In Situ Speech Visualization in Real-Time Interactive Installation and Performance

Golan Levin
Carnegie Mellon University
College of Fine Arts, CFA-300
5000 Forbes Avenue
Pittsburgh, PA, 15213 USA
+1.412.268.2000
golan@andrew.cmu.edu

Zachary Lieberman
Parsons School of Design
Design and Technology Dept.
2 W. 13th Street, 10th Floor
New York City, NY, 10011 USA
+1.212.229.8908
zlieb@parsons.edu

ABSTRACT
Although we can sense someone's vocalizations with our ears, nose, and haptic sense, speech is invisible to us without the help of technical aids. In this paper, we present three interactive artworks which explore the question: "if we could see our speech, what might it look like?" The artworks we present are concerned with the aesthetic implications of making the human voice visible, and were created with a particular emphasis on interaction designs that support the perception of tight spatio-temporal relationships between sound, image, and the body. We coin the term *in-situ speech visualization* to describe a variety of augmented-reality techniques by which graphic representations of speech can be made to appear coincident with their apparent point of origination.

Keywords
Art, interactive installation, audiovisual performance, speech visualization, speech analysis, augmented reality, head tracking, computer vision, phonesthesia, sound-image relationships.

1. INTRODUCTION
The topic of speech visualization has generally been treated as an issue that arises in the scientific disciplines of phonology and psychoacoustics, or more recently, computational audition. Practitioners in those fields have developed or refined a range of widely adopted and useful visualization techniques, including waveform graphs, spectral and formant plots, autocorrelograms, cochlear and pharyngeotracheal diagrams, etc., which help solve the problems those practitioners encounter. It is important for us to state at the outset, therefore, that the projects we describe here diverge considerably from such solutions, insofar as our speech visualizations are (A) explicitly non-utilitarian and (B) designed solely to establish a perceptually and aesthetically plausible, interactive fictional universe in which speech is somehow visible.

In creating such a fictional world, our work builds on the seminal design insight of what Ben Schneiderman termed "direct manipulation interfaces": that interactive systems work best when they eliminate layers of functional or spatial indirection, and thus merge the loci of control and display [8]. Borrowing a solution from the graphic language of comic strips, our projects attempt to do this, in the realm of interactive speech visualization, by shifting the apparent display-position of such visualizations to the same physical location as their source of control: the mouth. Although our techniques were developed for interactive artworks, our solutions for the coincident analysis and display of speech are in principle adaptable to more traditional or scientific problems.

The 3rd International Symposium on Non-Photorealistic Animation and Rendering, June 7-9 2004, Annecy, France.

2. BACKGROUND
At the heart of our investigation into artistic speech visualization is our interest in *phonesthesia*, or phonetic symbolism. According to this idea, the sounds of words tend to reflect, to some extent, associated connotations from other perceptual domains such as shape or texture. A classic illustration of the phonesthetic principle can be found in Wolfgang Köhler's pioneering psychology experiment [4] from 1927, in which he asked subjects, "which of the figures below represents the sound *maluma*, and which one represents the sound *takete*?" Nearly all viewers respond with the same answer—suggesting rich research opportunities for both cognitive psychology and artmaking. Although Köhler did not break down his results into the kinds of quantitative perceptual mappings that could be profitably used in a generative graphics algorithm, his experiment suggests that high-frequency spectral content could be a good starting point for the creation of synaesthetic mappings between shape and sound. In point of fact, his work forms an important underpinning for many of the mappings we implemented.

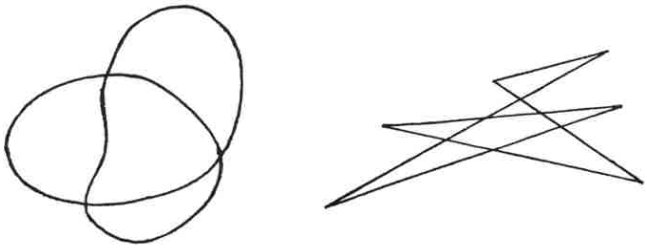

Figure 1. Kohler's 1927 phonesthesia experiment [4].

