

Beyond Using Voice as Voice

Sama'a Al Hashimi
Lansdown Centre for Electronic Arts, Middlesex University
Hertfordshire, England

Most existing voice-controlled systems are actually speech-controlled. They exploit the linguistic more than the paralinguistic (i.e. non-speech) potential of voice as an input mechanism. The purpose of this paper is to suggest ways in which paralanguage can be usefully exploited in interactive media, and to explore some characteristics of voice that can be employed in controlling interactive applications. The paper discusses existing voice-controlled systems and suggests new forms of artwork aimed at the use of paralinguistic vocalisations for expressive interaction. It presents an interactive game, *Sing Pong*, which is based on paralinguistic input as a means of generating and interacting with the visual output. It also provides an empirical evaluation based on the observation of players' interactions and an informal analysis of audience feedback about *Sing Pong* during an exhibition in London (September, 2004). It is suggested that the use of vocalizations as input might create what could be called *vocal disinhibition* or *vocally-induced catharsis*. For some players vocal control was an opportunity to express themselves more openly and to use their voices in ways that would normally feel forbidden in a public context.

Keywords

Paralanguage, vocal input, voice visualisation, vocal disinhibition, catharsis

Introduction

Over the years, the attempts to emulate human-human interaction in the field of interactive media and to anthropomorphize computers may have blurred researchers' view of the possibility of utilizing the non-human aspects of machines. Recent research into speech recognition has focused on attempts to perceive and interact with computers as anthropomorphic machines. This has led many users to the restricted assumption that only the speech aspect of voice can be used as an input to make computers accessible. It has, perhaps, limited many developers' realization of the potential ability of non-speech voice to be used as an input mechanism. Technology, from such a viewpoint, has been looked at as a mirror which should reflect human-human interaction rather than as a glass through which may be darkly glimpsed the many innovative human-computer interaction methods that are yet to be envisioned. It is time to look at what is behind the mirror as well as at what is in front of it; it is time to surpass making computers barely do what we can do, and give further attention to programming them to do what we can not do.

As humans, we can perceive and interpret vocal paralanguage but we can not easily measure its characteristics. Vocal paralanguage either separately or as an accompaniment of speech includes vocal characteristics (pitch, volume, timbre, etc.), emotive vocalizations (laughter, screaming, crying, etc.), and vocal segregates (fillers, pauses, exclamations, etc.). We are more accurate than computers in perceiving and interpreting each others' paralinguistic vocalizations, but

computers can complement our perceptions by their accuracy in capturing and processing voice signals and measuring their characteristics. During our human-human conversations, we usually respond to others' voices by voice. A computer, however, can respond multimodally to voice by visuals, movements, images, odors, or any kind of output. It can fundamentally 'transmediate' voice or even transform one medium into another.

Paralinguistic Vocal Control of Interactive Media

Although vocal paralanguage is neglected in the field of interactive media, its use as an input mechanism may enrich the vocabulary of interaction and widen the range of human input to computer-based systems. It may create a real-time and immediate causal relationship between the vocal input and the visual output and may thus facilitate continuity and direct engagement. It may direct an interactive system to performative extremes and allow it to be experienced as a holistic engagement, with the human body as a rich source of expressive input and performance.

Lately, there has been significant research interest and a considerable amount of work on audio-visualisation. Several developers have developed applications that translate musical cues into graphical feedback. They have employed various strategies to establish novel mapping techniques between visual properties and sound characteristics. Quite a few, however, have focused on voice-visualisation to establish meaningful mappings between vocal inputs and visual outputs. Among these developers are Levin and Lieberman who explored a variety of mappings in *Hidden Worlds* (2002). *Hidden Worlds* is an augmented multi-user installation in which six people sit around a table and use their voices to control visual data displayed through data glasses. Voices are transformed into "worm-like figures" displayed on the table. The position of each figure is determined by the spatial position of each user. The duration of the figure's appearance is mapped to the duration of the vocalisation, while the figure's diameter is mapped to volume. Pitch determines the figure's "flocking behaviour".

Another voice-visual installation which Levin and Lieberman developed, and by which our game *Sing Pong* was inspired, is *Messa Di Voce* (Italian for "placing the voice"). This is an installation that explores speech visualisation using an original technique that makes speech appear visually as if emerging from the speaker's mouth (Levin and Lieberman, 2003). In one of *Messa Di Voce's* sections, "Pitch Paint", performers use the pitch of their voices in order to paint on a projected screen. In another section, "Jaap's Solo", a performer uses his voice to emit bubbles which appear to emerge from his mouth on the screen. A computer vision technique is used to enable him to use his shadow in order to interact with the bubbles and move them.

