EECS225d Audio Signal Processing in Humans and Machines	**Spring 1995**

Lecture 28: March 24, 1995

Lecturer: Sidney Fells (University of Toronto) *Scribe: Michael Shire*

28.1 Introduction

Sid Fells described the operation and utility of the Glove-TalkII system. Glove-TalkII is a gesture-to-speech device. An operator would generate input to a speech synthesizer by motions of his hands. Adaptive neural networks learn the gestures of an operator and convert the operator's hand motions to speech parameters. A formant-based speech synthesizer takes the parameters and generates the speech in real-time.

Neural networks are used because they are adept at finding mappings between inputs and outputs. Since there would be a great deal of variation between the gestures of different operators and even the same operator, the neural network would need to adapt. An adaptive interface is useful for high-bandwidth devices. The idea behind the gesture-to-speech device is to model an artificial vocal tract with a different part of the body. This would be useful for someone who did not have use of his/her vocal tract.

28.2 Spectrum of Gesture-To-Speech Mapping

The movement times of the hands must be taken into account when mapping gestures to speech. Gestures must be quick enough to keep the speech output real-time and natural. An approximate granularity for speech is:

- 10-30 msec: Artificial vocal tract
- 100 msec: Phoneme generator
- 130 msec: Finger spelling
- 200 msec: Syllable generator
- 600-up msec: Word generator

For an artificial vocal tract, gesture would model the vocal tract dynamics. The times above are derived from formant rates and previous work by Von Kempelen, Bell, and Dudley.

28.3 Vocabulary

The types of gestures were intended to model an articulatory representation of speech. The mappings attempted to make the positions and states of the hands and fingers correspond roughly and intuitively to what happens in the vocal tract.

To create vowels, the hand was kept in on open configuration, corresponding to an open vocal tract. The XY position of the hand was loosely mapped to the tongue position and height. This determined the vowel sound. See figure 3 for XY position map for vowels.

For consonants, fingers were bent toward the thumb; a constricted hand corresponding to constrictions in the vocal tract. The selection of the finger determined which consonant. See table 2 for an example of static non-stop consonants.

Stop consonants were handled by buttons on the fingers. The neural networks were initialized with an initial articulatory model.

28.4 System Overview

A Cyberglove (right hand) and polhemus combination measured the spatial position of the hand in the XYZ plane. The polhemus measured all six degrees of freedom but roll, pitch, and yaw were not used. Contact switches for stop consonants were placed on the left hand. A foot pedal provided amplitude gain. The pitch of the voicing was mapped directly to the Z (up/down) position of the hand. In all there were 16 speech parameters at a rate of 100Hz fed into a parallel formant speech synthesizer. The parameters were quantized to six bits. There were 10 main parameters:

- ALF low frequency amplitude
- F1 first formant frequency
- A1 first formant amplitude
- F2 second formant frequency
- A2 second formant amplitude
- F3 third formant frequency
- A3 third formant amplitude
- AHF high frequency amplitude
- V degree of voicing
- F0 fundamental frequency

Three expert neural networks were used to determine what the operator was intending to input. One network dealt with vowel consonant decisions. A second dealt with vowels. The last dealt with consonants. A separate function weighted and combined the decisions and parameters of the three networks. See figure 2 for a system diagram.

• Voice/Consonant Decision Network This network produced the probability that the user intends to say a vowel. There were 12 inputs corresponding to the shape of the hand: the 10 flex angles, the thumb abduction, and the thumb rotation. There were 5 hidden units. This net was trained on 2600 consonants and 700 vowels.

• Consonant Network This net had the same 12 inputs as above. 15 hidden units were normalized to RBF (Gaussian) units. The outputs of this net were the 10 main parameters mentioned previously. Training included 2600 patterns from the user on 15 different root sounds (phonemes). There was high formant error but the output speech still sounded good.

• Vowel Network This network produced 8 formant parameters. The XY hand position served as input. There were 11 hidden units normalized to RBF units. Training was done on 330 vowels. One potential hazard was the absence of an absolute XY axis. The centers for each vowel were fixed. Analysis of the network after training showed that the vowel space was discretized by the net so that different sections of space mapped to different vowels with discrete boundaries in between.

