a fixed mapping from the input devices. One subject has trained to speak intelligibly with Glove-TalkII. He speaks slowly with speech quality similar to a text-to-speech synthesizer but with far more natural sounding pitch variations.

In the final Glove-TalkII system [8], one of the main networks is the consonant network comprised of an input layer of 12 units corresponding to twelve hand sensors, a hidden layer of fifteen *normalized* RBF units connected to nine sigmoid output units. Normalized RBF units provide a much better topology for mapping hand gestures to consonant sounds than either unnormalized RBF units or sigmoid units. The RBFs were trained in two passes. First, an approximation of the RBF centers was obtained in a single quick pass through the training data. The

centers were clamped while the output layer was trained iteratively using conjugate gradient descent on the training data. After this first stage of training was complete, the RBF centers were unclamped and all the network parameters were optimized. The vowel network also used a hidden layer with normalized RBF activation functions which proved to be critical in the performance of the final Glove-TalkII system. The V/C gating network used sigmoid activation units. In this discussion of Glove-TalkII, the implementation, training, implications, and interpretation of normalized RBF units are covered. The normalized RBF architecture is compared to architectures having linear, sigmoid, sigmoid plus softmax, and unnormalized RBF activation functions in both gesture-to-vowel and gesture-to-consonant mapping. In addition, using a normalized RBF network in the context of a mixture-of-experts framework in a soft real-time environment is also discussed.

1 Introduction

Adaptive interfaces are a natural and important class of applications for neural networks. When a person must provide high bandwidth control of a complex physical device, a compatible mapping between the person's movements and the behavior of the device becomes crucial. With many devices the mapping is fixed and if a poor mapping is used, the device is difficult to control. Using adaptive neural networks, it is possible to build device interfaces where the mapping adapts automatically during a training phase. Such adaptive interfaces would simplify the process of designing a compatible mapping and would also allow the mapping to be tailored to each individual user. The key features of neural networks in the context of adaptive interfaces are the following:

- 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. 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.

In this chapter, neural networks are used to implement an adaptive interface, called Glove-TalkII, which maps hand gestures to control parameters of a parallel formant speech synthesizer to allow a user to speak.

There are many different possible 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 granularity. 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 (i.e. the artificial vocal tract [AVT]) high bandwidth control is necessary for good speech. Devices which implement this model of speech production are like musical instruments which produce speech sounds. The user must control the timing of sounds to produce speech much as a musician plays notes to produce music. The AVT allows unlimited vocabulary, control of pitch and non-verbal sounds. Glove-TalkII is an adaptive interface that implements an AVT.

Translating gestures to speech using an AVT model has a long history beginning in the late 1700's. Systems developed include a bellows-driven hand-varied resonator tube with auxiliary controls (1790's [16]), a rubber-moulded skull with actuators for manipulating tongue and jaw position (1880's [1]) and a keyboard-footpedal interface controlling a set of linearly spaced bandpass frequency generators called the Voder (1940 [5]). The Voder was demonstrated at the World's Fair in 1939 by operators who had trained continuously for one year to learn to speak with the system. This suggests that the task of speaking with a gestural interface is very difficult and the training times could be significantly decreased with a better interface. Glove-TalkII is implemented with neural networks which allows the system to learn the user's interpretation of an articulatory model of speaking.

This chapter begins with an overview of the whole Glove-TalkII system. Then, each neural network is described along with its training and test