Another voice-controlled application is the collaborative multiplayer game *Organum*. Players use their voices to control an avatar that moves through a three-dimensional model of the vocal tract. They moan, whistle, and hum to move the avatar along the x (left-right), y (up-down) and z (forward-backward) axes on the screen (Niemeyer et al, 2005). Each player controls the movement of the avatar along a different axis in order to avoid hitting body organs.

Like *Organum*, *SpitSplat* maps player's vocal characteristics to movement along axes. *SpitSplat* is a voice-controlled game which I developed with two other postgraduates in the Lansdown

Centre for Electronic Arts (LCEA). It was my first attempt to investigate the employment of vocal paralinguage as control interface. Voice is used to 'splat' the appropriate colour onto several moving targets (Figure 1). Players aim the “colour splat” at moving targets by altering certain qualities of their vocal expression including pitch, loudness and duration. Pitch and loudness correspond to the x-y coordinates. Thus, a high-pitched and loud vocalisation directs the “colour splat” towards the top right corner of the stage while a low-pitched and soft vocalisation directs the “colour splat” towards the bottom left corner.

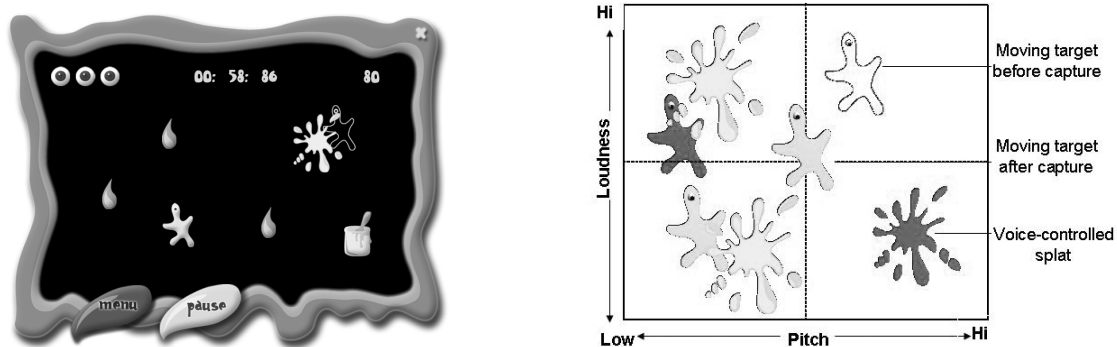


Figure 1: A screen shot and an illustration of *SpitSplat*; a game controlled by the pitch and volume of a player’s voice

More unusually, players have to demonstrate a high degree of control over their vocal cords, moving from high to low pitches and loud to soft voices. This technique may pave the way for future functional applications which could be used for therapeutic purposes by asthmatic and vocally-disabled users, or even as a training tool by vocalists and singers. Although the works discussed in this paper lie within the field of entertainment and/or the arts, the functionality of paralinguistic voice may prove to be of broader relevance.

The following section discusses the rationale for exploring paralinguistic vocal control of interactive media in more depth. It also presents the voice-controlled game, *Sing Pong*, and provides a brief empirical evaluation of it.

Sing Pong; A voice-controlled Pong

Sing Pong is a computer projected game based on using voice characteristics to control the traditional video game “Pong”. Initially created by Ralph H. Baer in 1966, “Pong” consists of two paddles and a ball. A Player controls a paddle by moving it using a keyboard or a joystick in order to hit the ball toward the other side of the playing field where the other player hits it back. *Sing Pong*, which I developed with another postgraduate at the LCEA, enables players to control the paddles using their voices. The game is displayed on a projection screen, and the two players stand in front of the projector (Figure 2). The volume of the player’s voice determines the height of the paddle. The position of the head’s shadow on the projection is tracked using a web-cam to determine the position of the vocal paddle (Figure 3). Thus, the paddle appears to emerge from the player’s mouth.

