Paralinguistic Vocal Control of Interactive Media

How Untapped Elements of Voice Might Enhance the Role of Non-Speech Voice Input in the User's Experience of Multimedia

A thesis submitted to Middlesex University in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Sama'a Al Hashimi

Lansdown Centre for Electronic Arts
Middlesex University

May 2007

Abstract:

Much interactive media development, especially commercial development, implies the dominance of the visual modality, with sound as a limited supporting channel. The development of multimedia technologies such as augmented reality and virtual reality has further revealed a distinct partiality to visual media. Sound, however, and particularly voice, have many aspects which have yet to be adequately investigated. Exploration of these aspects may show that sound can, in some respects, be superior to graphics in creating immersive and expressive interactive experiences. With this in mind, this thesis investigates the use of non-speech voice characteristics as a complementary input mechanism in controlling multimedia applications. It presents a number of projects that employ the paralinguistic elements of voice as input to interactive media including both screen-based and physical systems. These projects are used as a means of exploring the factors that seem likely to affect users' preferences and interaction patterns during non-speech voice control. This exploration forms the basis for an examination of potential roles for paralinguistic voice input. The research includes the conceptual and practical development of the projects and a set of evaluative studies. The work submitted for Ph.D. comprises practical projects (50 percent) and a written dissertation (50 percent).

The thesis aims to advance understanding of how voice can be used both on its own and in combination with other input mechanisms in controlling multimedia applications. It offers a step forward in the attempts to integrate the paralinguistic components of voice as a complementary input mode to speech input applications in order to create a synergistic combination that might let the strengths of each mode overcome the weaknesses of the other.

Acknowledgements:

I am extremely grateful to my director of studies and mentor, Gordon Davies, for his unstinting guidance and support throughout every stage of my Ph.D. research and experience. Without his contribution to the development of the source code none of my projects could have been completed, and I appreciate his help in solving all the technical problems that I have encountered. My gratitude to Gordon is as everlasting as his influence on my intellect. I am exceedingly grateful to Stephen Boyd Davis and Magnus Moar for their lavish supervision of my Ph.D. research. Their comments and suggestions were very helpful in refining this thesis and my practical projects.

I am indebted to Nic Sandiland for collaborating with me and teaching me all the necessary technical skills to bring *Expressmas Tree* to fruition. I am very thankful to Gordon Davies and Sumetha Nagalingam for their collaboration in the development of *Sing Pong*. I am also thankful to Gordon, Sumetha, and Athanasios Anthopoulos for their collaboration in the development of *SpitSplat*. I am thankful to Middlesex University for its continuous financial support and to Peter Williams, John Cox, Andrew Pomphrey and Dominique Rivoal for their constant technical support.

I am indebted to The University of Bahrain for sponsoring my studies and granting me the opportunity to pursue my postgraduate degree. I am very thankful to Dr. Fouad Shehab, President of Prince Sultan Al Saud Hearing and Speech Development Centre and Afrah Al Fardan, Instructor at the Centre for allowing their students to participate in my experiment.

I am infinitely grateful to my father for molding me into the person that I am today, for instilling in me an unquenchable thirst for knowledge, and for guiding me to the highest intellectual pursuits. I appreciate my mother's wholehearted support which encouraged me throughout my academic pursuits. I am embarrassed to write that I am grateful to her but honored to write that I love her

instead; I am confident that my love to her is immeasurably great. My gratefulness to her, however, remains humble no matter how great.

I am grateful to my parents-in-law. They are the reason I have a partner who shows such infinite compassion, patience and support.

I will forever be grateful to my partner, Qais, who waited for me to come back so that I could go forward.

Dear God,
Thanks for everyone I thanked.

Key Words:

Vocal input, non-speech sound, paralanguage, vocal telekinesis, voice-visualization, voice-physical

Declaration:

I declare that all the work in this thesis is my own and all sources have been acknowledged.

Sama'a Al Hashimi
May 18, 2007

Contents

1. Introduction

1. Introduction

> *"As one peels back the onion skin abstractions of media technology, voice as sound is revealed as the forgotten memory of a culture steeped in vision."* [Wendt, 2000]

The development of multimedia technologies such as augmented reality and virtual reality has revealed a distinct partiality to visual media. However, sound and particularly voice have many aspects that have yet to be adequately investigated:

"Expressions like 'I see what you mean', 'I'll see to it', 'Point of view', 'Seeing is believing', demonstrate how we equate sight with understanding, doing, and reality, fetishising the graphic. And yet the uncritical acceptance of the idea that ours is a visual culture has seduced us into believing that the voice has been somehow superseded, even though new technologies, as we'll see, have enhanced its importance rather than diminished it." [Karpf, 2006:209]

In this thesis my principal argument is that there are elements of voice which have not yet been tapped in the field of interactive media, possibly due to the recent preoccupation with the verbal channel of voice. These elements if appropriately exploited may enhance the user's experience of multimedia and offer certain benefits especially when used as a complementary input to speech recognition and other input mechanisms.

Vocal utterances can be detected by microphones, and computer software can be used to extract words – and ultimately some aspects of the meaning of those words – from the voice signal. This is the function of speech recognition. However, there are also paralinguistic aspects to vocal utterances – characteristics which may hold information, but without requiring access to the verbal content.

Most researchers and developers in the field of vocal input are currently focusing on speech recognition systems as an alternative input mechanism to keyboards and mice. Although research into speech recognition technologies has made remarkable progress, these systems are still not flawlessly reliable [Huerta et al.,

2004] [Cox and Walton, 2004]. They are usually affected by background noise and by the speaker's accent. Furthermore, the prevalence of research into speech recognition has led to the 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. Researchers have neglected paralinguistic components of vocalization.

Compared to the difficult process of detecting linguistic cues in verbal speech recognition, which often requires prior training of the software by the user, non-speech features of voice may be near-instantaneously detected through voice signal spectral analysis (e.g., [Quast, 2002]). Speech recognition systems may suffer from considerable latency and the user must wait for the recognition results after uttering a word [Igarashi and Hughes, 2001:155]. On the other hand, the use of paralinguistic vocal control in interactive media can display a real-time and near-immediate causal relationship between the acoustic input and the visual or physical output, and may therefore facilitate continuity and direct engagement. This was also noted by Igarashi and Hughes who believe that when the user produces a continuous vocal sound, s/he *"can continuously observe the immediate feedback during the interaction."* [Igarashi and Hughes, 2001: 155].

