ForTouch: A Wearable Digital Ventriloquized Actor

Sidney Fels
Electrical & Computer Engineering
University of British Columbia
Vancouver, BC, Canada
ssfels@ece.ubc.ca

Robert Pritchard
School of Music
University of British Columbia
Vancouver, BC, Canada
bob@interchange.ubc.ca

Allison Lenters
Electrical & Computer Engineering
University of British Columbia
Vancouver, BC, Canada
alenters@interchange.ubc.ca

We have constructed an easy-to-use portable, wearable gesture-to-speech system based on the Glove-TalkII [1] and GRASSP [2] Digital Ventriloquized Actors (DIVAs). Our new portable system, called a ForTouch, is a specific model of a DIVA and refines the use of a formant speech synthesizer. Using ForTouch, a user can speak using hand gestures mapped to synthetic sound using a mapping function that preserves gesture trajectories. By making ForTouch portable and self-contained, speakers can communicate with others in the community and perform in new music/theatre stage productions. Figure 1 shows one performer using the ForTouch. ForTouch performers also allow us to study the relation between gestures and speech/song production.

1. Hardware Design

The wearable aspects of the design of the ForTouch have been described in [3]. Here we outline the main technical components of ForTouch and focus on the primary elements added for the ForTouch's use as a new interface for musical expression (NIME). In particular, for use as a NIME, the needs and constraints of the performer were incorporated into the overall design of the system. As shown in Figure 2, the following hardware components are tailored to be worn by the performer: Cyberglove™, Polhemus Patriot tracker™, a TouchGlove, in-sole wireless footpedal, battery powered speaker, and a laptop computer.

The Cyberglove™ and Patriot™ tracker are worn on the right hand to track hand movements and gestures that are mapped to vowel and consonant sounds. The Bluetooth based TouchGlove is worn on the left hand and has eight contact sensors that are activated by the performer pinching the appropriate pad with her left thumb. These are mapped to plosives: 'B', 'D', 'G', 'J', 'P', 'T', 'K', and 'CH'. A Bluetooth-connected small momentary switch is placed within an insole that the performer wears in her shoe to turn on and off sound. The speaker is worn over the left side of the performer's chest so that she has a voice associated with her body. Currently we use an Apple™ MacBook 2.4GHz Intel Core 2 Duo running OS X 10.5.4 for the data processing and sound generation. The performer wears a harness that contains a backpack component that houses the laptop. The harness also supports the Patriot's transmitter station. All the components connect to the laptop in the harness either wirelessly or though cable tunnels in the vest the performer wears on stage as shown in figure 1.

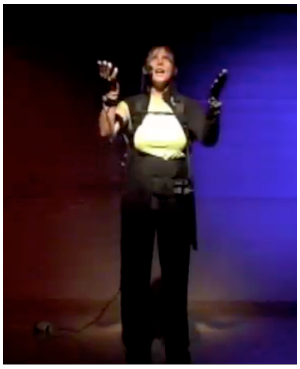

Figure 1: Photo from ForTouch performance, "What does a body know? in Japan, Jan 24, 2009; performer Marguerite Witvoet.

2. Software Design

The ForTouch's software components are based on the Glove-TalkII and GRASSP design but adjusted to be performer-friendly facilitating training by the performers.

2.1.1 Sound Mapping

As in Glove-TalkII, ForTouch uses three neural networks trained on samples provided by the user to map the right-hand gestures to vowel and consonant sounds. One normalized Radial Basis Function (RBF) network is specialized to map the X (front-back) and Y (left-right) coordinates of the right hand to vowel formants while another maps right hand finger movements to consonant