Several artists have explored the aesthetic possibilities of animated visualizations of speech. One important example of this is the Oscar-nominated short film *Reci Reci Reci* (*Words Words Words*) by the Czech animator Michaela Pavlátová [6]. In Pavlátová's delightful hand-drawn animation, the conversations of café patrons are represented as various kinds of abstract shapes and symbols that emerge from their mouths. These symbols are not only synchronous with the characters' speech sounds, but, in their forms and colors, reveal something about the characters' inner states as well. Grady Klein's animation *Afterbabble* accomplishes a similar result with more typographic means [3].

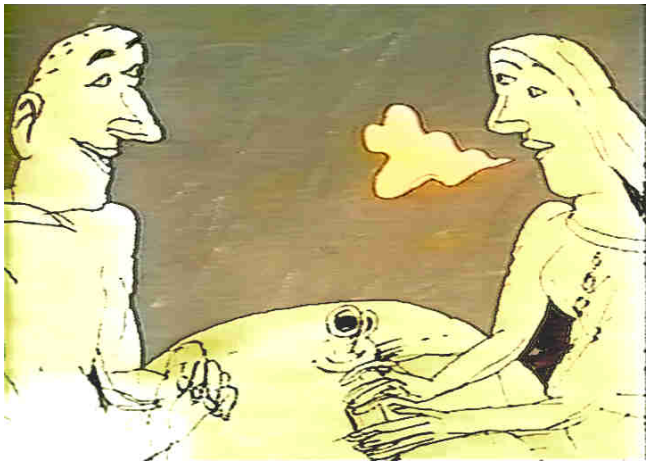

Figure 2. Pavlátová's animation *Reci Reci Reci* [6].

In the realm of interactive art and performance, the Japanese artist Toshio Iwai is well known for a long career of inventive and high-quality audiovisual works. In 1997, Iwai produced the masterful audiovisual performance *Music Plays Images X Images Play Music*, in collaboration with the composer Ryuichi Sakamoto, which explored a wide variety of possible relationships between physical sound-actions and their virtual graphic representations [2]. For this work, Iwai devised a performance system in which a projection screen, arranged in immediate proximity to a grand piano, displayed real-time animated graphics that appeared to emerge from the piano whenever one of its keys was sounded. Iwai has since developed a new performance, *Sho of Light* (2003), which employs a similar real-time audiovisual co-location between projected computer graphics and a Sho (traditional Japanese flute) performed by a leading Japanese flautist [1]. Although Iwai has published few details of this new work, its mechanism may in principle be quite similar to that used in our own performance, *Messa di Voce.*

3. INTERACTIVE ARTWORKS

In the summers of 2002 and 2003 we developed three interactive artworks (two installations and a performance) which dealt with the artistic possibilities of speech visualization. Although these artworks use many different graphical techniques to represent speech visually, they are motivated by a common goal of making these visual representations seem, perceptually speaking, to be as tightly coupled as possible to the speech sounds with which they are associated. To achieve this, all of the artworks involve the generation of graphic representations in real-time (that is, produced at very nearly the same instant as the utterance), and, additionally, in "real-space" or *in-situ*: that is, displayed so as to appear to emerge from the mouth of their associated speaker. In order to achieve such *in-situ speech visualizations*, we employed technologies ranging from stereographic 3D goggles with electromagnetic position sensors, to computer-vision-based tracking and projection systems.

3.1 Installation One: *Hidden Worlds*

We began our artistic investigation into speech visualization in the summer of 2002, when we were invited by the Ars Electronica Museum in Linz, Austria to develop an interactive installation on the theme of augmented reality. The result was *The Hidden Worlds of Noise and Voice*, or simply *Hidden Worlds*, which was installed at the museum for the 2002-2003 season.

3.1.1 Overview of "Hidden Worlds"

Hidden Worlds is an interactive audiovisual installation, or, alternatively, an augmented-reality speech-visualization system. Participants in *Hidden Worlds* are able to "see" each others' voices, which are made visible in the form of animated graphic figurations that appear to emerge from the participants' mouths while they speak.

In the installation, visitors wear special see-through data glasses, which register and superimpose stereoscopic 3D graphics into the real world. When one of the users speaks or sings, colorful abstract forms appear (through the goggles) to emerge from his or her mouth. The graphics representing these utterances assume a variety of shapes and behaviors that are tightly coupled to the unique qualities of the vocalist's volume, pitch and timbre.