At the same time, recent research into speech recognition has focused on attempts to perceive and interact with computers as anthropomorphic machines. These attempts to emulate human-human interaction in the field of interactive media and to arguably anthropomorphize computers may have overshadowed the possibility of utilizing the non-human aspects of machines. It might be more challenging to pass beyond making computers barely do what humans can already do, and give further attention to programming them to do what humans cannot do.

As humans, we can perceive and interpret vocal paralanguage but we cannot easily measure its characteristics explicitly. 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 currently more accurate than computers in 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 and interactions, we often respond to others' voices by voice and gesture. A computer, however, can respond multimodally to voice by visuals, movements, images, odors, or any kind of output. It may cause *"uniquely ephemeral dynamic media to blossom from the expressive 'voice' of a human user"* [Levin, 1999].

With this in mind, this thesis accompanies a set of artistic projects undertaken to investigate how to make creative use of the paralinguistic aspects of voice in controlling interactive media. These aspects can include vocal utterances which are non-verbal such as laughing, non-verbal aspects of verbal utterances such as the intonation given to words, and borderline cases such as vocal fillers like "er..". The emphasis of the present thesis and projects is on the non-verbal utterance, but it may be useful initially also to discuss paralinguistic aspects of the verbal utterance. In juxtaposition with my own projects, the theoretical framework of the thesis provides a literature review of existing audio-visual and voice-visual work and explores the various voice-visual mappings and techniques employed. The thesis also investigates the integration of the paralinguistic components of voice as a *complementary* input mode to speech input applications with the aim of creating a synergistic combination that might let the strengths of each mode overcome the weaknesses of the other. The thesis addresses the following research questions:

1. What are the aspects of voice that have not yet been adequately explored and exploited in the field of interactive media?
2. What benefits and what limitations are there in using paralinguistic rather than or along with linguistic input? And what is the future potential of paralinguistic voice input?

3. What are the factors that may affect users' preferences and interaction patterns during non-speech voice control, and by which the developer's choice of non-speech input to a voice-controlled system should be determined?

It is important to emphasize that in order to find answers to these enquiries, I have principally adopted an artistic, explorative and playful approach. However, in order to reflect upon this approach and undertake an evaluation of its outcomes, parts of the work are underpinned by experimental enquiry.

The general outline and structure of the thesis is as follows:

The first chapter presents a brief introduction to this thesis. It discusses the motives underlying the research. It presents the objectives that are to be achieved and the research questions that are to be examined, and outlines the theoretical framework of the thesis as a whole.

The second chapter covers the relevant background information and serves as a basis for subsequent discussion. It contains an overview of the diverse aspects of voice including the physiological, the artistic-performative, the paralinguistic, the communicative, the cultural, and the technological. Not all of these are dealt with in equal detail. In order to explore the variables of vocalization as a source of input to interactive media, it is necessary to explore the range of non-speech sounds that can be produced by the vocal tract, especially those that can be deliberately controlled by the user. It is also important to highlight the physiological and psychological factors that may underlie or affect the delivery of these sounds. The section on the physiological aspects, therefore, looks at the underlying physiology of voice including voice production, aural voice processing and neurophysiological voice processing. The section on the artistic-performative aspects describes vocal forms of art that involve non-speech vocalizations such as singing, chanting, sound poetry, and ventriloquism. The objective of this section is to consider paralinguistic vocal control of interactive media in the light of art and performance, and to facilitate later discussion about the expressive role of paralinguistic voice input in transforming users into

performers. The section on the paralinguistic aspects deals with the various constituents of vocal paralanguage including voice characteristics, emotive vocalizations, vocal segregates, and what I will argue can be called 'holistic' interjections. The same section also investigates the controversy about the classification of vocalizations. It also presents a table that compiles the paralinguistic information conveyed by various voice characteristics based on studies and experiments undertaken by a number of researchers. The section on the communicative aspects addresses vocal forms of communication such as vocal mimesis, onomatopoeia, and parent-infant communication. It shows that beyond the use of vocal imitation and infant-directed speech for communication purposes, these vocal forms of communication may also be of great significance to studies of melody recognition and prosody-dependent speech recognition. The section on the cultural aspects outlines the remarkable uses of vocal paralanguage that exist across various cultures for different communicative purposes. In this section I argue that there are vocalizations that are specific to certain cultures but which, if exploited in the field of interactive media, can be developed into newly *cross*-cultural means of human-computer interaction. The section on the technological aspects gives a brief account of various voice-related technological advancements such as voice recording, voice analysis, voice recognition, , speech recognition, recognition of non-speech aspects of voice and electronic manipulation of voice. It sets the wider context for some of the relevant technological aspects that are involved in the development of the projects discussed in later chapters.

The third chapter investigates paralinguistic vocal control of interactive media. First, however, it notes the small amount of research which has been done in relation to paralinguistic vocal *responses* to interactive media, and argues the potential role of exploring the vocal responses evoked by human-computer interaction in improving studies on vocal input. The rest of it is divided into a section about voice-visual media and another about *voice-physical* media. The section about voice-visualization first introduces some significant audio-visual works by others working in the field and analyzes the various audio-visual mappings employed. It next explores some remarkable voice-visual works and

analyzes the novel voice-visual mappings implemented. It then describes and evaluates my own voice-visual projects; *SpitSplat* and *Sing Pong*. The section about *vocal telekinesis* first introduces significant audio-physical applications. It then explores the few existing *voice-physical* works, and finally describes and evaluates my *voice-physical* projects; *sssSnake, Blowtter, and Expressmas Tree.*

Before a discussion of the potential benefits of paralinguistic voice input in the fifth chapter, the fourth chapter reflects some factors that have appeared as possible limitations to, or at least determinants of, the ability of some users to generate non-speech voice. The chapter presents the results of my experimental approach to the examination of the factors that may affect users' preferences and interaction patterns during non-speech voice control, and by which the developer's choice of types of non-speech input should be determined.

