

# Building a Portable Gesture-to-Audio/Visual Speech System

Sidney Fels<sup>1</sup>, Robert Pritchard<sup>2</sup>, Eric Vatikiotis-Bateson<sup>3</sup>

<sup>1</sup>Electrical and Computer Engineering, <sup>2</sup>School of Music, <sup>3</sup>Department of Linguistics  
University of British Columbia, Vancouver, BC, Canada

ssfels@ece.ubc.ca, bob@interchange.ubc.ca, evb@interchange.ubc.ca

## Abstract

We have constructed an easy-to-use portable, wearable gesture-to-speech system based on the Glove-TalkII and GRASSP gesture-controlled speech systems and a vizeme based face-synthesizer. Our new portable system is called a Digital Ventriloquized Actor (DIVA) and refines the use of the formant speech synthesizer. Using a DIVA, a user can speak using hand gestures mapped to both synthetic sound and face using a mapping function that preserves gesture trajectories. By making DIVAs portable and self-contained, speakers can communicate with others in the community and perform in new music/theatre stage productions. DIVA performers also allow us to study the relationship between visible gestures and speech/song production.

**Index Terms:** speech synthesis, gesture, face synthesis, wearable computing

## 1. Introduction

Technologies such as GloveTalkII [1] and GRASSP [2] are sensor-based systems that allow people to speak using their hands to control a speech synthesizer. However, our personal voice includes gestural, aural, visual, and postural elements, most of which are absent in the GloveTalkII/GRASSP systems due to their technological constraints. Also absent is the ability to be mobile while communicating making it impossible for people to walk around and talk with those systems. To overcome these limitations, we have developed a wearable system called Digital Ventriloquized Actors, or DIVA. Once mobile, a DIVA enables a person to sing and speak wherever they want. This system will be used in artistic productions, linguistic studies, and the daily communicative needs of each individual.

GloveTalkII was designed solely to allow a user to generate simple understandable speech, while GRASSP developments allowed greater expressiveness through speech and sound synthesis. The DIVAs are the culmination of a progression that allows individuals to embody technology for sophisticated vocal expression [3].

GloveTalkII is a gesture-controlled real-time speech synthesis system that used two glove-based inputs devices (a Cyberglove and a custom TouchGlove), a six degree of freedom (dof) magnetic tracker (Polhemus Fastrak) and a foot pedal, all connected to a parallel formant speech synthesizer [4]. GRASSP is a refactoring of the GloveTalkII code into Max/MSP with expanded capabilities for sound control and synthesis. In both of these systems the sensing technologies were connected through wires and stationary magnetic fields preventing speakers from being able to move and gesture freely. We expand on our description of our reconstruction [5] to fit within a mobile, wearable version suitable for performances and daily communication including animated face control. The wearable requirements are constrained by

the need to create a comfortable, aesthetically designed suit that will house all the technology. The development of a vocal, wearable computer suitable for stage performance, linguistic studies and everyday speaking provides insights into how wearable computing can be used for expressive content.

## 2. Related Work

Since the 1950's almost all speech synthesis systems have been text based and non-mobile from a performance point of view. While some systems have been made to mount on wheelchairs, such as the blink-switch operated speech synthesizer used by the well-known cosmologist, Steven Hawking, none to date have been incorporated for performance that are made to be worn. Further, systems such as [6][7][8] provide voice synthesis methods for performances, but are intended to be played as a stage mounted instrument which does not allow a performer to move around as they perform.

DIVA's design encompasses the notion that for a person's voice to be expressive they will need to embody the device [9]. Hence, we require specific aesthetic elements that still allow us to be able to contain the technology. From the textile perspective, creating a durable suit with the appropriate aesthetics is a challenging problem for wearables [10][11]. We have the added issue that the wearable device forms the individual voice and will be used on stage. Our approach pays attention to the size, weight, distribution, placement and fasteners to support comfort and durability, even during long usage. Our aesthetic needs include demands from the speaker as well as from the artistic director, composer, and librettist requiring a complex design process used by the clothing/product designer that follows the critical design path up to distribution stages since we plan on making only three DIVAs.