Hidden Worlds permits up to six visitors to participate in the consensual hallucination, enabling a wide range of engaging audiovisual and conversational play. For those who are not equipped with the data-glasses, a projection at the center of the installation makes visible the "shadows" of the virtual spoken forms.

Figure 3. Modified Sony *I-Glasses* in *Hidden Worlds*.

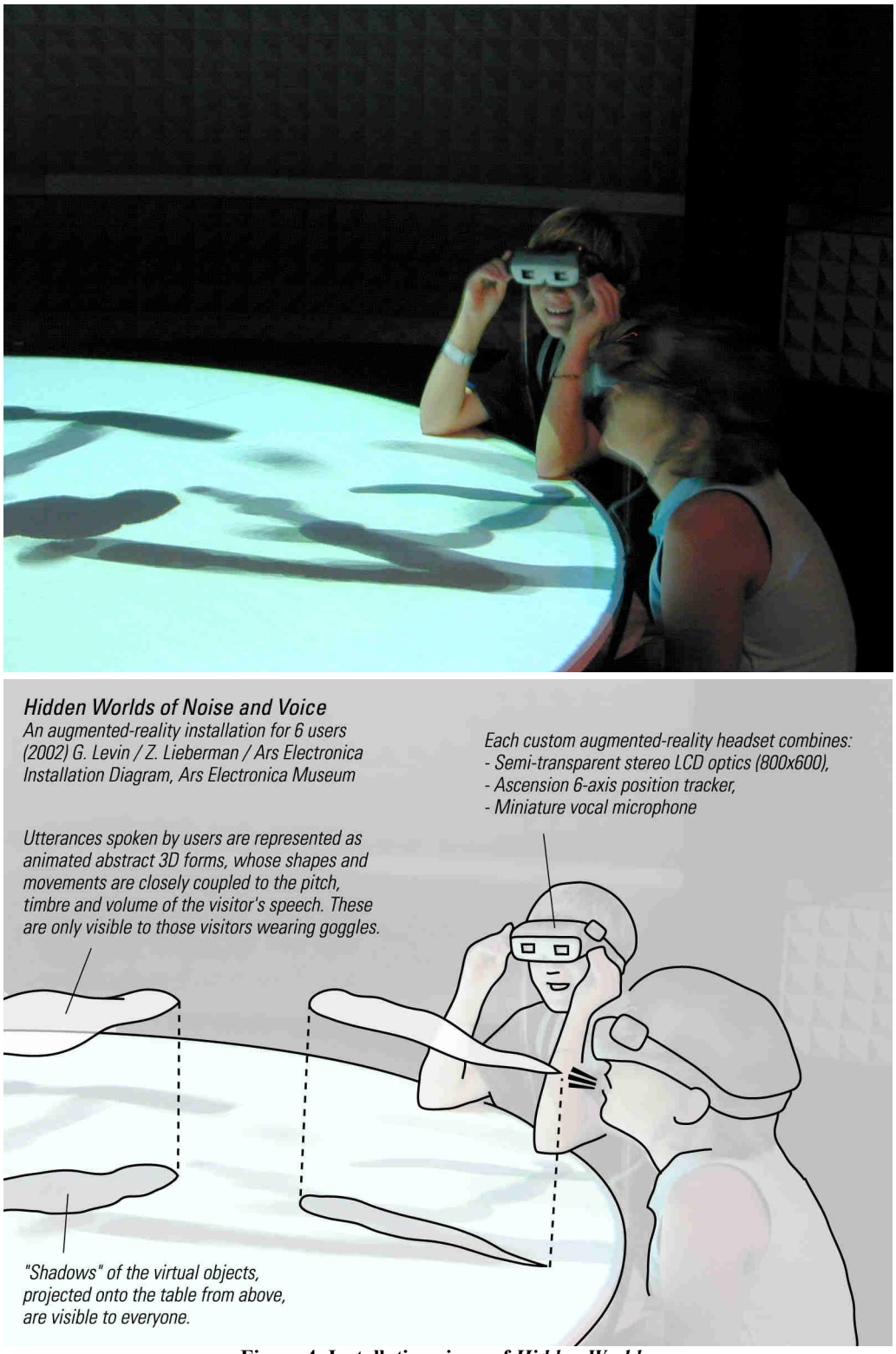

Figure 4. Installation views of *Hidden Worlds.*

3.1.2 Implementation of "Hidden Worlds"

The *Hidden Worlds* installation offers an *in-situ* speech visualization through the use of *augmented reality* technology, defined by Lev Manovich as the "overlaying of dynamic and context-specific information over the visual field of a user" [5].