The fifth chapter is a discussion of the potential roles of paralinguistic voice input in inducing cathartic experiences, in singing and expressive communication, in voice therapy and speech therapy, in augmenting the awareness of voice characteristics in the hearing-impaired, and in transforming users into performers.

In the concluding chapter, the objectives, arguments and findings are briefly restated and their implications are assessed in an attempt to prompt the development of forms of expressive interactive media in which the ultimate objective is literally to incorporate the user into the installation.

2. Key Aspects of Voice

2. Key Aspects of Voice

In order to exploit the paralinguistic successfully, I found that I needed to understand many different aspects of vocalization. This chapter therefore gives an indication of each of these issues in turn, as an essential foundation for what follows. In subsequent chapters, many of the concepts presented in this chapter are revisited and discussed in the context of my own practical work.

2.1 Physiological Aspects

This section provides a review of the physiological basis of vocalization including voice production, aural voice processing, and neurophysiological voice processing. Some of the processes that voice production may involve, such as voiced as opposed to unvoiced sound, are discussed in the context of my own projects. Some aspects that underlie aural voice processing have helped me acquire the necessary background information to conduct a study on deaf children as will be discussed in the fifth chapter.

a. Voice Production

The underlying physiology of voice is based on the vibration of the vocal folds in the larynx or voice box. The larynx is at the top of the windpipe or trachea (Figure 1). The vocal folds or vocal cords are two bands of muscle which vibrate when a person generates voice. The air that goes in and out of the lungs during breathing causes the vocal folds to move away from each other, and the air that goes out of the lungs during phonation causes the vocal folds to come together. The length and thickness of the vocal folds can be changed while vocalizing. Lengthening and relaxing the vocal folds while vocalizing generates a low-pitched voice, and tightening and shortening the vocal folds generates a high-pitched voice. If the vocal folds vibrate 250 times per second, they are said to vibrate at 250 Hertz (Hz).

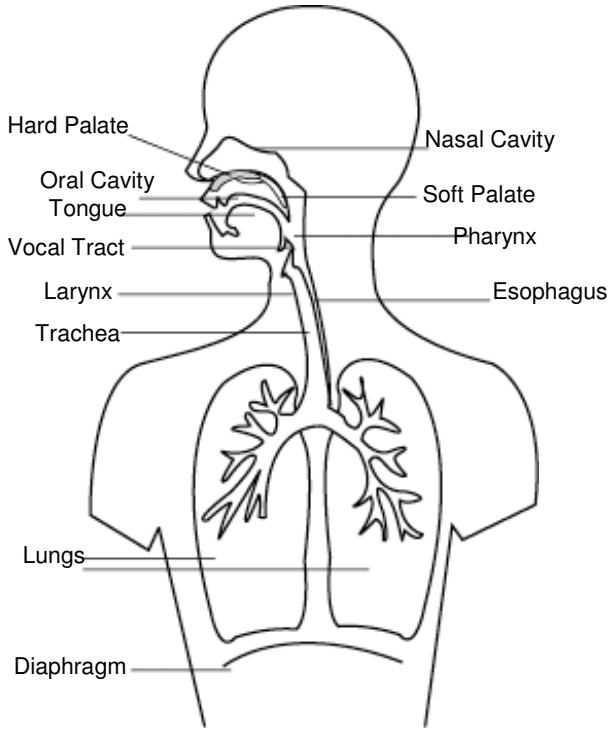


Figure 1: Anatomy of the voice [Based on Hooshmand, 2000].

Gender difference in the production of voice is due to the difference in size between the vocal folds of adult males and females. This difference occurs during puberty, when the length of the vocal folds increases by 3-5mm in females and by 5-10mm in males [Scherer, 2000]. The thicker and longer vocal folds of males vibrate at a lower frequency. The distinction is further accentuated by differences in larynx size and the space between the vocal folds. The average normal speaking pitch for males is 150 Hz while it is around 200 Hz for females [Scherer, 2000].

The process of generating voice involves breathing, sound creation, articulation, control, and delivery. Voice generation starts from breath control for which the diaphragm and lungs are responsible. The air stream then passes through the trachea into the larynx, and then into the glottis and epiglottis and into the mouth. The glottis is the area between the vocal folds. The shape of the mouth determines some of the features of the sound produced. The process of modifying the airflow by reshaping and moving the mouth along with other body

organs (i.e. articulators) such as the lips, tongue, larynx, teeth, and nose to form sounds is called articulation.

The three methods of expressing voice are voicing, place, and manner [Hale and Kissock, 2002] [Hall, 2002]. Voicing is determined by the status of the vocal folds. If the vocal folds vibrate, for example when uttering the letter 'z', the sound produced is voiced. If the sound produced does not involve vibration of the vocal folds, such as when uttering the letter 's', the sound is voiceless.

The place of articulation is determined by the articulators that meet and obstruct the vocal tract in order to produce a specific sound. When the upper and lower lips meet during articulation, the sound produced is bilabial. When the teeth and the lower lip meet, such as when uttering the consonant 'v', the sound is labiodental. When the tip of the tongue meets the front upper teeth, the sound is dental. When the tip of the tongue meets the gum ridge behind the front upper teeth, the sound is alveolar, and when the tongue meets the area behind the ridge, the sound is palatoalveolar or postalveolar. When the tongue meets the hard palate, the sound is palatal and when it meets the soft palate (also known as velum), the sound is velar.

Manner is determined by the kind of obstruction formed when the articulators meet or contact each other. When there is no air flow as a result of complete closure of the oral and nasal cavities, the sound produced when the cavities reopen is a plosive or an oral stop. When only the oral cavity is closed, the air flows through the nasal cavity and the manner by which the sound is produced is termed nasal stop. When friction is produced as a result of a partial obstruction of the airflow, the sound produced is a fricative. Fricatives can either be sibilant or non-sibilant. When the sound produced is that of a high-pitched hissing noise like 's', the sound is a sibilant. When the sound produced is a result of a combination of a plosive followed by a fricative, such as when uttering 'ch', the sound is termed affricate. When the airflow is released through the sides of the tongue, the sound produced is lateral. When the vocal tract is narrowed and the tongue approaches the roof of the mouth but without creating an obstruction that

causes friction, the sound produced is an approximant. When approximants do not involve a significant degree of obstruction just like vowels, they may be classified as vowels. Any other sound that involves some degree of obstruction is termed a consonant.

