Crosstalk: Making People in Interactive Spaces

Simon Biggs
University of South Australia
School of Art, Architecture and Design
Adelaide, SA, 5000, Australia
+61.8.83020418
simon.biggs@unisa.edu.au

Sue Hawksley
Independent dance artist and researcher
Adelaide, South Australia
sue@articulateanimal.org.uk

Garth Paine
Arizona State University
School of Art, Media and Engineering
Tempe, Arizona, USA
+1.480.9650972
Garth.Paine@asu.edu

ABSTRACT

Crosstalk is an interactive performance work by media artist Simon Biggs, choreographer Sue Hawksley and composer Garth Paine. The work employs real-time multi-modal sensing and interaction systems, including three-dimensional tracking of multiple performers combined with multi-source voice recognition for speech to text and an interactive multi-channel data driven sound score.

The three artists have previously collaborated on *Bodytext*, an interactive multimedia solo performance work in which a spoken (described) and performed dance are simultaneously interpreted by both the performer and the computational system. *Crosstalk* developed out of the processes undertaken in *Bodytext*, specifically the 'drama of the performance', which arose from an antagonistic but interdependent human/machine relationship. Created for two performers, *Crosstalk* engages social relations as articulated through performative language acts. The project explores ontologies of self-hood within the generative potential of a technologically mediated social space. The elements in the system, including performers and machines, affect how each adapts from state to state, as the various elements of the work - language, image, movement and sound - interact with one another.

Developed as an enactment of the proposition of 'making people', inspired in part by contemporary anthropological ideas, *Crosstalk* begins with two dancers speaking descriptions of each other. Automatically transcribed in real time into a virtual three-dimensional world, using speech to text software, these descriptions become textual objects that inhabit the environment and interact with other elements within the system, both human and non-human. The spoken texts form the foundation for an evolving sonic environment. When moving the performers collide with the text objects, causing them to also move. As the text objects interact, they re-write each other, facilitating the emergence of new textual and sonic material, created through the recombinant computation of the texts in the collided objects. Through this generative mechanism, the interaction situates each dancer as a product of their initial perception, the evolving environment, their interaction with it and their interrelationships. *Crosstalk* thus presents the multidimensional emergent properties of perception, interaction, place making and identity.

MOCO'14, June 16-17 2014, Paris, France.

http://dx.doi.org/10.1145/2617995.2618006

In both *Crosstalk* and *Bodytext,* the performers are enmeshed in a public/private drama within an interactive system, to which the audience is witness. Movement gesture is further enacted through multi-channel spatialisation of audio, enveloping the audience in the 'dance as sound' through the placement of the loudspeakers around them. As the dancers move, bodily and vocal sounds are acquired and processed, the resulting sounds and texts dynamically shifting around the audience. The spatialisation of texts and sounds immerse the audience within the morphology of the dancers' gesture and the act of revealing the dancers' inner voice. In some presentation scenarios (such as art gallery settings) the audience members can interact directly with the *Crosstalk* system and become part of the ecology of the work.

Categories and Subject Descriptors

J.5 [Arts And Humanities]: Performing arts (e.g., dance, music)

Keywords

multi-modal full body interaction, speech recognition, immersive performance environment.

1. INTRODUCTION

Language, a key factor in defining the human, is a central thread in *Crosstalk*. Language, in all its forms, allows us to represent, even create, our world and ourselves. Within *Crosstalk* speech and physical movement are captured and dynamically reconfigured through a system where each acquired element can modify another, leading to the generation of novel elements that could not have been foreseen prior to their emergence. A symbolic and abstract ecology evolves through the use of both interpretative and generative grammar systems. This is an environment possessing generative properties that facilitate people's interactions and engagement with one another and the creation of new elements emergent from the multimodal dynamics and recombinant nature of the system.

