Paralinguistic Microphone

Alex McLean
Interdisciplinary Centre for
Scientific Research in Music
University of Leeds, UK
a.mclean@leeds.ac.uk

EunJoo Shin
Incheon Catholic University
Yeonsu-gu
406-849 Incheon, Korea
eunjooshin@gmail.com

Kia C. Ng
Interdisciplinary Centre for
Scientific Research in Music
University of Leeds, UK
k.c.ng@leeds.ac.uk

ABSTRACT

The Human vocal tract is considered for its sonorous qualities in carrying prosodic information, which implicates vision in the perceptual processes of speech. These considerations are put in the context of previous work in NIME, forming background for the introduction of two sound installations; "Microphone", which uses a camera and computer vision to translate mouth shapes to sounds, and "Microphone II", a work-in-progress, which adds physical modelling synthesis as a sound source, and visualisation of mouth movements.

Keywords

face tracking, computer vision, installation, microphone

1. INTRODUCTION

The human voice is a highly adapted carrier of language, but in the digital age its articulation of paralinguistic qualities is often not considered. This is because much of the expressive range and subtlety of the voice lies outside what is commonly notated in the typewritten word. Through interactive sound installation, we have developed an approach which focuses on the sonorous qualities of the voice as carrying paralinguistic communication. In the following we consider our work against a diverse background, including psychology of perception, and build a theoretical basis for wider consideration of related works in New Interfaces for Musical Expression and related fields.

2. SOUND AND SHAPE

In the human vocal tract, the relationship between sound, shape and articulation is clear. This relationship is visceral, and firmly grounded in perception; watching lips move can create a very real experience of hearing sounds which are not there (e.g. McGurk-McDonald effect, McGurk and MacDonald 1976). This has been shown to generalise to watching abstract movements (Rosenblum and Saldaña 1996), demonstrating shared resources for movement and sound in our perceptual faculties.

The relationship between sound and shape is a recurring subject of interest by artists and musicians. One example in digital art includes Takeluma, an alphabet which is based on mouth shape (Cho 2005), with reference to similar properties of Hangul, the native, yet invented alphabet of the Korean language. What these systems have in common is that they notate sound with shape. As already noted, there is strong indication that perception of speech is informed by visual perception of shape, and kinaesthetic perception of its articulation, complementing sonic perception via the cochlear. This relates to the use of vocable words in music, where musicians use words to describe instrumental articulations, connecting their voice to their instrument in a process which often amounts to onomatopoeia (Chambers 1980; McLean and Wiggins 2008).

As Neumark (2010) describes, the voice is both sonorous and signifying, and both embodied and between bodies. These are apparent paradoxes, but our present work brings attention to sonorous qualities as meaningful in their own right, in as much as abstract, orientational metaphor is considered meaningful. Orientational metaphors are those which express concepts in terms of each other, via spatial relationships with the body, forming a coherent system of meaning (Lakoff and Johnson 1980; Gärdenfors 2000). The paradoxical ground between the voice as both embodied and shared between bodies is where our work sits, and for us is a question of resonance, analogous to the two hemispheres of the brain making a whole, through mutual oscillation (Buzsaki 2006).

3. PROSODY IN NEW INTERFACES FOR MUSICAL EXPRESSION

The connection between mouth shape and sound is a recurring theme in the NIME proceedings, for example the Mouthesizer was demonstrated in the first NIME workshop in 2001 (Lyons and Tetsutani 2001), controlling filters based on analysis of mouth width and height via computer vision. Further developments have included the control of physical models (de Silva, Smyth, and Lyons 2004) and whole-face tracking, including mouth shape, in musical parameter mapping (Ng 2004). Voice-controlled synthesis has been pioneered by Janer and Peñalba (2007), using vocable words based on *scat singing* in Jazz as a control mechanism. In somewhat related work, McLean and Wiggins (2008) has explored the use of *vocable words* by describing sounds with onomatopoeic text. In these latter two examples, phonetics are implicated for their role in describing movements which both underlie the production of sound, and inform its perception.

4. MICROPHONE

Microphone is an artwork by *Communications*, a collaboration between two of the present authors. Microphone was first installed at the Unleashed Devices group show at the Watermans gallery London in Autumn 2010, inviting par-

NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.