It is believed that the voice *"can only reproduce what one can hear"* and that changing the ability to hear may change one's voice [Tomatis 1991 and 1996 quoted in Karpf, 2006: 32]. The next section deals with how the ear processes the voice.

b. Aural Voice Processing

Since hearing is argued to have a significant influence on the development of vocalizations [Clement et al., 1994], it is necessary to provide a brief explanation about how the ear processes the voice. Some researchers, for instance, claim to have noticed a difference in the development of vocalizations between deaf and normally hearing children during their first year [Clement et al., 1994].

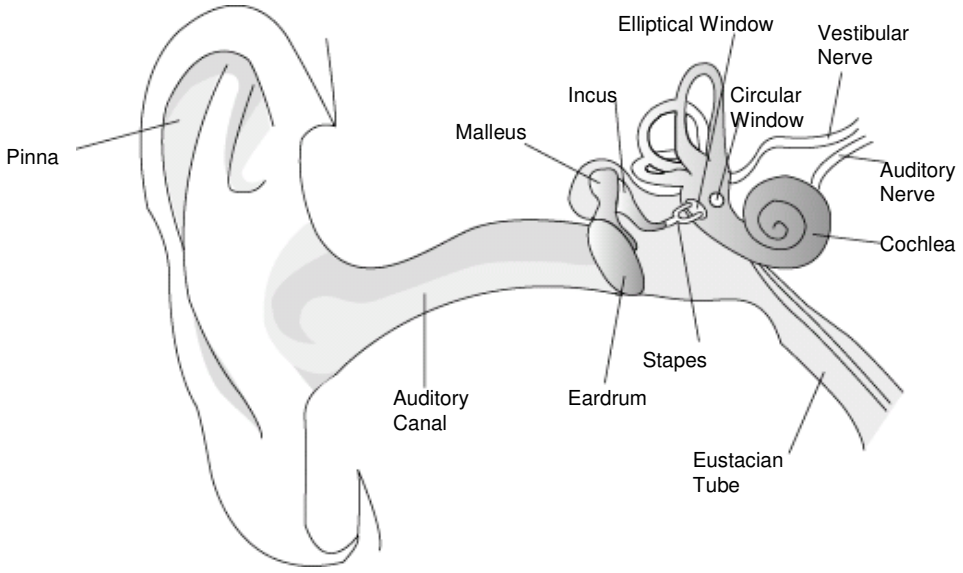


Figure 2: Anatomy of the ear [Based on Gray et al., 1973 and Hooshmand, 2000].

Normal infants are thought to start vocalizing through babbling which starts before the age of eleven months [Clement et al., 1996], while no babbling was found before that age in deaf infants [Clement et al., 1996]. It has been reported, however, that deaf infants produce repetitive hand movements as a way of babbling with their hands [Clement et al., 1994].

The three main parts of the ear are the outer ear, the middle ear, and the inner ear. The outer ear or pinna is the visible part of the ear (Figure 2). It collects sound and sends it through the ear canal to the middle ear. When the sound waves reach the tympanic membrane (also known as the eardrum), the membrane, which is located between the outer and middle ear, vibrates. The ear drum is attached to the ossicles. These are the three ear bones (malleus, incus, and stapes) which transmit the sound waves to the oval window and then to the cochlea in the inner ear. The thin fluid in the inner ear transmits the pressure changes of the voice through the cochlea to the hair cells. Each hair cell is sensitive to a certain frequency at which it may resonate. When a hair cell senses the pressure changes of the voice, it sends them as nerve impulses through the auditory nerve to the cerebral cortex in the brain. The next section looks at how the brain is thought to process the human voice.

b. Neurophysiological voice processing

Neurophysiological voice processing, the ability of the brain to perceive, analyze, and produce vocal information, plays an important role in our daily social interactions. We can still extract social and emotional information regardless of whether the voice we listen to does or does not contain speech information. It is thought that different elements of vocal information are processed in different areas in the brain.

The temporal lobe is the part of the brain that is mainly responsible for receiving auditory information from the ears and then communicating with Wernicke's area in the parietal lobe (Figure 3). Results from recent brain imaging research

on neural processing of voice and speech parameters generally suggest that tonal and musical qualities and affective voice processing are usually associated with the brain's right hemisphere [Loewy, 2004][Karpf, 2006:58]. On the other hand, speech processing is often associated with the brain's left hemisphere [Lattner et al., 2005]. The left hemisphere is known to be responsible for phonetic analysis and the planum temporale in it has been observed in functional magnetic resonance imaging (fMRI) to be activated in response to consonant-vowel syllables but not to steady vowels [Jäncke et al., 2002].