To accomplish this, we use headsets consisting of a home-brewed combination of an Ascension *Flock of Birds* 6-degree-of-freedom position tracker; Sony *I-Glasses* which had been modified (with additional view-holes and custom half-silvered optics) to be semi-transparent, and a miniature AKG vocal microphone mounted within the bridge of the Sony glasses. With appropriate coordinate offsets, this combination of technologies allows us to surmise the location and orientation of our visitor's viewpoints and mouths. The viewpoint estimates are used to provide our visitors with personalized, geometrically "correct" views of the shared 3D synthetic world, while the mouth position estimates are used as the origination points for the localized 3D speech visualizations.

Visitor utterances are segmented, analyzed and then represented graphically as noodle-like "sound-gestures." We employ a variety of speech analyses to parameterize the shapes of the sound-gestures, but the most perceptually salient of these are duration (mapped to the visual length of sound-gestures) and volume (mapped to changes in the sound-gestures' diameter). When a sound-gesture is created, it initially emerges from the location of its speaker's mouth; thereafter, however, it gradually submits to the influence of a flocking simulation [2], which directs it to swim around the visitors' heads. The character of its flocking behavior is influenced by other aspects of its associated utterance (such as pitch and spectral centroid).

Figure 5. A view photographed through a *Hidden Worlds* eyepiece. At upper right, the underside of a synthetic "sound gesture" superimposed into the scene; at lower left, the "shadow" of the sound gesture cast on the table by the overhead video projector; at upper left, the child (elbows on table) who created the sound gesture.

One of the innovations of the *Hidden Worlds* system is the use of a video projector to augment the narrow visual field of the visitors' stereographic goggle displays. Like many consumer-grade (i.e. non-military) 3D goggles, the Sony *I-Glasses* suffer from an extremely small (15-degree) field of view; using such a viewport is roughly analogous to inspecting the world through a postage-stamp sized hole positioned 10cm in front of one's face. By projecting the virtual "shadows" of the virtual 3D forms onto a broad central table, each visitor's peripheral vision was extended considerably: the *I-Glasses* provided a detailed and colorful foveal view, while the projection around and beneath the goggle views provided contextual information about the location and movements of surrounding sound-gestures. The table projection has the additional benefit of providing an enticing partial view for museum visitors who have not yet participated in the installation.

3.2 Installation Two: *RE:MARK*

A smaller installation entitled *RE:MARK* was produced as a companion piece to the larger *Hidden Worlds* project. Designed for only two participants, *RE:MARK*, like *Hidden Worlds*, presents an interactive visualization of its users' speech. Unlike *Hidden Worlds*, which primarily attends to the noiselike aspects of vocal sounds, *RE:MARK* shifts this inquiry towards the more symbolic domain of the spoken and written word.

3.2.1 Overview of "RE:MARK"

In *RE:MARK,* sounds spoken into a pair of microphones are analyzed and classified by a phoneme recognition system. When a phoneme is recognized with sufficient confidence, the written name of the phoneme (for example, *oh*, *ee*, *ah*, etc.) is projected on the installation's display. If the user's sound is not recognized by the system's classifier, then an abstract shape is generated instead, according to parameters derived from the timbral (spectral and formant) characteristics of the vocalization. Among other mappings, sounds with high-frequency spectral centroids are represented with pointier, more irregular forms.