Crosstalk is an interactive performance environment where participants (whether performers who have rehearsed to become expert *interactors* or casual viewers visiting for a relatively short period of time) are able to interact, through various technical systems, with each other and virtual graphical and aural elements. The input technologies employed include three-dimensional infra-red motion tracking (Microsoft Kinect), voice acquisition (wireless INVISIO® M3 in-ear bone conduction headset) and speech recognition (Dragon Dictate), allowing for multi-modal interaction through movement, speech and realtime sound synthesis (Symbolic Sound, Kyma[1]). Output technologies include large-scale multi-screen video projection and multi-channel surround sound, facilitating an immersive experience.

[1] See http://www.symbolicsound.com viewed April 11, 2014

Custom software was developed employing the Processing JS development environment for the gesture tracking and visual environment and Kyma for the real-time sound, both running on Mac OSX and communicating over Open Sound Control (OSC). The Processing development included wrapping of the JBullet physics engine as a Processing library (this development work undertaken by Simon Biggs and Hadi Mehrpouya at the University of Edinburgh), which facilitated full-body physical interaction between the *interactors* and other virtual objects within a three-dimensional virtual world employing full physics simulation. Two three-dimensional worlds are computed (one per dancer) and video projected at ninety degrees to each other, rendered on large-scale screens placed around the *interactors*. The position of the text objects, collision events from each virtual world and calculated height, velocity and joint location data for each dancer is sent over OSC to Kyma and used to drive realtime sound synthesis and sound spatialisation.

The virtual visual space is primarily, although not exclusively, composed of three-dimensionally rendered graphical textual objects. These are created using real-time dictation; when the *interactors* speak, the words appear in the virtual space co-located with the speaker's head. Keywords also trigger the playback of whispered echoes of each spoken text in the audio system. The movement of the *interactors* causes the text objects to move. On collision, text objects (sentences) read one another, exchange textual elements and rewrite themselves, the texts therefore evolving as they move about within the virtual space, interacting with one another and the *interactors*. This recombinant process employs interpretive and generative grammar systems that allow the emergence of a dynamic language environment. The text objects also become stimuli for sound composition when they make contact with the *interactors* and when keywords are either spoken or emerge from the recombinant process. Furthermore, their location within the three-dimensional world is used to drive sound spatialisation algorithms, using vector-based panning around a multi-channel surround sound loudspeaker arrangement.

2. TECHNICAL DEVELOPMENT

The software development for the gesture tracking and visual environment of Crosstalk was undertaken in the Processing JS programming language. Processing was chosen as it is an interpreter language allowing rapid prototyping of software within an object-oriented programming architecture. The OOP architecture was the preferred mode for development as it facilitates development of multi-agent systems where numerous instances of objects are to be produced, on the fly, through unforeseeable occurrences of interaction between diverse agents (human and non-human).

Class structures were created that allowed for the run-time instantiation of models of individual *interactors* as software agents and on-the-fly instantiation of the text objects derived from the *interactors*' speech (acquired employing wireless microphones with Dragon Dictate speech recognition software, functioning system-wide). The instantiation of the graphical elements that compose, at certain points in the performance, the visualisation of interactions between the multiple agents within the system, were handled by another Class. Three-dimensional data of the *interactors* was acquired from the performance environment via two Microsoft Kinect infrared 3-D sensors. The SimpleOpenNI library was used to process and interpret the three-dimensional data sets acquired from the Kinect sensors and this data was used by the instantiated skeleton models of the *interactors* to control their graphical actions (during the performance these models are not visible).

A Class was defined for the interpretative grammar system, which each individual text object inherits in order to read other text objects and modify itself. This 'reading' process involves each text creating a grammatical model of the texts it interacts with and comparing that model with it's own internal grammar model. This comparison allows texts to exchange syntactic elements, allowing for changes in the semantic form of the text without significant change to the syntactic structure. A dictionary of words is included in the software, based on a general but relatively modest English vocabulary augmented with all the words created by the choreographer and dancers during the development of the performance, ensuring enhanced speech to text accuracy and correct syntactic analysis by the system.