28.5 Learning to Speak

A rehearsed video presentation was presented to the class. A pianist trained for 100 hours. Among the demonstrations were the individual phonemes, the "ABC" song, numbers, and a recitation of "Sam I Am" by Dr. Seuss. The steps for learning the system were as follows:

- Initial setup (wearing equipment)
- Initial network training
- Simple words and phoneme formation
- Words formed with pitch control
- Short segments of speech ("abc", "how are you?")
- Reading ("Sam I Am")
- Fine tuning, movement control, phrasing
- Conversation

28.6 Conclusions

The Glove-TalkII demonstrated that intelligible speech was achievable through gesture-to-speech processing. It was observed during experimentation that pitch control greatly improved intelligibility. Some problems encountered included achieving a high speaking rate, producing 'R's, gesturing polysyllabic words, and producing reliable and accurate vowels (absence of absolute XY axis). It was speculated that an inverse mapping would be of great use in training an operator. Also, no part of the system necessitates use of neural network and so other techniques may be explored also. Neural networks do seem to be most appropriate for learning complex mappings.

Some open questions deal with how children and non-native speakers would learn and if a more structured leaning process than the one used would help. It was observed that after 100 hours, the operator's skill had not reached any indication of a plateau. Interestingly, many tongue-twisters were in no way difficult to produce but new difficult to pronounce "hand-twisters" were introduced.

References

[FH95] Sidney Fels and Geoffrey Hinton, "Glove-TalkII: An Adaptive Gesture-to-Formant Interface" 1995.

[FH94] Sidney Fels and Geoffrey Hinton, "Glove-TalkII: Mapping Hand Gestures to Speech using Neural Networks" Dissertation, University of Toronto, 1994.