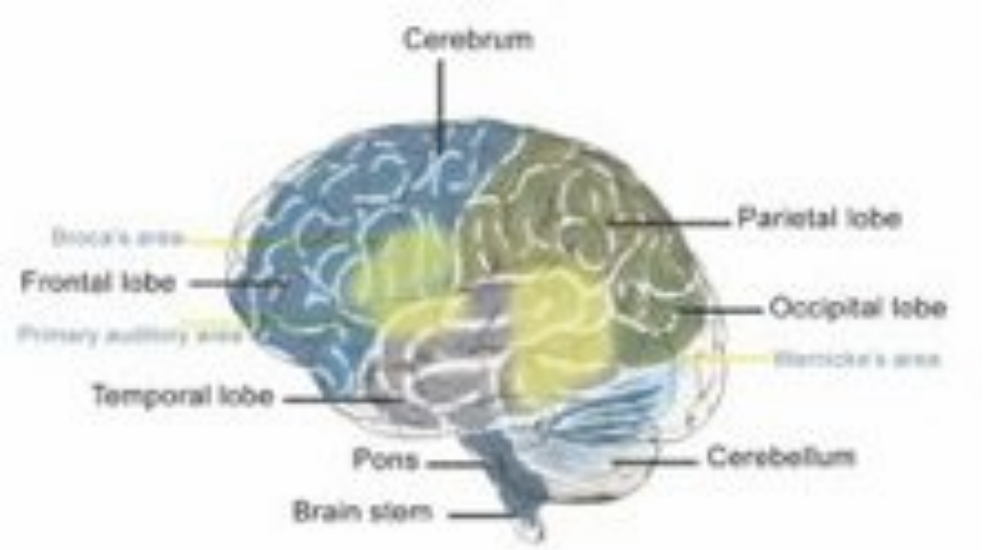


Figure 3: The main parts of the brain and the areas mainly involved in neurophysiological voice processing [Based on Gray, 1973 and Hooshmand, 2000].

The left hemisphere "*produces, processes, and stores individual speech sounds*" while the right hemisphere "*produces and processes the intonation and melody patterns of speech*" [Givens, 1998b]. The left hemisphere's inability to articulate speech because of brain damage or lesions is aphasia. There are two types of aphasia; sensory aphasia and motor aphasia. Sensory aphasia is mainly caused by damage in Wernicke's area, which is at the back of the parietal lobe and is responsible for understanding language. Patients with sensory aphasia may not understand speech but may still produce speech sounds. Motor aphasia is caused by damage in Broca's area which is on the left side of the frontal lobe and is responsible for generating speech. Patients with motor aphasia may understand language but may not be able to produce meaningful words because of the damage in Broca's area which controls the muscles of the lips, mouth, and

larynx. On the other hand, the right hemisphere's inability to encode or decode the emotional tone of speech is aprosodia. Patients with aprosodia have a monotonous voice incapable of producing prosodic vocalizations or communicating emotional speech [Givens, 1998b] [Karpf, 2006: 59]. Prosody denotes the musical elements of speech including its melody (intonation) and rhythm (stress and timing) [Thompson et al., 2004: 48].

Recent fMRI studies of emotional vocalizations outside the context of speech such as laughs and cries have suggested the significance of structures such as the amygdala and anterior insula in processing emotive vocalizations [Belin et al., 2004][Karpf, 2006:61]. Similar studies have been done to explore whether the brain responds differently to vocal sounds and non-vocal sounds. In these fMRI studies, vocal sounds and non-vocal sounds induced a similar response in most auditory cortical regions. However, activation elicited by vocal sounds seemed to be significantly greater than activation induced by non-vocal sounds [Belin et al., 2004]. The same studies also aimed to investigate the difference in the brain's processing of instrumental tones and singing, but no difference was observed [Belin et al., 2004]. Nevertheless, studies about the way the brain generates and processes vocal information are still in their infancy. Many aspects of neurophysiological voice processing have not been discovered yet and many of the already constructed theories await further elucidation.

Michael Hunter and his colleagues at The University of Sheffield carried out neuroimaging experiments that aimed to explore how the brain processes and identifies gender and other human qualities from the voice of a speaker. In their fMRI scanning experiment on male subjects, they found that male and female voices activate different areas in the male brain. Their future work will involve implementing the same fMRI paradigm on female subjects. According to their studies, the right superior temporal sulcus (STS) is associated with perceiving femininity in the voice, and the precuneus is associated with perceiving masculinity [Sokhi et al., 2005: 577]. They also claimed that female voices caused more activity in the auditory cortex than male voices did, and that due to their acoustic *"complexity"*, female voices are much more difficult for men's

brains to interpret than male voices. The parameters that define this complexity, however, have not been fully described [Noyes and Frankish 1989 quoted in Sokhi et al., 2005: 577]. Some writers humorously reflected that claim as a reason for men's inability to hear their wives' orders (for example [Mcginty, 2005]). Some researchers also related the *"complexity"* of the female voice to the relative difficulty for computers to *"recognize"* and *"synthesize"* it [Noyes and Frankish 1989 quoted in Sokhi et al., 2005: 577]. Perhaps because male voices are less *"complex"* to perceive and easier to synthesize, 71 percent of the schizophrenics who hear hallucinatory voices report that these voices are male voices, while only 23 percent report that they hear female voices [Sokhi et al., 2005: 572]. In other words, it seems to be easier for the brain to produce an imaginary male voice than an imaginary female voice during hallucination.

Gender is not the only information that the voice can convey and that the brain processes to determine certain aspects about the speaker. The voice carries many other identity and affective cues that allow listeners to recognize individuals and emotional states [Belin et al., 2004]. Voice as a result is sometimes viewed as the *"auditory face*" that allows the brain to determine the gender and the age of the speaker as well as some physical characteristics such as height and weight and also some psychological characteristics such as trustworthiness [Belin et al., 2004]. Von Kriegstein and her colleagues suggest that *"voice and face modules in the brain are functionally connected"* and that the brain forms a "cross-modal" association between voice and face information [Von Kriegstein et al., 2006]. It is not surprising, therefore, to find this "audio-visual coupling" [Von Kriegstein et al., 2006] manifesting itself technologically through the recent move towards manifold forms of audio-visualization and voice-visualization in human-computer interaction.

In order to investigate untapped aspects of voice, which is the main objective of this thesis, it is germane to explore the vocal aspects that *have* been tapped including the artistic-performative, the cultural, the communicative, and the technological.

2.2 Artistic-Performative Aspects

The voice in performance has tended to be confined to the domains of singing, chanting, acting, poetry, and ventriloquism. With the aid of technology, however, performative voice can encompass a vast new area including breathing, whistling, humming, and a multitude of other non-speech sounds. In this section I take a broad-brush approach to explore these artistic-performative aspects of voice. The aim is to pave the way for my later more detailed arguments about the potential of interactive work that exploits non-speech sounds for expressive and performative ends in human-computer interaction.

a. Singing

When the vocalizer is singing, a greater than normal degree of control occurs before delivery. This process involves controlling muscle tension in the neck and the face as well as changing the position of the jaw in order to produce a specific voice register. *"Voice register"* is related to the part of the body that resonates most [Kob, 2002: 9].