To allow for the creation of a convincing three-dimensional physics simulation the Java-port of the open-source game development Bullet Physics Library was wrapped as a Processing library and included in the Processing application development environment. This allowed all the graphical objects in the system, rendered using the OpenGL library, to be modeled in parallel in the Physics engine. The models of the *interactors* and the text objects derived from their speech were rendered in this way. Once modeled, the outcomes of the interactions of the objects within the physics engine were retrieved through a callback function and employed to control the objects in the final OpenGL rendering.

Other key elements in the software included developing a Class to handle UDP communications between the multiple computers being used in the work, with care taken with the particular requirements for communication between the computers acquiring interaction data and visualising subsequent outcomes and the computer controlling the Kyma system, which produced the real-time three-dimensional audio.

All of the sound and music for Crosstalk was generated from the spoken text of the dancers. Invisio bone conduction headset microphones were employed for voice acquisition because they are not sensitive to airborne sound and thereby do not transmit the sound from the loudspeakers back into the system. Paired with Shure body-pack wireless microphone transmitters this approach gave clear voice commands and source material even when the loudspeaker sound levels were quite high. The music composition comprised realtime analysis and re-synthesis of spoken words, including spectral manipulation, the synthesis of musical textures using formant and resonant filters on noise sources, stutter effects, the playback of whispered samples of the dancers voices and granular synthesis which, in some cases, also included morphed spectrum using large oscillator banks, where the formants and frequency deviation of the timbre of the realtime voices was manipulated by realtime skeleton data (e.g. the distance between hands and between feet, position in space and inter-relationships of the dancers). The text manipulation was grouped according to *doing* words and descriptive or *ID* words. Pre-recorded whispered versions of key descriptive words were triggered by those words being spoken by the dancers and were often replayed simultaneously with the live speech (delayed by the recognition system) and in a spectrally altered manner (heightening the whisper tones and extending the textures) or using a randomised stutter effect driven by the dynamic of the dancers activity.

All of the sound was created in realtime, with the synthesis algorithms laid out in a timeline but with the time of execution controlled by the dancers' movements, such as thresholding of head height or velocity of movement. A *wait-until* function would look for these conditions and dynamically instantiate the next set of algorithms, which could include new sounds and/or

augmentation of existing algorithms. All of the dancers movement data was received from the tracking computers over OSC, with scaling, smoothing and thresholding completed in Kyma. For this purpose a *CrossTalk_GlobalOSC* class was developed which contained all of the data, direct and cooked, and created global variables which could be accessed by any sound algorithm.

The musical score was devised as timbrally distinct layers, each of which had a unique characteristic of *liveness* in the space. The morphology and positioning, density and spectral content were composed so that each element acted on a body or agent within the physical space, producing a visceral improvisation with the dance. To this end, four spatialisation algorithms were developed which produced different characteristics of movement through the sound field from jittery and fast to fixed with slow movement, to constant and dependable movement in addition to spatialisation, which was driven directly from the position of the dancer in the space. Spatialisation was achieved by placing loudspeakers around the audience so as to create an immersive environment using vector-based panning through the sound field.

The key principle that underpinned the process of software development throughout the project was a rigorous adherence to an OOP approach to multi-agent interaction. This was important not only from a technical point of view, aiding clear and concise development, but also artistically, reflecting in the structure and conceptualization of the software the key philosophical idea that underpins the work - the apprehension of human and non-human *interactors* as agents defined by and defining of networked relations, as made apparent through movement and speech.

3. PERFORMANCE

The choreographic form of the work addresses questions of presence and perspective. The performance is an improvisation, comprising four different scores. These were developed in conjunction with the technical systems by choreographer and performer Sue Hawksley, working with dance artists Lucy Boyes (in residency at the Bundanon Trust, NSW Australia), and Angel Crissman and Michaela Konzal (at Arizona State University). Each section presents a distinct perspective on the performers - first to third person, individual and inter-related. The following outline addresses the sections in a linear order; however the performers may use voice commands to select and change between sections in any order.