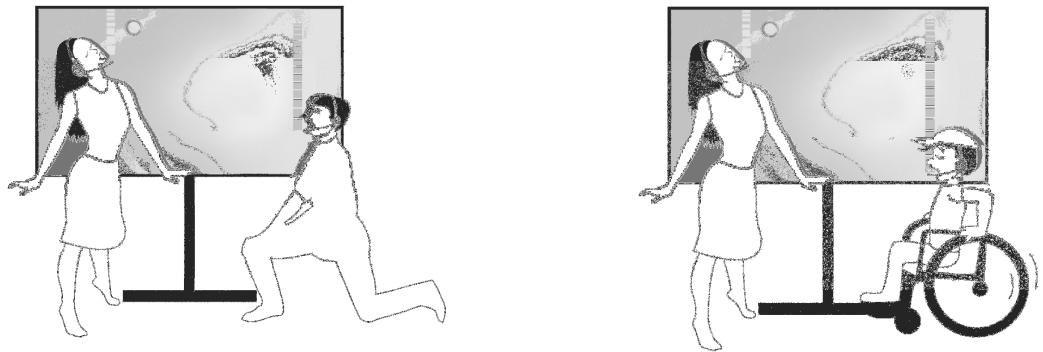


Figure 2: An illustration of the voice-controlled game *Sing Pong*

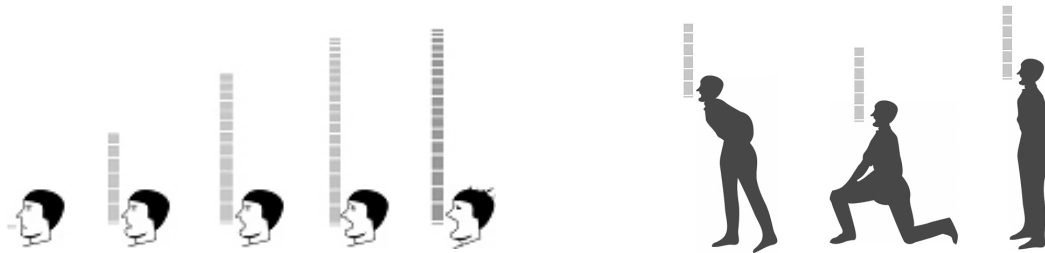


Figure 3: The height of the paddle is mapped to the volume of the player's voice and the position of the paddle is mapped to the position of the player's shadow

In this game all devices are hidden and the interface extends to the space in front of the screen. It embraces the players as well as the physical space allocated for movement and tracking. The button that initialises the play is a round mat positioned on the floor, on which players stand in order for their shadows to appear in the middle of the projected screen. The shadows' appearance in the middle of the screen initialises the game. Thus players are fully aware of their real as well as their virtual presence. They are prompted to move backward and forward within the physical space, and to stand and kneel in order to control the position of the vocal paddle in the virtual space. Thus, the game induces physical motion, and allows players to involve their bodies as well as their vocal skills while playing.

From the spectators' point of view, the playing space and the players are an integral part of the game. Their shadows are images, their voices are sound effects, their movements are animations, and their interactions are physical events. Moreover, any movement within the physical space allocated for playing and any interference from the audience may alter the play, and include them in the interface. During its exhibition in London, some spectators distracted the play by putting their fingers in front of the projector. Many spectators enjoyed walking in front of the projector and interfering with the players' shadows either to make them lose or to help them hit the ball. One of the spectators was a kid who enjoyed projecting his shadow to interfere while his dad and sister played. In *Sing Pong*, therefore, audience members are not passive spectators, and the action on the stage is not all there is. Spectators may take on an active role and their actions around the stage can also affect the performance. In addition, any alteration in lighting and any change in the position or size of objects or subjects within the interaction space can significantly

alter the game. Thus, the actual interface is no longer determined by what the player sees within the screen but by what the camera sees within its field of view.

The game requires a dark space, and this is probably one of the factors which encourage people to play it in front of others, and to project their voices as visuals on the screen. In such a voice-visual installation, the voice is projected towards the screen as visuals rather than towards the audience as pure voice. The microphone which is normally used to amplify and exhibit the voice is now used to transform the voice into visuals, and thus conceal it aurally by displaying it visually. This shifts the players' focus from their voices as voices to their voices as visuals. They don't think of their voices as an audible element of which they are the source as much as they think of it as a disembodied visible element of which they are the controllers. Unlike typical singing which requires the singer to face the audience and watch their reactions, *Sing Pong* players' vocal engagement with the visuals diverts their attention and anxiety from thinking about what spectators think of them towards thinking of play. They do not face the audience as singers but each other as contenders. This is probably what eliminates shy player's shyness. On the other hand, outgoing players consider such a game an opportunity to release their energy, express their emotions, and impress or grab others' attention. All of this leads me to infer that paralinguistic vocal control of interactive media may play a significant role in disinhibiting players and in inducing their cathartic experiences.