The different ranges of voice are usually described as follows [ibid]:

- Pulse register: A creaky voice where the range of the notes produced is lower than the pitch range of normal speech (also known as vocal fry).
- Chest voice: The sound resonates primarily in the chest and the pitch produced is within the range of normal speech.
- Middle voice: The range of notes produced combines chest and head resonance and blends their registers.
- Head voice: The sound resonates in the mouth or the skull – as when shouting – and is at the higher end of the normal vocal range.

- Falsetto: A sound generated by a man higher than his normal vocal range, and 'falsely' sounding like a woman's voice. Yodelling, for instance, is a form of singing that involves shifting voice registers from chest voice to falsetto.

Most singing involves manipulating the voice to produce sensible musical sounds. Not all forms of singing, however, involve sensible words. There are other styles which are based on the production of non-speech vocalizations. These styles include voice instrumental music, throat singing, scat singing, and Mieskuoro Huutajat. I describe these below.

Voice instrumental music, otherwise referred to as mouth music or *vocalise*, is a wordless form of singing that treats the voice as a tonal instrument capable of expressively producing instrument-like rhythm and timbre.

Throat singing, on the other hand, is a style of singing that is used in Mongolia, Tuva, Tibet and close-by regions. It is also referred to as *overtone singing* or *harmonic singing*. Throat singing involves the production of more than one tone simultaneously. Generally, one tone is a low hum and the other is a series of high flute-like melodies.

It is believed that throat singing originally developed as a form of sound mimesis when the first throat singers attempted to mimic natural sounds *"whose timbres, or tonal colors, are rich in harmonics, such as gurgling water and swishing winds."* [Levin and Edgerton, 1999]. These throat singers are thought to believe in *"Tuvan animism"*; the belief that natural objects and animals are inhabited by spirits and possess a spiritual power, and that by imitating their sounds this power is absorbed [Levin and Edgerton, 1999].

Another form of throat singing was developed among the Inuit of Northern America. This type of singing does not involve producing harmonics but rather involves two women singing while facing each other, and sometimes also holding each other. The Inuit's throat singing is practiced as an artistic form, a

musical performance, or even as a game. One player starts vocalizing and the other player responds competitively until the player who runs out of breath or laughs first loses the game. The first player is usually the leader who leaves little gaps between vocalizations for the other player to quickly fill in. Such game-like uses of voice are explored in my own interactive projects as will be discussed in the third chapter.

Scat singing is also a form of singing in which the lyrics of the song are replaced by nonsense syllables. It is employed by jazz singers. A well known scat song is *"Mah Na Mah Na"* written by the Italian film and soundtrack composer Piero Umiliani in 1968 for a Swedish documentary called *Svezia, Inferno e Paradiso* (Sweden; Hell and Heaven).

In 1987, around 20 men formed Mieskuoro Huutajat (Men's Choir Shouters) in Finland (Figure 4). These men performed by shouting and screaming in unison [Knuuti, 2003] rather than by singing conventionally.

Figure 4: Shouting Men of Finland performing on ice by shouting in the Baltic [BBC News, 2004].

The role of vocalizing and particularly shouting in self expression, which will be extensively discussed in the fifth chapter and demonstrated by participants' interactions with my own projects, is expounded by the Choir's director as follows:

"Shouting is a very natural way of expressing yourself; it's much more natural than, say, singing. Shouting is the easiest way to make as much noise as possible." [Sirviö quoted in Knuuti, 2003].

Singing is not the only art form involving non-speech vocalizations. These also form part of the art of magical chanting or even religious chanting like the art of reciting the Quran as discussed next.

b. Chanting

Chanting is the rhythmic, and often repetitive, recitation of vocalizations during sorcery, performing rituals, singing, or supporting a football team.

Magical chants are utterances often used to ward off evil spirits. Some of these chants are nonsense utterances such as the following incantations: abracadabra, alakazam, shazam, presto-chango, sim sala bim. These words were believed by many to have a mystical power capable of changing something from one state to another. The recitation of these spells has a unique prosodic character that is based on an iterative and rhythmic style.

Chants, voices, and ritual songs referred to as aboriginal songlines were also used by the aborigines during 'dreamtime'. Dreamtime or the time of creation is the period before earth existed when there were only spiritual entities and no physical objects were created. It was believed that communicating with the spirits by chanting songlines while travelling across lands would make the spirits create the upcoming rocks, water, and natural landmarks, thus bringing the world into existence.

Religions of the Book such as Christianity, Islam and Judaism encourage chanting under appropriate circumstances.

Football matches – and other sports besides – also involve chanting and shouting to encourage football players and to out-chant the rivals. These chants may be as

simple as shouting the name of the team, shouting the score, or even shouting the name of a particular player in the team repetitively. Some football vocalizations also consist of non-speech utterances such as boos and the sound 'oooo' which is usually used to discourage the rival.

Chanting is also used in game-play in South Asia. The utterance "Kabaddi-Kabaddi" is chanted during a popular game in Bangladesh and Punjab, called "Kabaddi" [International Kabaddi Federation, 2006]. The utterance "Kabaddi" means "holding breath" in Hindi and is a key part of the game. The game involves two teams which compete against each other by trying to conquer each other's territory and tag as many opponents as possible. One of the players in the first team must enter the territory of the other team and tag any of its players by touching them. This player can stay in the other team's territory as long as s/he keeps chanting "Kabaddi-Kabaddi" while holding his/her breath. Only when s/he fails to hold his/her breath, can players from the other team tag him/her before s/he goes back to his/her territory. However, if the player succeeds in returning to his/her territory while chanting "Kabaddi-Kabbadi", then this player can stay in the game.

c. Sound Poetry

Sound poetry is usually taken to refer to vocal compositions in which more emphasis is placed on the non-verbal musical aspects of human vocalizations than on the verbal aspects. This form of poetry, which frees the word from its semantic dimension, is also referred to as abstract poetry, phonetic poetry, or sound art. Sound poetry was part of the Dada movement which was formed in Zurich and Switzerland partly as an "anti-art" reaction to the First World War. Dada, which also influenced painting and theatre, was characterized by irrationality, cynicism, meaninglessness, and negation of conventional laws and tradition.