The first texts spoken in Crosstalk reveal information about what can already be seen by the audience and performers, e.g. visual references to the two performers' frame, height, eye and hair colour and identifying features such as “a small red tattoo on her left wrist”. This encourages an objective method for looking at people which does not require language to provide information. The performers' movement gestures, however, reveal whispered phrases in the sound-field containing private information that can only be known if it is told. The phrases include names, place of birth, likes and dislikes, pets and friends, and descriptions of characteristic movement preferences; “diving into space to open up and roll”, “poppy bursts and flurries of energy”, "establish the space with lines and curves”. Hearing these movement preferences described reinforces what can already be seen in the dancers' movement patterns. This is reinforced from an external, third-person perspective through the triggering of whispered samples of these spoken phrases and subsequent spectral manipulation and spatialisation, revealing these phrases as sonic objects controlled by the distance between and velocity of hands and feet, head and upper body, as tracked by the Kinect sensors. The performers have some degree of control over what is heard, for example by opening and then reducing the width between their hands before a phrase is fully spoken, but a randomised algorithm also re-synthesises the pitch and timbre of the voices, generating a richly textured sonic environment.

In another section known as the *Spine score*, the spoken texts describe the activity required to engage the system; “do only what is necessary”, “keep the action plain and simple”, “read and respond to the words”, “avoid the focus of attention getting stuck”. The performers' spoken words get written onto the two screens of the projected display. The locations of the written texts on the displays correspond to the location in the three-dimensional virtual space they and the performers occupy. Other texts indicate effective ways to navigate the virtual space and to locate text-objects; “transport all of yourself to find the point of contact”, “add just enough extension and impulse to reach”, “start patterns to create whirlpools”. Additional sounds may be triggered/created when the *interactor's* body makes contact with a virtual text-object or when objects meet other objects. Upon meeting, text-objects read and rewrite each other, thereby generating new elements for the score. Using a spoken command, objects can be returned to the spatial location where they were first spoken, while another voice command causes the speech to be recorded and replayed whenever that location is revisited, affording the performers the capacity to construct a textual environment with known landmarks (evoking the concept of a spatialised memory-theatre). An array of sampled speech is stored in association with an X/Y coordinate location. When the dancer re-visits that location the original spoken text is the trigger to replay, whilst simultaneously the resultant audio stream becomes source for realtime sonic processing and spatialisation driven by the momentary quality of activity at the time of playback.

When both *interactors* occupy the floor-area tracked by both Kinects their combined presence allows for the connections between them to be visualised as drawn lines, generating a third 'figure' which is animated by their collective actions. For different modulations of the figure, the performers organise their movement according to fundamental patterns such as Head-Tail Connectivity and Core-Distal Connectivity (Hackney, 1998), which lend different qualities and form to the improvisations. The visualised forms are similarly hybridised so that one figure is composed of the two upper torsos of the dancers and the other of the two lower torso components, shifting the differentiation of agency between them.

Further voice commands permit the relocation of the point of view from the external perspective to the head of the speaker, or the head of the other performer. This is achieved by positioning the virtual camera that is employed to visualise the on-screen three-dimensional scenes, at the location of the virtual model of a performer's head, the camera pointing towards the location of the other performer's virtual head. This allows the creation of a vector along which vision can be 'performed' by the two dancers. The movement possibilities of the performers in this section become highly contingent and interdependent. In order to read the words spoken by the other, one performer must adjust their position, and so then the other must move again. While making these constant readjustments of their relative positions, both must also attend to their position and facing relative to the Kinects, taking care not to occlude the other. At this time there are many collisions of body and text objects that cause the musical score to thicken in texture and become more free flowing and dynamic. In contrast to the improvisation in the *Spine score*, where the movement vocabulary tends to be abstract, in this *Connected score* the movement vocabulary is functional, determined by the practical requirement to see and to be seen and maintain a relative connection between

people and technology. In an environment where the contingent relations between things is foregrounded we can witness the system making itself make itself.