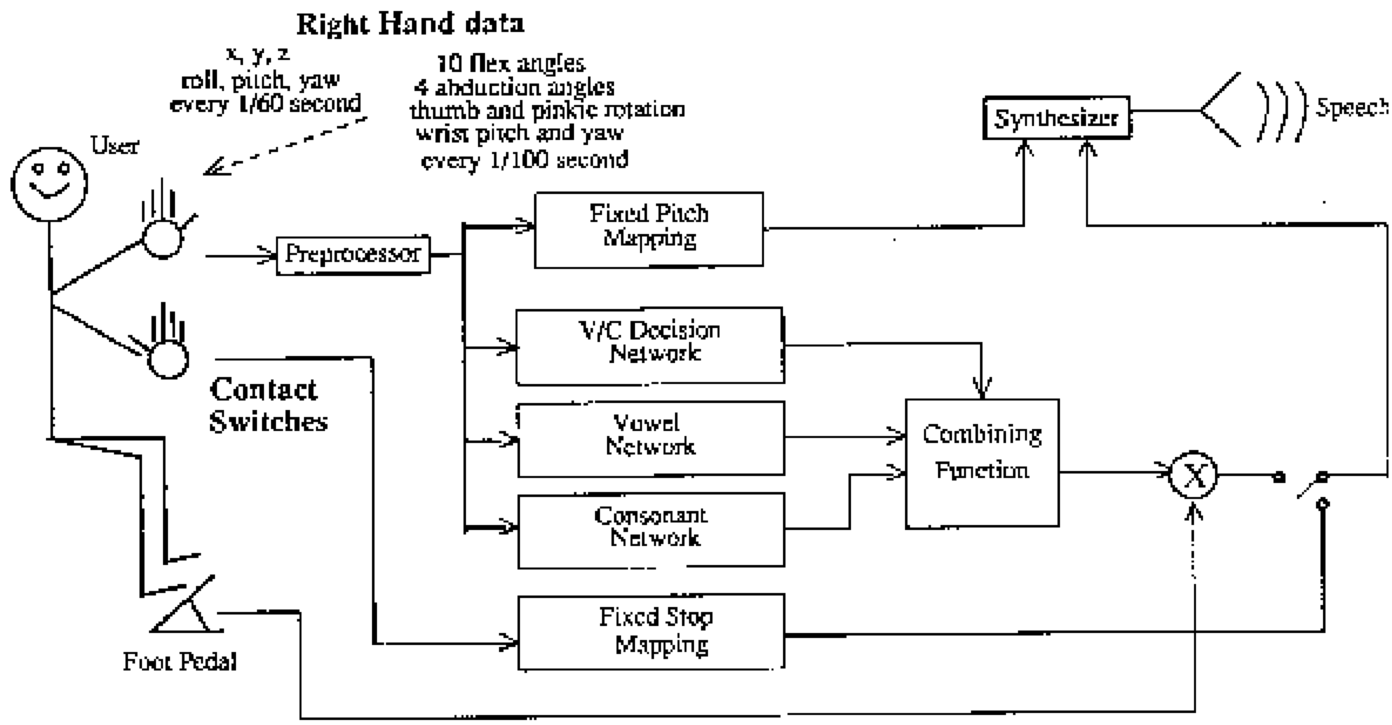

Figure 2: Block diagram of Glove-TalkII: input from the user is measured by the Cyberglove, polhemus, keyboard and foot pedal, then mapped using neural networks and fixed functions to formant parameters which drive the parallel formant synthesizer [14].

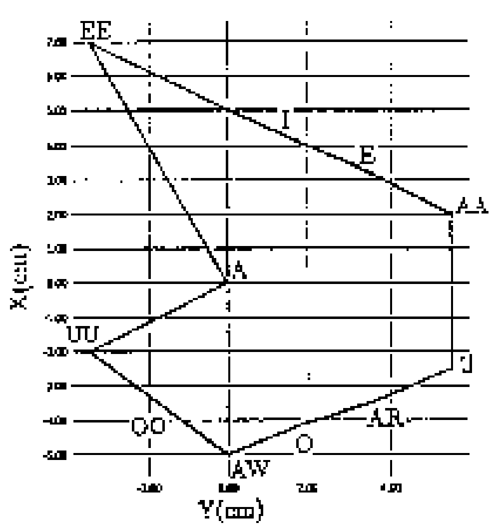

Figure 3: Hand-position to Vowel Sound Mapping. The coordinates are specified relative to the origin at the sound A. The X and Y coordinates form a horizontal plane parallel to the floor when the user is sitting. The 11 cardinal phoneme targets are determined with the text-to-speech synthesizer.

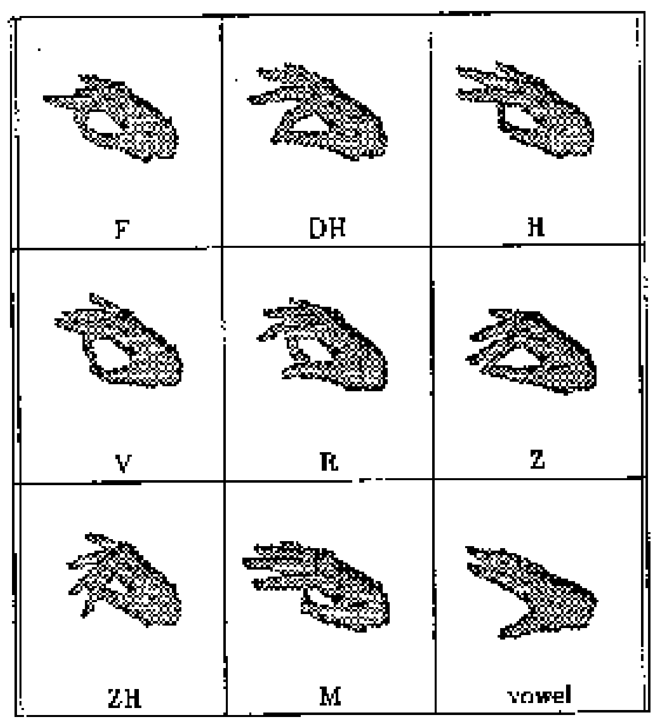

Table 1: Examples of static gesture-to-consonant mapping. Note, each gesture corresponds to a static non-*stop* consonant phoneme generated by the text-to-speech synthesizer and the neural networks provide the continuous interpolation.

Figure 28.1: Figure 3: Block diagram of Glove-TalkII