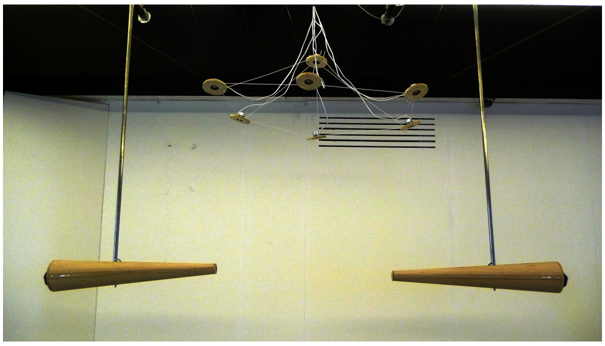

Figure 1: First version of Microphone installed at the Watermans gallery, London. Shows wooden microphone devices hung from ceiling, and multichannel speaker arrangement. An additional speaker was installed inside each of the microphones.

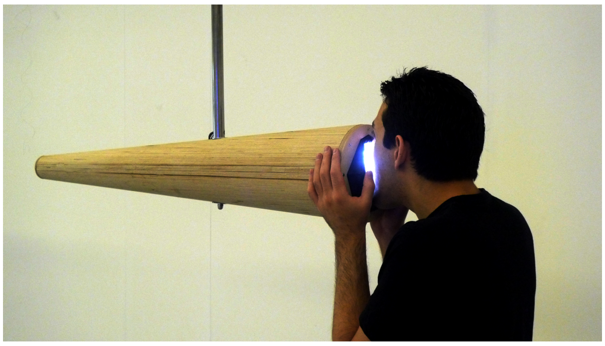

Figure 2: First version of Microphone in use. Mouth area is lit from sides to trivialise contour detection of the mouth cavity.

ticipants to communicate with each other across the gallery, using two large microphones. The title of this piece is somewhat provocative, in that the work does not involve microphones as they are normally understood. That is, the Microphone devices do not capture sound with a conventional electronic transducer, but with a digital camera, where software is trained to produce vowel formants from mouth shapes.

Microphone invites participants to communicate using sound as a medium for paralinguistic gesture, where standard words are not communicable (at least, not in our native Korean and English languages), so that focus is brought on the role of movement in communication. We argue that this evokes a feeling that is visceral, of a vocal organ encoding patterns of movement into sound, and of that sound being perceived in terms of those movements.

Microphone uses computer vision techniques, using standard OpenCV contour detection to identify a polygon representing the shape of the mouth. This is made robust by the design of the Microphone, which surrounds the mouth with light from the side, making the face a controlled stage. From the detected polygon, the parameters of *roundness* and *area* are derived, along with the *aspect ratio* (height/width ratio of the minimum enclosing rectangle), and the *convex hull* area. These measures are used in combination to create a 4D metric space, in which five vowels *a*, *e*, *i*, *o* and *u* are pinpointed in a training phase, using stereotypical pronunciation. Then in use, a participant's mouth shape is located in the space, and formant values are taken as a midpoint between the three closest vowels, weighted by city-block distance. Thus, the range of sounds are mapped continuously. Because these multiple measures are combined, the interaction has subtlety, so that holding the mouth open in a fixed position while moving the tongue produces a corresponding modulation in the sound.

If the mouth is closed, then no sound is produced. Furthermore, contours which are too small, too large or with a centre of gravity too far from the centre are ignored, so that for example nostrils are not falsely identified as mouths.

The sound produced by Microphone is synthesised by SuperCollider (McCartney 2002), and straightforwardly applies a formant filter to a noise sound source, creating a vocal-like sound. The sound is panned across eight channel speakers between the two microphones, including one inside each of the Microphone devices. Due to the space constraints of a group show, the speaker layout meant that the spatial aspects of the work were not discernible by the participants, and only by third parties standing between the microphones.

For more details of Microphone please refer to the video documentation, and the free/open source software and hardware schematics, which are available on-line: `http://comms.me/`.

4.1 Reception

Microphone was installed at the "Unleashed devices" group show at Watermans gallery in London, 2010. Due to the constraints of working in a group show, our exact specification could not be met, in particular the microphones were installed closer together than anticipated, and so the sound spatialisation aspects of the work were not easily perceivable.