4. CREATIVITY AND AUTOPOIESIS

Anthropologist James Leach writes of how people make things, including people, through the social performative. He observes how the people of the Rai Coast of Papua New Guinea, through ritualised processes of exchange, 'create' individuals and bind each other in social groups. Leach has observed "the role of 'creativity' in the ways people generate new places in the landscape" (Leach 2003), for example through land management (gardening) and musical instrument making (drums), and has argued that,

> ... in so doing, they also generate new people, who emerge from these places, and objects which facilitate or even participate in these creative processes. Making people and places involves relations to other people and to spirits and ancestors that embody, through song/design/dance complexes, the generative potential of land itself. (Biggs & Leach 2004)

Emerging from a another intellectual tradition and addressing agency and generative systems in a very different manner, is the work of Humberto Maturana and Francesco Varela, whose concept of autopoiesis, developed in part in relation to Alan Turing's work on computability, represents a core concept in third order cybernetics. Drawing on empirical research in cellular biology considering how single cell organisms reproduce themselves, Maturana and Varela developed a theory for how symbolic systems could evolve with the characteristics of agency that Turing envisaged when he conceived of his Turing Machine. The term autopoiesis (from the Greek 'poiesis', to create) refers to the capacity for something to create itself.

> An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it (the machine) as a concrete unity in space in which they (the components) exist by specifying the topological domain of its realization as such a network. (Maturana and Varela 1980, p.78)

The *Crosstalk* project conceives of language as an autopoietic machine which creates language. However, it also treats people who are within and interact with it as equivalent with all the other elements in the system. As such, people are not distinguished as discrete elements within the autopoietic system - they are part of a holistic generative system that makes language as well as other things, including people.

In *Crosstalk* the aim is to create a space where people interact within a generative computational system, primarily but not exclusively manifested as an interpretative and generative language machine, and with each other. The manner in which the symbolic (computational) and biological (human) elements interact within the system are conceived of as a unity and no distinction of ilk is sustained - people, texts, graphical objects and sounds are all treated similarly and their agency considered equivalent in status. So far as the system is concerned there is no difference in the agency of the people or the symbolic objects (texts and other elements) that inhabit the environment. Each contributes to and is a function of the ecology of the work. As a whole, the environment is conceived of as autopoietic.

A design intention in Crosstalk is for the *interactor's* awareness of their contingent and constructed status to be foregrounded within the context of the environment and interactions they experience. The *interactors* are placed in a context where they can become aware of their agency in relation to the computational systems, their interactions with one another, and through their movement and speech and the resultant sonic environment. The distinction between how movement and speech can carry conscious and unconscious intent and be interpreted explicitly or implicitly, at the boundary of the liminal, is another important consideration in the work, especially within the context of understanding how people compose themselves through interaction in multi-agent environments.

5. ENGAGEMENT AND AGENCY

In the *Crosstalk* sonic environment, interactions are monitored and created through the remediation of the *interactors'* speech and their movement in the space, generating contextual cues representative of the energy of the interactions and the nature of the spoken and recombining emergent texts. The sound acts as an environmental constraint (Paine, 2007), both reflecting and constraining the nature of the interactions. The positioning of the projection screens and sound system is designed to highlight the audience's perception of themselves as part of the emergent ecology of the system. Movement and gesture is enacted through the multi-channel placement of the loudspeakers around the audience. Spatialisation of the audio is designed to envelop the audience in the 'dance as sound'. As the dancers move, bodily and vocal sounds are acquired and processed, the resulting sounds and texts dynamically shifting around the audience. The spatialisation of texts and sounds act as a similitude of agency, immersing the audience within a morphology of the dancers' gesture.

The validity of audiences watching others performing within interactive works, rather than participating in a work as an *interactor*, is often questioned. Isn't interaction primarily about the viewer interacting with the work? How can the viewer gain a sense of the interactive dynamic of a work if they are not directly engaged? If this dynamic is important in the comprehension of the work then surely the audience must have this direct engagement with it? *Crosstalk* is not conceived of as a performance only work. It also exists as an installation where viewers can directly interact with the work. After some performances audience members have entered to interact with the environment, augmenting their experience of the work. However, the performers have worked within the system for extended periods of time and thus developed a fine tuned understanding of its qualities, capabilities and limitations. This allows them, through their knowledge and insight, to reveal subtle aspects of the system that are emergent within it.