## 3. Design Approach

One of the primary functions of our design process was to assess how the object potentially shapes the relationships between individuals that come into contact with it [3]. The intent was to design a new mobile system that not only supported communication between individuals in public and private settings, but also was a performance mechanism in which the voice and the story that the individual projects is one which is both received and responded to by other performers as well as by the audience.

To this end we required that the gestural quality of the actions involved in producing the system's voice be emphasized without visually negating the technology and origins of the sound produced. The physical design of elements used to house the new mobile DIVAs system reflects these requirements.

Previously, GloveTalkII users were required to sit in a chair while wearing the Cyberglove®, TouchGlove and the

tracker, with the foot pedal conveniently placed on the floor close by. The stationary GRASSP system used an identical setup connected to a Macintosh desktop computer. For DIVAs we needed to make the system mobile and completely self-contained with software all running on a laptop computer located on the performer. To achieve this we modified each component as necessary to be self-contained, including adding power supplies.

All of the components of the GRASSP/Glove TalkII noted above were transferred to a wearable interface. Different means of integrating each component in the DIVA were considered when developing this new mobile format as described in the following sub-sections. Figure 1 is a diagram of these components.

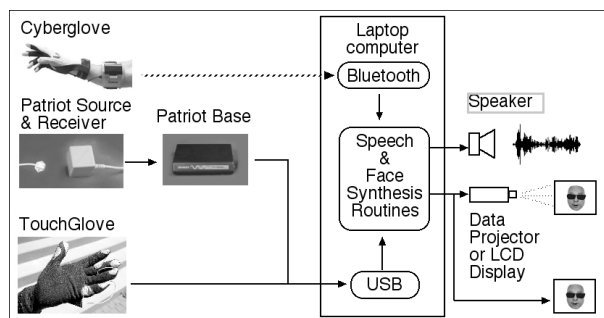


Figure 1: Diagram of the mobile DIVA components.

### 3.1. Hardware Components

#### 3.1.1. *CyberGlove®*

The Cyberglove® is a commercial product that provides measurements of eighteen or twenty-two finger, wrist, and abduction angles. This data is used to generate liquids, glides, fricatives, affricates and nasals – that is, all the English language consonants except the stops. It comes in a wired or a Bluetooth wireless version. We use the wireless version to remove the need for providing complex wiring to the performers hands from the Cyberglove® control device.

#### 3.1.2. *Polhemus Patriot® (Magnetic tracker)*

Currently, a wireless version of the magnetic tracker is not available that supports general mobility, and thus we developed a method to mount the wired tracking system on the performer. There are three main components: 1. the magnetic field source that is a 2" cube with a heavy wire attached, 2. a small receiver that is a 0.5" cube with a light wire attached and 3. a control box where all the wires attach. We created a custom harness (see section 3.6) to hold these. Of particular difficulty is that all the position and orientation measurements are relative to the magnetic source, and thus it must be mounted to the harness in a way that provides stability and repeatability. As well, stitching was required to route the wires from the back harness over the right shoulder and down the front body to connect the source and receiver without hampering movement.

#### 3.1.3. *TouchGlove*

The TouchGlove, shown in figure 2, consists of a customized Isotoner® glove with one touch pad on the thumb and two touch pads on each of the fingers. By placing the thumb pad in contact with one of the finger touch pads, a stop consonant, such as 'B', 'D' and 'G', is generated. Early versions of the TouchGlove used circular conductive fabric

pads sewn onto the surface of the glove to provide touch points. This method of attaching the touch pads on the top surface hampered the user's ability to easily locate and achieve contact with the pads during use. The redesign of this interface began with a consideration of the ergonomics of the gestures required of the TouchGlove: specifically, how the thumb's side interacted with all the fingers during closing and touching motions [9]. The result of this was a decision to change the thumb touch pad made out of Zelt Conductive Fabric® from a circular form to a tear shape. With this configuration, as the thumb moves and rotates across the finger touch points, the gradual variation in surface area allows for greater precision of touch.