The term "catharsis" was originally used by Aristotle to refer to the emotional expression, purgation, and cleansing which an audience experiences after watching a tragedy. The term is occasionally used in contemporary psychological discourse to describe the purgation and emotional release of personal frustrations and tensions through expression, revelation, and discharge of "the light fears and angers of social embarrassment" (Heron, 1977). *Sing Pong*, in this context, also provides an outlet for emotions and encourages social and *vocal disinhibition* through expression and discharge of embarrassment. The fact that it is a two-player experience encourages players to make voices in public more than it would if it was for one player. This is because the experience is shared and explored by both players while their voices overlap; masking and at the same time augmenting each other. Moreover, the act of raising ones' voice which is usually inhibited and considered of negative consequences is in *Sing Pong* encouraged and rewarded by winning the game. The *vocal disinhibition* and the *vocally-induced catharsis*, which players seem to experience, is therefore a result of the negative becoming positive and the disapproved becoming approved in the game's context. In *Sing Pong*, disinhibition is not initially experienced by the player during the game but is also experienced beforehand when the player-to-be is a spectator. Being a spectator of a voice-controlled installation does not only involve watching players play but also involves watching other spectators' reactions. These reactions determine whether the spectator chooses to become a player or not, and the disinhibition process starts when the spectator perceives other spectators' positive reactions. As a result, the game involves a hidden form of social interaction and of social learning (theory) through which people learn by watching others who are rewarded. It also involves some forms of technology-related psychotherapy because it makes certain actions and expressions more tolerable socially, and it increases the viewer's willingness to try new interactive experiences.

Conclusions

Vocal paralinguistic control introduces a new style of interaction between the human and the machine. It enables users to interact with a machine in an expressive manner by executing explicit vocal expressions. In *Sing Pong*, the interaction sometimes led to unexpected dramatic excitement and creative improvisations. Some people chose to whistle rather than generate “ahhh, ooh” voices. Others started jumping in order to raise the height of the paddle. Many players really enjoyed shouting and attracting other visitors' attention. Some players placed two fingers right in front of the projector in an attempt to control a paddle with each finger. Such improvisations and vocal experimentations may never be possible in a speech recognition-based game that restricts players to limited languages and particular accents. Paralinguistic vocal control proved to be a transcultural mechanism aimed at encouraging the development of performance-centered applications and achieving the highest level of audience engagement and participants' immersion, regardless of their languages and accents. Paralinguistic vocal control also stimulates going beyond or perhaps complementing the verbal form of communication to convey inner thoughts and emotions and translate them into visible or audible representations. Voice as an input mechanism can also enable dual task performance; it may allow the player to sing while dancing, or to shout while running and jumping. It also frees the user's hands and movements and encourages their use as improvisational controllers of new input mechanisms within the real space rather than as traditional controllers of keyboards and mice within the virtual space. It thus stimulates other expressive and multimodal means of interaction, and extends the level and scope of interaction vocally, modally, and spatially. It may potentially complement and facilitate other effective forms of input within a single application. Furthermore, this technique potentially introduces a dramatic approach to multimedia applications. Games which employ vocal input do not only target players as first-level users or performers but also engage spectators as second-level users or as an audience. In *Sing Pong*, the players are not just being entertained but are entertainers as well.

It seemed that players' pleasure in interacting with *Sing Pong* could mainly be due to the invisibility of the computer which is usually visible, and to the visibility of the voice which is usually invisible. For that reason, my future work will involve an exploration of a variety of novel voice-physical mappings which will extend beyond the graphical output to include physical feedback such as changes in the position, size, temperature, brightness (and other aspects of colour), speed, direction, and height of real objects. I mainly aim to explore the possibility of physical control of inanimate objects with minimal vocal input, or what I would refer to as *Vocal Telekinesis*. I am currently developing a voice-physical version of the classic ‘Snake’ game. The game consists of an installation table on which the user places a coin on the top surface. A virtual snake is projected on the table. The user's voice characteristics are programmed to move the coin around in order to avoid collision with the snake. The position of the user around the table determines the direction of the coin's path, and this encourages physical activity as well as vocal activity during the game.

I think that it is now time to learn new input mechanisms rather than teach the computer our old input mechanisms; to focus on the user as an input source rather than on the computer as an output source.

References

Heron, J. (1977); Catharsis in Human Development; Human Potential Research Project; University of Surrey

Levin, G. and Z. Lieberman (2003); Messa Di Voce; <http://tmema.org/messa/messa.html> (May, 2004)

Levin, G. and Z. Lieberman (2004); In-Situ Speech Visualization in Real-Time Interactive Installation and Performance; Proc. 3rd International Symposium on Non-Photorealistic Animation and Rendering, June 7-9 2004, Annecy, France.

Niemeyer, G., Perkel, D., and R. Shaw (2005); Organum the Game
<http://www.sims.berkeley.edu/~dperkel/organum/index.html> (April, 2005)