We observed and filmed use of the microphone during a related public "dorkbot" event, attended by around 100 people interested in electronic art. We decided not to provide participants with instructions for how to interact with the work, aside from the title "Microphone", allowing people to explore the operation of the devices for themselves. We were also on-hand to answer questions and later gave a presentation on the work. People engaged with the work, often using it in pairs, sometimes experimenting with call-and-response between the two devices. Often a third party stood between the two microphones, where the spatial movement of the generated sounds was most audible. Although we did not conduct structured questionnaires or interviews, feedback was good. The most persistent user was a young girl with a cochlear implant, which raises some interesting hypotheses around accessibility.

In addition we were able to capture images from the cameras, from which individuals could not be identified, and which were destroyed at the end of each day. They were primarily used to monitor the correct working of the microphones, and from these it was clear that the microphones were generally used as anticipated, using the mouth. However there were many cases where people put their ear to the microphone, with the ear hole identified as a mouth shape and triggering sounds.

4.2 Analogue and Digital

There is a question about whether Microphone is digital art or not; it applies digital technology to map between modalities, but the resulting interface is analogue in practical terms. Indeed, entirely analogue means may be used to much the same ends; there are long traditions of using the mouth as musical instrument, either alone (e.g. scat singing or other forms of vocables), or augmented with instruments such as the Jew's harp. The digital foundations

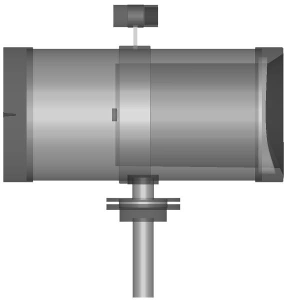

Figure 3: Concept design sketch, showing side view of Microphone II device. Pico projector is mounted to top of microphone, in order to display visualisation.

of this artwork allow expectations to be confounded in interesting ways, but we argue that outwardly, Microphone is an entirely analogue artwork.

4.3 Microphone development

The initial Microphone installation was designed as an extensible framework. The CNC-milled wooden body was designed with extra cavities to encase sensors and other hardware as they came available, and as the conceptual aspects of the work continued. However although the work has been invited to two major international festivals, the cost of intercontinental shipping proved too great for their production budget. This, along with an interest in *device art* supported by the increasingly widespread adoption of consumer-grade 3D printing hardware, motivated the development of a second iteration of the microphone.

5. MICROPHONE II

The second version of Microphone is in the latter stages of iterative design, towards more portable hardware and a richer interaction. We have maintained the same creative limitation of a camera-driven microphone, but are developing the software, refining the configuration of speakers, and redesigning the physical aspects of the interaction towards a smaller, floor-standing device.

Figure 3 shows a concept design sketch of Microphone II. It shows the device in much the same configuration before, in a compact design, but still with adequate focal length between the mouth and the camera. It also shows a pico projector, which will project visualisations of the mouth shapes as shown below.

As the Figure 4 shows, the projections will throw upon either side of a circular display. This will use netting, allowing around 50% of the light through. As a result the participants will see their co-communicator's mouth shapes, but their own shapes will be brighter. As half the light will pass through the netting, the mouth movements will fill the whole space.

5.1 Physical modelling in Microphone

As well as developing the physical and sculptural aspects of the work, we have also developed the software aspects, in terms of synthesis and control. As the schematic shown in Figure 5 shows, the mechanism for controlling a formant filter from mouth shape remains, but the source sound is now generated from a physical model of a drum membrane. Energy is injected into the physical model from an optical flow algorithm, so that motion from across the mouth area is mapped onto the drum surface. This is analogous to making sound by rubbing a face across a drum skin, although to create a more interesting range of timbres, points of movement are translated to short bursts of noise. The effect is more akin to grains being dropped on the surface, with the velocity of the facial movement exaggerated as velocity of the grains.