A second change to the TouchGlove was the placement of the finger touch pads lower than the base fabric of the glove. This surface modulation allows for tactile feedback, providing the individual with non-visual cues as to the relative position of thumb and finger - a place to slot the thumb and finger pads together. A Twiddler2® could be used in place of the glove, but we chose to use a glove so that listeners/audiences could better see the hand shape for given sounds.



Figure 2: Contact surfaces on the TouchGlove.

#### 3.1.4. *Connect box*

The TouchGlove is connected to the laptop through a connect box. Each touch pad is wired to the box, and the placement of the wires called for a thoughtful approach to problems of system and aesthetics. Because the finger pads are located on the palm-side surface of the glove, wiring to the pads must be robust enough to accommodate a certain amount of movement. At the same time the wiring should not constrict the movements of the performer. An additional problem is the location of the connect box that dictates where must the wires go.

We placed the box close to the back shoulder on the left hand side of the person. A separate connecting wire was imbedded in the sleeve of the harness/jacket connecting the touch pad wires at the wrist to the connect box on the back. We considered other placements, such as on top of the wrist, under the wrist and part-way up the arm, but they were discarded due to either weight distribution, fabric twisting, aesthetic or fragility issues. An alternative approach would be to make the system wireless, however we did not have the technology available to make this easily.

#### 3.1.5. *Footpedal/On-Off Control*

Perhaps the most obvious problem in making the system mobile is finding a substitute for the foot pedal used in Glove-TalkII/GRASSP. Various sensors were considered including

breath, pressure, visual, and proximity sensors. Work with Glove-TalkII, suggested that an on/off mechanism is as effective as a continuous controller. Thus, we finally decided on expanding the touch pads on the TouchGlove to include a conductive strip along the edge of the palm. Another conductive strip was sewn into the left hand side of the harness/jacket. When in contact the two conductive strips send a signal to the system that turns off the speech synthesizer. Thus, moving the hand out of contact with the body turns on the speech synthesizer. The resulting gesture is one often evident in the body language of people when they begin speaking.

### 3.1.6. Speaker

Each DIVA configuration can have a variety of locations for the position of the single loudspeaker. Normally the loudspeaker is mounted on the upper chest as this provides the best position to correlate the speech with the visual information coming from gesture, posture, and other visual cues, but in performance it is possible to locate the loudspeaker elsewhere on the body depending upon the aesthetic, artistic, and practical demands of the situation. Regardless, the loudspeaker is treated as the equivalent of the human mouth, and if necessary on stage the sound being emitted can be amplified by a standard microphone setup. Notice that it is possible to directly take the line-out feed from the computer, but this violates the aesthetics of DIVA being self-contained.

### 3.1.7. Video Display

A small video display is required for displaying the face synthesis during performance. While we can project the face directly from the face synthesizer onto the stage, aesthetically, we want the whole system to be self contained. We plan to mount a small, 5" VGA LCD panel (VertexLCD) opposite the speaker placement. Note, that when DIVA is used as a speaking device (i.e. not in performance mode) it may not be necessary to have the face synthesis. We have not included this on the DIVA at this time.

### 3.1.8. Harness/Jacket

The aesthetic of the harness/jacket developed in response to the demands of the hardware being used combined with the stage performance and individual needs. Figure 3 shows the jacket from the front and back. The harness is designed to hold a laptop computer, a small magnetic tracking system with two sensors and a source, small rechargeable batteries, a connect box, and a piezoelectric speaker, with all equipment easily and quickly accessible for trouble shooting and performance. We included a rubber track system to lift the computer off the back of the individual to provide air circulation so that the laptop does not overheat. Additional heat deflecting capacity has been provided using an insulating foil coated layer on the surface of the harness.