Audience feedback after performances indicated that the audience found it compelling to witness the dancers thinking 'on-the-hoof' as they sought to cope with the multi-tasking demands. Mark Coniglio, artistic director of Troika Ranch, suggests that the audience of interactive performances is present to the "on-the-fly artistry" of the performers (Coniglio 2005, p.8). The audience's understanding that:

> … the performer has a virtuosic command of his or her instrument and that he or she is creating something new in the moment of performance adds yet another layer of

> 'liveness' to the experience, and [that] this is … a core rationale for adding interaction to the mix in the first place. (ibid, p.6)

A strong argument for presenting *Crosstalk* as a performance to be viewed is that the artists can share their refined level of understanding of the potentials of the system, developed through a lengthy, in-depth exploration of the complexity and subtlety of the interactivity. Improvisation within an adaptive and emergent environment such as *Crosstalk* requires movement precision together with a fluency in attention to its numerous, often conflicting, sometimes liminal, elements. To facilitate this the choreographic research for *Crosstalk* incorporated training practices that functioned like a 'perception gym'. These included movement exercises to heighten sensory awareness of in-depth (internal), surface, situated and imagined body layers – bones, skin and space (both actual and imagined) – and involved explorations of the practices of listening, including blindfolded sound-walks and the mapping and drawing of sound-maps. The *Spine score* is written as a 'user-manual' - the performer saying and being told what to do. This also provides a method whereby non-expert users can participate in the system as moving *interactors*. Obviously, that the score rewrites itself makes for an unpredictable user-manual, where the directions emergent within the system are evolving in ways that are not only distinctive but sometimes challenging or even impossible to enact.

Generative language systems often lead to the emergence of illegible or unreadable texts. However, in *Crosstalk* the texts originate in human speech (not as the product of computer software or database) and, although subject to a recombinant modification that can render them illisible, remain as evidence of speech acts that are not only presented as a record but also intended to be read and acted upon by the performer. The requirement that the performer seeks to interpret the meaning of each text element and, furthermore, take direction from it to inform their subsequent actions, invests the text, illisible or not, with an agency that functions to assure the status of the text as an element in the system. In the *Crosstalk* environment everything is invested with agency, regardless of the origins or value of an element. As a generative system, where things, including people, are produced by one another, it might therefore be considered a people making machine.

6. ACKNOWLEDGMENTS

Arizona State University, School of Arts Media and Engineering, Tempe, Arizona, USA.
Research and Knowledge Exchange Office, Edinburgh College of Art, University of Edinburgh, UK.
Bundanon Trust Artist's Residency Programme, New South Wales, Australia, https://www.bundanon.com.au

7. REFERENCES

[1] Biggs, S. and Leach, J. 2004. *Autopoiesis, Novelty, Meaning and Value*, London, Artwords.
[2] Coniglio, M. 2005. The Importance of Being Interactive. In *New Visions In Performance*, G. Carver and C. Beardon, Eds. Taylor & Francis e-library, pp.5-12.
[3] Hackney, P. 1998. *Making Connections: Total Body Integration through Bartenieff Fundamentals*, New York, Routledge.
[4] Leach, J. 2003. *Creative Land: Place and procreation of the Rai Coast of Papua New Guinea*, Oxford, Berghahn Books
[5] Maturana, H. and Varela, F. 1980. *Autopoiesis and Cognition: the Realization of the Living*, Dordecht, D. Reidel Publishing Co.
[6] Paine, G. 2007. Playing and Hearing Sonic Environments. In *Hearing Places: Sound, Place, Time and Culture*, R. Bandt, M. Duffy and D. MacKinnon, Eds. Newcastle, England, Cambridge Scholars Press, pp.348-368.