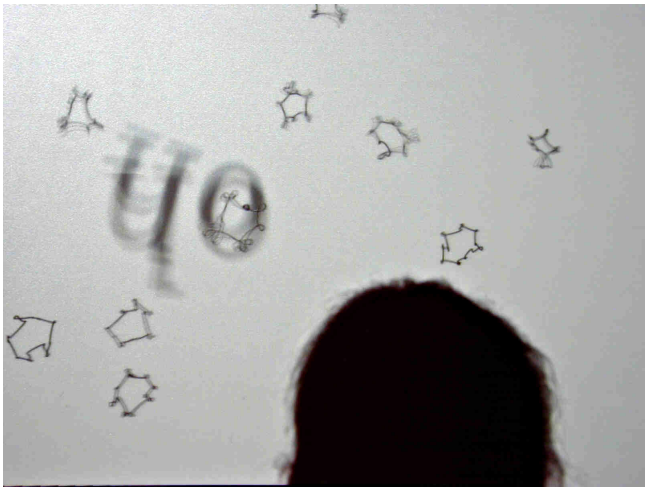

Figure 6. Installation view of *RE:MARK*'s screen, showing a combination of textual and graphic shapes emerging from the shadow of the participant's head.

As the visitor speaks, the corresponding written phonemes and abstract forms are rendered as silhouettes. These graphics are animated such that they appear to emerge from the shadow of the speaker's head, as determined by a computer-vision system. The result is a playful fiction, in which the ostensibly "magic light" of the video projector makes the shadows of one's speech visible.

The installation's visitors become actors in a shadow world of reactive cartoon language.

3.2.2 Implementation of "RE:MARK"

RE:MARK employs an elementary phoneme classifier, which works by comparing the Mel-Frequency Cepstral Coefficients (MFCC) of the user's vocal signal with those from a set of pre-stored phoneme models, and selecting the best match according to a least-squares difference. As such, the system is able to recognize about a half-dozen vocal sounds with approximately 90% reliability across a broad range of speakers.

The speech analysis system works in tandem with an equally simple computer-vision system, which estimates the location of a user's mouth by selecting a point somewhere above the centroid of the user's thresholded silhouette. When a user speaks, graphical visualization elements, consisting of alphabetic text fragments (when a phoneme is recognized) and timbrally-parameterized abstract polygons (when no sound is recognized) are birthed by the vocal analyzer and placed at the presumed mouth position.

A blurred version of the user's silhouette is additionally used, by the graphics animation system, as a gradient field for the application of forces to the free-floating graphical elements. These forces propel the visualizations away from (the shadow of) the speaker's head, thus making room for new elements. This implementation also permits the speaker an additional interaction, in which it becomes possible to propel the graphics across the screen with the shadow of one's body.

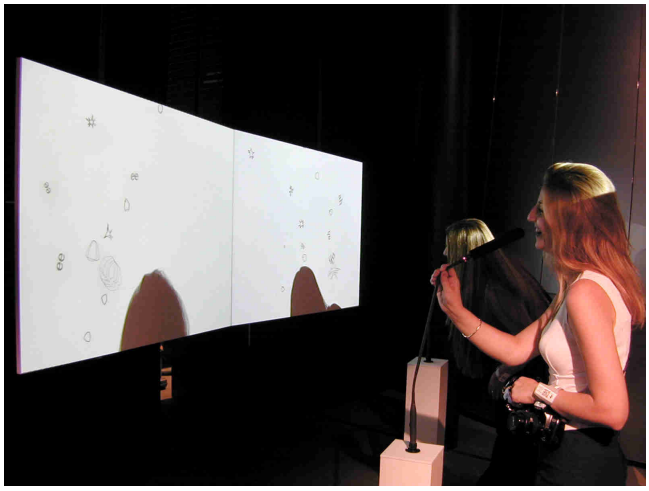

Figure 7. Installation view of *RE:MARK*.

RE:MARK was quickly executed during the same time period that we were completing the much more logistically complex *Hidden Worlds* installation. Plans to replace *RE:MARK*'s rather modest phoneme recognition algorithm with a more robust commercial solution, among other developments, are underway.

3.3 Performance: *Messa di Voce*

Our installations *Hidden Worlds* and *RE:MARK* were designed for easy apprehension by a lay audience. In this regard they appear to work extremely well, and seem especially popular with museum visitors under 10 years old (and other visitors who are not too shy to bark or oink). In the course of developing these installations, however, we realized that our software, with further refinement, had the potential to support even more nuanced interactions.