While wearing the harness the individual needs to execute a variety of small and large hand and arm gestures across the horizontal and vertical planes of the body. In addition, as more than one user was envisaged to be using the system, it was essential that the design easily accommodate different body sizes, shapes and postures [9]. Fittings and pattern development of the prototypes reflect these requirements related to movement and body type along the lines of [10].

To secure the bulky tracker source, the lower part of the harness consists of a wide polyethylene belt fitted to the hips. The tracker source is attached with plastic bolts to the front

right of the belt. This placement allows the hips to carry as much of the weight as possible and prevents the source from moving relative to the wearer. The belt's front closure consists of Velcro® straps which allow for adjustments during the design and testing period.



Figure 3: Harness back and front views.

## 3.2. Software Components

DIVAs software components are based on the Glove-TalkII design coupled with the GRASSP sound control and run on a single laptop computer. We adjusted the sound control framework, added a face synthesis component and made the system user friendly to facilitate training by the performers for DIVAs. We describe these three components below.

### 3.2.1. Sound Mapping

As in Glove-TalkII, DIVAs use three neural networks that are training 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 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 respond to a hand posture associated with a cardinal sound such as an 'EE', 'L', or 'SH'. There are 11 cardinal vowel sounds and 15 cardinal consonants mapped this way. The left touch glove 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. The foot pedal controls the master volume. Collectively, all the formant outputs drive a parallel formant speech synthesizer. A LSI hardware research synthesizer was used in Glove-TalkII. This was too heavy and large to be used in a mobile platform and thus was re-implemented in software to run on the laptop in the Max/MSP environment.

### 3.2.2. Face Mapping

We have added a face synthesizer to DIVAs to provide additional expressive capacity. Currently, we have implemented a very simple approach to allow the performer to control face gestures coordinated with the sound gestures. We base our current gesture-to-face mapping on the principal component (PC) approach by [12] and drive the visualization by connecting our Max/MSP software to the artisynth toolkit [13]. In the same manner as we do for the sound control, we use the outputs of the normalized RBF functions from the vowel network and the consonant network to weight the

associated cardinal vizemes that we have created for each of the sounds. Thus, a weighted sum of the PCs for each of vizemes is sent to the face synthesizer running on Artisynt as the performer is speaking. Figure 4 shows an example of the current face used for the DIVA. We plan to investigate other means for controlling face representations.

At the moment, the visualization of the face appears on a computer monitor that can be projected anywhere at any size on the stage during performance. For the mobile platform, we intend to place a small, battery operated display on the performer to be consistent with its aesthetics.

### 3.2.3. Simplifying the User Interface

Two main user interface limitations in the Glove-TalkII/GRASSP systems require attention to make DIVAs effective. 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 have built a new user interface to simplify all of these tasks. Key to the simplification is the creation of user profiles that keep track of all the training sessions and configuration parameters for each user. This provides a coherent mental model for performers to know which parameters do what, what each mode meant and how to easily organize their training data. Our primary concern is to allow performers to focus on the task at hand, such as performing or training without spending time or cognitive effort fiddling with the software. We also include simple methods to monitor and check that the system is working correctly as we anticipate that during performance confirmation that everything is working will be critical.

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 DIVA, 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. We accomplish this 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 rather than optimize them. Thus, the network training time is short and works with as many examples as the user wants to supply. We are evaluating the effectiveness of this approach.



Figure 4: Example of some Vizemes (neutral, L and UU as in "boo") in Training Set (images from Artisynt software).

## 4. DIVA Aesthetics

Because the hands and arms create the gestures that generate the speech, we deemphasize the technology on the

front of the user while emphasizing it on the back. For this, most of the glove and tracker wiring on the front of the user is concealed by the sleeves and body of the jacket resulting in a "non-techno, non-futuristic" presentation and an emphasis of the gestures which control the speech production. Conversely, on the back various Velcro® straps harness the laptop, tracker, and batteries, holding the hardware securely while allowing for quick adjustments and easy access. The open harness emphasizes the technology behind the sound, especially with the wiring being clearly in view.