It has been claimed that the sound poem *Gadji Beri Bimba*, which was recited by Hugo Ball at Cabaret Voltaire in Zurich in 1916, was the first sound poem [Wendt, 2000]:

> *"Gagji beri bimba*
> *glandridi laula lonni cadori*
> *gadjama gramma berida*
> *bimbala glandri galassassa laulitalomini […].*" [Ball 1916 quoted in Finch, 2003]

Hugo Ball and other sound poets who included Richard Huelsenbeck, Tristan Tzara, and Marcel Janco also performed a group poem called *L'Amiral cherche une maison à louer* ('The Admiral is Looking for a House to Rent') [Wendt, 2000]. They performed it by whistling, singing and making noises [Wendt, 2000].

Another sound poem was also performed by Tristan Tzara and Hans Arp at the Cabaret Voltaire in Zurich. Steve McCaffery, who formed a sound poetry group in Toronto describes the performance as the following:

> *"a high energy, performance oriented cacophony of whistling, singing, grunting, coughing and speaking.[…] Defying categorization as either theatre, music or poetry, it emphasized the improvisatory, spontaneous and aleatoric possibilities of multivocal expression"* [McCaffery, 1978].

Kurt Schwitters and Raoul Hausmann followed Hugo Ball's style. Hausmann referred to his work as *optophonetic* and used different typefaces and font sizes to typographically notate the written expressions and to indicate the sound associated with each utterance (Figure 5). Hausmann, however, was not the first person to create a visual representation for sound poetry. The Italian futurist Filippo Tommaso Marinetti is thought to have preceded him in creating a visual representation of sound poetry called *parole in liberta* (words in freedom) [McCaffery, 1978].

Figure 5: Raoul Hausmann's visual representation of sound poetry (left) and Kurt Schwitters' *Bird Song* (right) [Wendt, 2000] .

Schwitters was influenced by Hausmann's *bbbbetFmsbw* written in 1918. His *UrSonata*, which contains the following verse, is considered by most commentators to be as significant as Ball's *Gadi Beri Bimba* in the history of sound poetry:

> *"Ooooooooooooooooooooooooooooooo Bee bee bee bee bee bee bee bee Ooooooooooooooooooooooooooo Zee zee zee zee zee zee zee zee […]."* [Finch, 2003]

Steve McCaffery suggests that it is one of the longest sound poems though it consisted mainly of the short letter 'W' which was "*performed with the full gamut of pitch, tone, volume and emotional intensity"* [McCaffery, 1978].

A recent audio-visual reflection of Schwitters' *UrSonata* was performed in England by the Dutch sound poet and vocalist Jaap Blonk together with the artist Golan Levin [Levin, 2005]. Speech recognition technologies were used to project real-time subtitles that corresponded with Blonk's vocalizations in timing and timbre. Their work represented a modern multimedia adaptation of Hausmann's *optophonetic* and Marinetti's *Parole in Liberta*.

Many other technological manifestations of sound poetry exist. The earliest and simplest of these are the tape recorded sound poems of Henri Chopin and François Dufrêne which are discussed in section 2.6.

d. Ventriloquism

Ventriloquism is an artform in which the ventriloquist deceives the viewer by manipulating, projecting, or throwing his/her voice to give the auditory illusion that it is generated from another source. When the false source is near to the ventriloquist, for example when using a dummy, it is referred to as "near" ventriloquism. When the source is distant, for example when throwing the voice, it is referred to as distant ventriloquism. The artform is thought to have developed from gastromancy, an ancient Greek divination in which the diviner speaks to the consulter without moving his lips in order to deceive the consulter into thinking that the source of the voice is a spirit possessing the diviner.

A puppet is usually used as the false source of the voice while the ventriloquist speaks or makes non-speech noises by keeping the lips and jaw still. This is possible through the substitution of phonemes that involve obvious oral articulation through lip movement by similar sounding phonemes that can be generated by the tongue or at the back of the throat rather than visibly articulated. The perceptual illusion achieved as a result of the interaction between the auditory and visual illusion is referred to as the McGurk effect, which plays an important role in the perception of speech and in lip reading. McGurk conducted an experiment that involved a synchronized playback of a video recording of a person uttering the sound 'ga', and an audio recording of a person uttering the sound 'ba'. He observed that the resulting sound was perceived as 'da' [McGurk and MacDonald, 1976].

Ventriloquism also involves the generation of non-speech sounds such as bird, animal, or other natural sounds.

As section 2.2 reflected, voice has not only been used as a communication tool but has also been treated as an instrument for performance and expression. Its application in the domain of performance, however, seems to be mostly confined to forms of art such as singing, chanting, poetry, and ventriloquism. Its application in the domain of human-computer interaction seems to be mostly confined to verbal communicative forms of interaction, namely speech recognition. The paralinguistic and performative uses of voice in human-computer interaction have not significantly been exploited. Later, I suggest ways of extending the capabilities of the human voice in interactive media and in performance.

2.3 Paralinguistic Aspects

The term *vocal paralanguage* refers to non-speech vocalizations either separately or as accompaniments of speech. When accompanying speech, vocal paralanguage can be an integral prosodic constituent of the verbal utterance (in the form of stress, intonation, rhythm and other patterns), it can be separately interposed between verbal utterances, or both. Vocal paralanguage includes voice characteristics, emotive vocalizations, vocal segregates, and what I will argue can be called 'holistic' interjections. Voice characteristics include timbre, duration, loudness, intensity, pitch, tempo and arguably envelope. Emotive vocalizations include laughing, crying, sighing, cracking, whispering, yelling, moaning, groaning, sneezing, and coughing. Vocal segregates include vocal fillers (such as uh-huh, ooh, um, uh), silent pauses, and other hesitation phenomena.

Albert Mehrabian claims that 55 percent of communication is through body language, 38 percent through vocal paralanguage and specifically vocal tonality, and only 7 percent through actual words [Mehrabian, 1981:76].