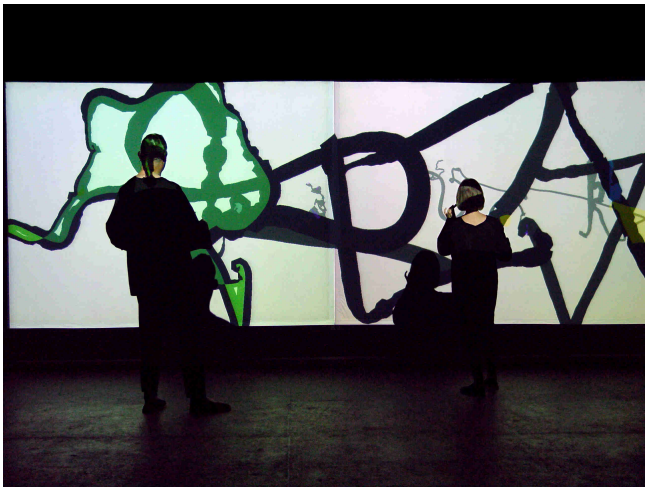

Figure 8. "Pitchpaint", a section of *Messa di Voce* in which the performers paint lines by modulating the pitch of their voice.

In summer 2003, we sought to take our core ideas about speech visualization to a new level of sophistication, by developing a "professional" version for use by a duet of virtuosic vocalists. Our foremost challenge in doing so, as artist-engineers, would be to create interactive systems that could be deeply instrumental for these vocalists, and *commensurately expressive*. The result of our effort was *Messa di Voce*, a concert performance in which the speech, shouts and songs produced by two vocalists are augmented in real-time by custom interactive visualization software. Created in collaboration with Joan La Barbara and Jaap Blonk—two singer/composers known for their experimental vocal techniques—the performance touches on themes of abstract communication, synaesthetic relationships, cartoon language, and writing and scoring systems, within the context of a sophisticated and playful audiovisual narrative.

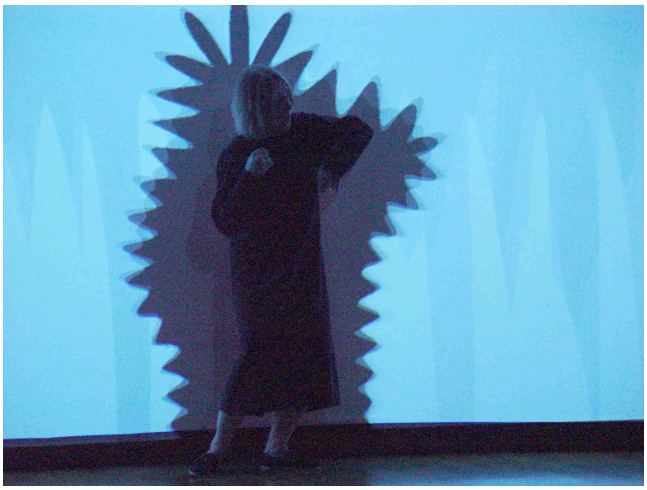

Figure 9. "Insect", a section of *Messa di Voce* in which a performer's silhouette is affected by her speech and song.

The core mechanism of *Messa di Voce* is similar to that of the *RE:MARK* installation, incorporating an integration of real-time computer vision and speech analysis algorithms. In *Messa di Voce*, a combination of computer vision techniques is used, not only to track the locations of the performers' heads, but to estimate the

orientations and positions of their bodies as well. The analysis computer also captures the audio signals coming from the performers' microphones, and extracts features such as pitch, spectral content, and autocorrelation data. In response, the computer displays various kinds of visualizations on a projection screen immediately behind the performers; these visualizations are synthesized in ways which are tightly coupled to the sounds spoken and sung by the performers. Owing to the head-tracking system, these visualizations can be projected such that they appear to emerge directly from the performers' mouths.