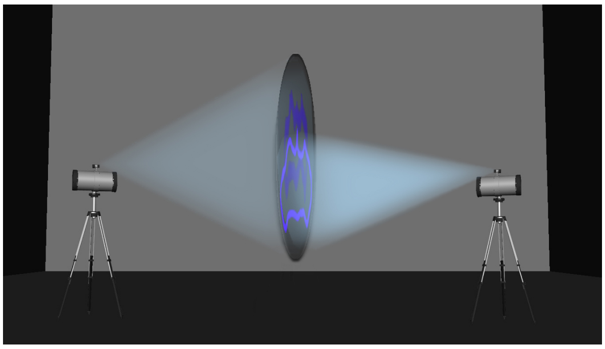

Figure 4: Concept design sketch, showing circular projection screen placed between the two devices.

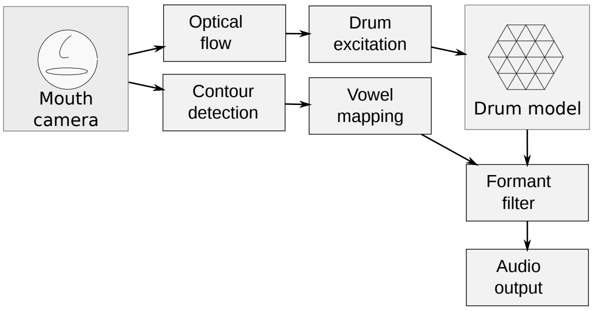

Figure 5: Structure of Microphone II software, showing optical flow as manipulating a simulated drum as a sound source, and mapping of mouth shape to a formant filter.

Sound output can be taken from any point on the membrane, allowing multichannel output. When listening on stereo nearfield monitors or headphones, with channels taken from opposite sides of the membrane, the resulting experience is of being close to the vibrating surface. We anticipate that adding additional channels would accentuate this effect, and we intend to build multichannel speakers within the microphone devices themselves.

6. CONCLUSION

We have outlined our approach to exploring paralinguistic communication through device art, introducing our first installation and describing ongoing work towards a future installation. By furthering the NIME tradition of capturing mouth movements as sound, we have produced work exploring paralinguistic communication. We anticipate that our proposed introduction of physical modelling synthesis driven by facial movements will support rich, yet still nonlexical communication.

7. BIBLIOGRAPHY

Buzsaki, Gyorgy. 2006. *Rhythms of the Brain.* Oxford University Press, USA.

Chambers, Christine K. 1980. "Non-lexical vocables in Scottish traditional music."

Cho, Peter. 2005. "Takeluma: An Exploration of Sound, Meaning, and Writing."

Gärdenfors, Peter. 2000. *Conceptual Spaces: The Geometry of Thought.* The MIT Press.

Janer, J., and A. Peñalba. 2007. "Syllabling on instrument imitation: case study and computational methods." In *Proc. of 3rd Conference on Interdisciplinary Musicology.*

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By.* University of Chicago Press.

Lyons, Michael J., and Nobuji Tetsutani. 2001. "Facing the music: a facial action controlled musical interface." In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 309–310. New York, NY, USA: ACM.

McCartney, James. 2002. "Rethinking the Computer Music Language: SuperCollider." *Computer Music Journal* 26: 61–68.

McGurk, H., and J. W. MacDonald. 1976. "Hearing lips and seeing voices." *Nature* 264.

McLean, Alex, and Geraint Wiggins. 2008. "Vocable Synthesis." In *Proceedings of International Computer Music Conference 2008.*

Neumark, Norie. 2010. "Introduction: The Paradox of Voice." In *Voice: Vocal Aesthetics in Digital Arts and Media*, ed. Norie Neumark, Ross Gibson, and Theo van Leeuwen. MIT Press.

Ng, K. C. 2004. "Music via motion: transdomain mapping of motion and sound for interactive performances." *Proceedings of the IEEE* 92 (apr): 645–655.

Rosenblum, L. D., and H. M. Saldaña. 1996. "An audiovisual test of kinematic primitives for visual speech perception." *Journal of experimental psychology. Human perception and performance* 22 (apr): 318–331.

de Silva, Gamhewage C., Tamara Smyth, and Michael J. Lyons. 2004. "A novel face-tracking mouth controller and its application to interacting with bioacoustic models." In *Proceedings of the 2004 conference on New interfaces for musical expression*, 169–172. Singapore, Singapore: National University of Singapore.