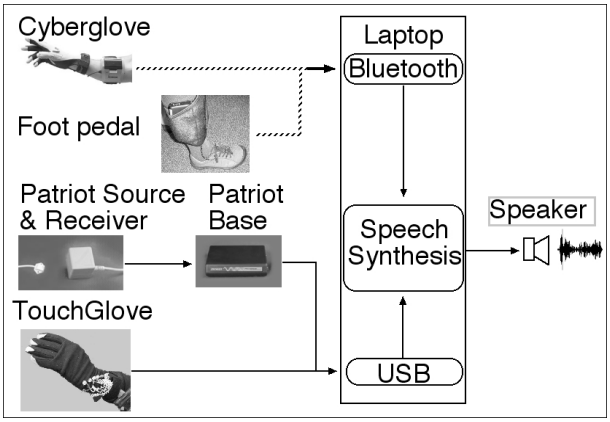

Figure 2: Block diagram of the ForTouch

formants. The third network blends these two formant outputs together based on how much of a vowel or consonant shape the performer's hand is in. The centres of each RBF in each network are set to the respond to a hand posture associated with a cardinal sound such as an 'EE', 'L', or 'SH', as defined by equation (1). There are 11 cardinal vowel sounds and 15 cardinal consonants mapped this way. The left TouchGlove behaves like eight buttons to trigger preset stop consonant formant trajectories to be delivered to the formant synthesizer. Right hand movement on the Z axis (up-down) controls pitch through a fixed mapping that has customizable frequency range, quantization and physical ranges. The foot pedal controls the master volume. Collectively, all the formant outputs drive a software version of a parallel formant speech synthesizer [4].

The normalized response NR of each RBF is given by the following equations, where M and σ are the mean and standard deviations over all training samples, for each sensor parameter i; and I is the current real-time value for that same sensor.

$$R_i = \frac{\Sigma (M_i - I_i)^2}{\sigma_i^2} \quad ; \quad NR_j = \frac{R_j}{\Sigma R_j} \qquad (1)$$

We simplified our approach for obtaining the means and standard deviations for the RBFs as discussed next.

2.1.2 Simplifying the User Interface

Two main user interface limitations in the Glove-TalkII/GRASSP systems require attention to make the ForTouch more effective for performance. First, setting up the system to speak/sing, tuning the many parameters to personalize the sound and entering training data was very complicated and required considerable technical expertise. We built a new user interface to simplify all of these tasks. Key to the simplification is the creation of *performer profiles* that keep track of all the training sessions and configuration parameters for each user including presets for performances. This approach provides a coherent mental model for performers to know which parameters do what, what each mode means and how to easily organize their training data and save/retrieve configurations for use in performances so they can concentrate on performance issues rather than software issues.

For our second simplification we reduced the amount of training data required. Originally, hundreds of samples for each cardinal sound were required from the user to train Glove-TalkII and it also required many hours of gradient descent optimization to train the weights of the networks. In GRASSP, we reduced training to a single example per sound, but this proved to be ineffective as the mapping was too sensitive to specific hand postures to make it easy to control. In ForTouch, we allow users to supply as many examples as they want and easily add new examples to fine-tune the sounds to take into account variations of hand posture for cardinal sounds. Further, we have reduced the network training time to be essentially instantaneous by calculating the RBF centres based on the means (and standard deviations) of the hand postures for each of the cardinal sounds and fix the linear mapping between the normalized radial-basis function outputs to the formants for these sounds

3. ForTouch Aesthetics and Performance

Currently we are creating a stage work from Cadell's libretto *What does a body know?* and have had one public performance of the first movement. We have also created videos of English and Japanese speech. Our libretto uses the current vocabulary of our single performer and the subtext of the piece is the singer's discovery and teaching of the system. During the performance the singer uses the ForTouch and her own voice and often the same words occur sequentially or simultaneously. This primes the ear for comprehension and supports the connection between the system and the body. In a similar manner in the videos we provide sub-titles to prime the listener's ear.

4. Conclusions and Future Work

Currently the DIVA system is being tested and modified. Design and performance issues are still being identified as well as sound and face synthesizer changes. More work needs to be done on the design, integration and robustness of the touch pads and their wiring. We continue to refine the harness in order to achieve the best weight distribution and comfort for the performer during normal use. As well, we continue to work with easily adjustable speaker and video display configurations that will be used in different circumstances (performance monitoring, mobile speech, mic-ed public speaking). The resolution of these issues combined with testing of the training paradigm will allow us to move forward with the next stage of development, performance and experiments for understanding the role gesture plays with the intelligibility and expressiveness of a DIVA.

5. Acknowledgments

We are grateful to SSHRC, The Canada Council for the Arts, NSERC, MAGIC and ICICS, and the DIVA team (www.magic.ubc.ca/VisualVoice.htm).

References

[1] Fels, S; Hinton, G.; Glove-TalkII: A neural network interface which maps gestures to parallel formant speech synthesizer controls. IEEE TNN. Vol 9(1):205-212. 1998.

[2] Pritchard, B.; Fels, S.; GRASSP: gesturally-realized audio, speech and song performance, NIME, 272-276, 2006.

[3] Day-Fraser, Helene; Fels, S.; Pritchard, R.; Walk the Walk, Talk the Talk, International Symposium on Wearable Computing (ISWC), 117-118, 2008.

[4] Rye, J.M.; J.N. Holmes; A versatile software parallel-formant speech synthesizer, JSRU Research Report No. 1016, Nov 1982.