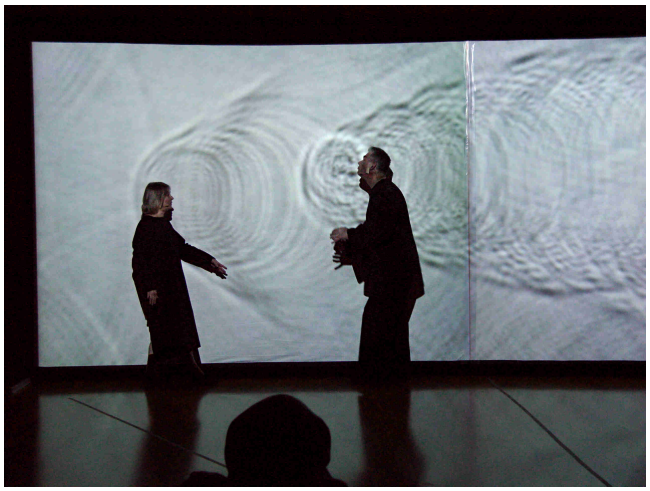

Figure 10. "Ripple", a section of *Messa di Voce* in which the performers' voices affect a water-ripple simulation.

In some of the visualizations, projected graphical elements not only represent vocal sounds visually, but also serve as a playable interactive interface by which the sounds they depict can be re-triggered and manipulated by the performers. An example of such an interaction can be found in Jaap Blonk's solo, in which the sounds he creates are represented by small black circles. When Jaap subsequently "touches" one of the balls with his shadow, the ball "releases" a playback of the sound associated with it.

Figure 11. "Jaap's Solo", a section of *Messa di Voce*.

The *Messa di Voce* concert makes use of a wide range of different graphic visualizations, including particle systems, elastic spring meshes, fluid simulations, cloud simulations, etc. Several of our techniques were adapted or implemented directly from recently-published computer graphics articles, such as Jos Stam's well-known "Stable Fluids" paper [9]. In our concert, these techniques and ideas take shape in a series of twelve brief vignettes which explore different symbolic, tactile and audiovisual aspects of phonesthetic relationships. The entire *Messa di Voce* performance generally runs 30 to 40 minutes in length.

More information about these projects can be found at:
`http://www.tmema.org/messa`
`http://www.flong.com/remark`

4. ACKNOWLEDGMENTS

The works described in this paper were commissioned by the Ars Electronica Festival, and made possible through the generous support of SAP, the Siemens Artist-in-Residence Program at the Ars Electronica Futurelab, Art+Com AG, la Fondation Daniel Langlois, Eyebeam Atelier, Ars Electronica Futurelab, the Lower Manhattan Cultural Council, the Rockefeller MAP Fund, and the New York State Council on the Arts. This work would not have been possible without the unflagging faith, support, and inspiration of Gerfried Stocker and Horst Hörtner of the Ars Electronica Center, Linz; and our collaborators Jaap Blonk and Joan La Barbara.

5. REFERENCES

[1] Iwai, Toshio and Miyata, Miyumi. *Sho of Light: The sound of the Sho returned as light.* Audiovisual performance at Digital Arts Festival, Tokyo, 2003.

[2] Iwai, Toshio and Sakamoto, Ryuichi. *Music Plays Images X Images Play Music.* Audiovisual performance at *Ars Electronica Festival*, Linz, Austria, 1997.

[3] Klein, Grady. *Afterbabble: Or, what happens to our words when we're done with them*. Self-published animation, 2003. http://silvertone.princeton.edu/~grady/

[4] Köhler, Wolfgang. *Gestalt Psychology.* Liveright Publishing Corporation, New York, 1947.

[5] Manovich, Lev. *The Language of New Media*. MIT Press, 2001.

[6] Pavlátová, Michaela. *Reci Reci Reci* ("*Words Words Words*"). Color animation, 8 minutes, 35mm. Kratky Film, Prague, 1991.

[7] Reynolds, Craig W. Flocks, Herds, and Schools: A Distributed Behavioral Model, in *Computer Graphics*, 21(4) (SIGGRAPH '87 Conference Proceedings), 1987, 25-34.

[8] Schneiderman, Ben. Direct manipulation: A step beyond programming languages. *IEEE Computer*, 16:5743, August 1983.

[9] Stam, Jos. Stable Fluids, in *Computer Graphics*, 33, (SIGGRAPH '99 Conference Proceedings), 1999, 121-128.