## 5. Current 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.

## 6. Conclusions

The development of a mobile audio-visual speech synthesis system controlled by gesture for use in performance and everyday speaking raises design and aesthetic issues and problems. Solutions to the problems must be responsive to – and supportive of – the performer's need to create speech and song as easily, effectively and expressively as possible, while still promoting and maintaining the vision of the clothing and performance designers' visions. Wearable technology and a simplified interface is critical to allow emphasis to shift away from a *user* using a system, as in Glove-TalkII, and a *performer* playing GRASSP to a *person* with a voice enabled with a DIVA. We believe that by continuing our design approach we will be successful in achieving all of these aims.

## 7. Acknowledgements

This work has been funded by an Artist-Researcher grant SSHRC and a New Media Initiative grant from the Canada Council for the Arts and NSERC. Support has also been provided by the UBC Media And Graphics Interdisciplinary Centre (MAGIC) and by the UBC Institute for Computing, Information and Cognitive Science (ICICS). We are grateful to the contribution from Helene Day-Fraser, Ying Yin, Allison Lenters, Marguerite Witvoet and the rest of the DIVA team.

## 8. References

- [1] Fels, S; Hinton, G.; Glove-TalkII: A neural network interface which maps gestures to parallel formant speech synthesizer controls. IEEE Trans on Neural Networks. Vol 9. No. 1, pp 205-212. 1998.
- [2] Pritchard, B.; Fels, S.; GRASSP: gesturally-realized audio, speech and song performance, NIME, 272-276, 2006.
- [3] Verbeek, P.; Materializing Morality: Design Ethics and Technological Mediation. Science, Technology & Human Values. Vol. 31, No. 3, 361-380, 2006.

- [4] Rye, J.M.; J.N. Holmes; A versatile software parallel-formant speech synthesizer, JSRU Research Report No. 1016, Nov 1982.
- [5] Walk the Walk, Talk the Talk, Day-Fraser, Helene, Fels, S., and Pritchard, R., International Symposium on Wearable Computing (ISWC), 2008, in press.
- [6] D'Alessandro, N.; D'Alessandro, N.; Le Beax, S.; Doval, B.; Real-time CALM Synthesizer: New Approaches in Hands-Controlled Voice Synthesis, NIME06, 266-27, 2006.
- [7] Cook, P. R.; Leider, C.; Squeeze Vox: A New Controller for Vocal Synthesis Models", International Computer Music Conference (ICMC), 2000.
- [8] Cook, P. R.; SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software Synthesis System, Computer Music Journal, 17:1, pp 30-4, 1992.
- [9] Fels, S.; Designing for Intimacy: Creating New Interfaces for Musical Expression. Proc of the IEEE. Vol 92. No. 4, pp 672-685. 2005.
- [10] Knight, J.F.; Deen-Williams, D.; Arvanitis, T.N.; Baber, C.; Sotiriou, S.; Anastopoulou, S.; Gargalakos, M., Assessing the Wearability of Wearable Computers, ISWC, 75-82, 2006.
- [11] McCann, J.; Hurford, R.; Martin, A.; A design process for the development of innovative smart clothing that addresses end-user needs from technical, functional, aesthetic and cultural view points, ISWC'05, pp 70 - 77, 2005.
- [12] Kuratate, T., Vatikiotis-Bateson, E. and Yehia, H. C., Estimation and animation of faces using facial motion mapping and a 3D face database, in Computer-graphic facial reconstruction, eds. Clement, John G. and Marks, Murray K., Elsevier Academic Press, Amsterdam}, pp. 325-346, 2005.
- [13] Fels, S., Lloyd, J., Stavness, I., Vogt, F., Hannam, A., Vatikiotis-Bateson, E., ArtiSynth: A 3D Biomechanical Simulation Toolkit for Modeling Anatomical Structures. Journal of the Society for Simulation in Healthcare. Volume 2. No. 2. Pages 148. 2007.