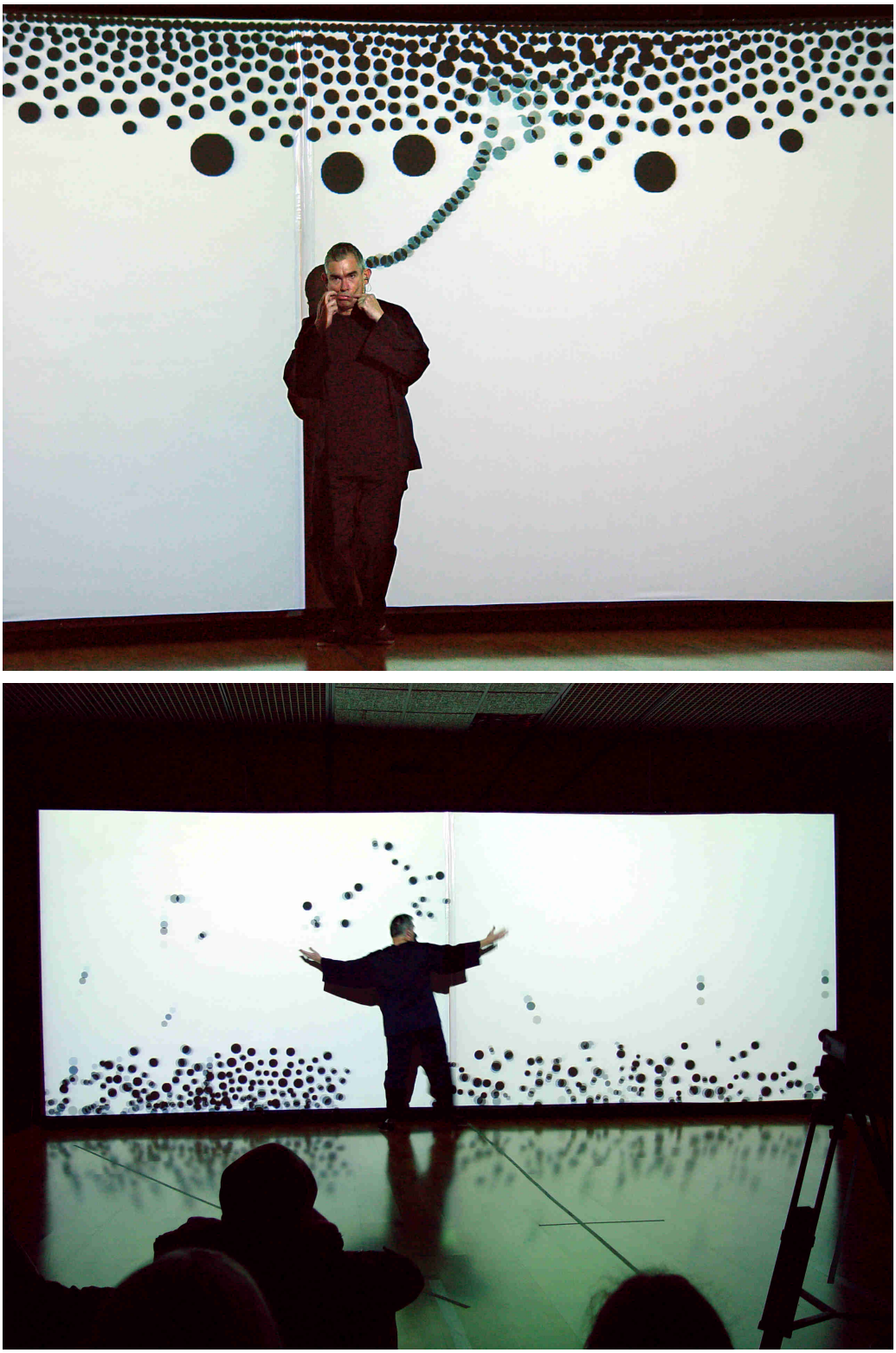

Figure 12, 13. More images from "Jaap's Solo" section of *Messa di Voce*. In the upper image, Jaap emits a stream of bubbles by making a special cheek-flapping sound. As his sounds grow more vigorous, his bubbles fill up the screen. But the resulting cloud of jostling sound-bubbles is unstable. Turning to admire his work, his cloud bursts — raining bubbles that "release" his cheeky sounds when they fall onto him or crash to the ground below. In the lower image, he struggles to contain the noisy torrent, but, failing this, storms off in distress.