Vocal paralanguage is an important social and cultural cue. It may help the listener recognize many aspects about the speaker and the subject even if the

content of speech is incomprehensible. During a phone call, the voice timbre and pitch, not the verbal content, is what indicates that the speaker is a female or even a particular familiar female. That is why one may still recognize the person talking over the phone even if this person is a friend pretending to be someone else. That is also why one may recognize the voice of a famous singer on the radio even if the song is new and unfamiliar. Thus vocal paralanguage is very important in identifying identity, gender, and even age. It is significantly useful in identifying the emotional state of the speaker and understanding the non-explicit meaning of one's speech.

Saying "she is VERY NICE" with a rising pitch could denote that she is really very nice. However, saying "she is veeeerrrry niiiice" with a stretched out intonation and with falling pitch could connote sarcasm. Words can thus convey the opposite meaning when intonation and other paralinguistic cues are used. Language alone may not convey non-linguistic information such as gender and age, nor extra-linguistic information such as attitude and emotion. This might actually be one of the main factors that make the improvement of speech recognition systems moderately slow. Quast suggests that there is great potential in adding further non-speech dimensions to speech recognition systems; these dimensions can make recognition more accurate and allow the extraction of more information [2003]. However, the non-speech aspects of voice which include prosody are "*still not widely used in automatic speech processing systems, especially not in commercial systems*" [Batliner and Nöth, 2003]. Prosodic models have been applied "*only occasionally, rather in basic research, but almost never within an existing end-to-end system*" [Batliner at al., 2001] [Batliner and Möbius, 2005].

The non-verbal vocalizations which people generate form one of the most immediate modes of expressive affective communication [Beeman, 1998]. Language itself is deficient in conveying affective states to others, and it requires significant enhancements from other dimensions of communication [Beeman, 1998]. Thus, prosody – the variations of frequency, loudness, stress, and rhythm

in speech – can sometimes be more significant in conveying meaning than speech per se.

Some studies, however, have shown that only adults give more weight to vocal paralanguage than to linguistic content:

> *"children of preschool and early school age give less weight to vocal paralanguage than to lexical content in interpreting affective discrepancy […] For example, a speaker saying 'You're my favorite person' in an angry voice would be interpreted as feeling happy […].In contrast, adults appear to base affective judgments of sarcastic or joking utterances on paralinguistic, rather than lexical, content […]"* [Friend and Bryant, 2000].

Other evidence of the inefficiency of language alone in the communication of affective information is the current use of emoticons in text-based chatting services to enable the reader to recognize the writer's affective state and attitude. Moreover, vocal signals such as pitch, volume, speech rate and intonation can be better than words in communicating interest and attraction towards members of the opposite sex. If a man greets a woman with a deep-toned, low-pitched voice that rises in intonation at the end of the greeting, then this kind of voice may indicate attraction or interest [Fox, 2003]. However, if the greeting is short, high-pitched and monotonous then this could be an indication of lack of interest [Fox, 2003].

Anmol Madan, a Ph.D. candidate at Massachusetts Institute of Technology (MIT), is developing the *Human Interest-Meter* which allows for the use of mobile phones to measure interest in males and females. Through measuring the stress and activity features of speech, Madan intends that the interest meter can be used for different purposes including advertising and online dating. Madan and his colleagues at MIT also developed a *Speed Dating* system that predicts if a person is romantically interested or not during a conversation (Figure 6). The system, which is integrated into a personal digital assistant (PDA), measures attraction and turn-taking cues in real-time. Detecting interjections such as 'aha' and 'yup' indicates attraction and improves the possibility of the other person's acceptance. [Madan et al., 2005].

Vocal signals are very significant *"turn-yielding cues"* [Fox, 2003] in conversations. A falling intonation and a drop in volume at the end of a sentence may indicate that a speaker is done talking and is giving the turn to the conversation partner to speak.

Figure 6: Two users experimenting with the *Speed Dating* system to measure the other person's interest and attraction [Madan et al., 2005].

Another significant use of the subtleties of voice is during presentations where a monotonous voice may cause boredom in the audience. A loud low-pitched voice combined with various acceleration, deceleration, and pausing techniques may reflect a sense of credibility and keep the audience interested. In this context, Will Stoltzman developed *ElevatorRater* [Stoltzman, 2005]. This system measures voice characteristics during a speech, and rates the persuasiveness of the speaker.

Emerging multimedia technologies, however, still have not fully exploited our natural abilities to perceive sound characteristics and to generate paralinguistic or non-speech vocalizations. The following paragraphs offer a more in-depth

exploration of these paralinguistic elements before suggesting how they can turn out to be an expressive source of input to interactive media.

a. Vocal Characteristics

Voice characteristics are properties of the analog voice signal and are associated with laryngeal anatomy, vocal tract configuration, and the vibration of vocal folds. The main characteristics of voice include frequency, volume, timbre, rhythm, and duration.

The frequency of the voice corresponds to the number of times per second the vocal folds come together during phonation [Scherer, 2000]. It is affected by the mass, length, and tension of the vocal folds. It is strongly related to pitch but one distinction is that "*frequency describes a physical phenomenon, while pitch describes a perceptual phenomenon*" [Ballora, 2006]. In other words, pitch can be subjectively perceived by the ear, whereas frequency is measurable; one can change the perceived pitch without changing frequency value [Karpf, 2006: 35]. The frequency levels that the human ear can hear are in the range between 20-20,000 Hz.

Pitch is argued to show the most emotion-dependent fluctuations [Nwokah et al., 1999]. Mithen suggests that when one is emotionally upset, it is not easy to stop the pitch of his/her own voice from rising [Mithen, 2005]. Pitch range is the variation in average pitch within a voice signal. When a voice signal has a very narrow pitch range, it is called a monotone. According to Van Leeuwen "*the wide pitch range conveys 'excitement', 'surprise', 'anger'; while the narrow pitch range conveys 'boredom', 'misery'*" [1999].

Experiments conducted by Saffran and Griepentrog have shown that we are all born with perfect pitch but this ability is replaced in many of us by relative pitch as we grow up [Saffran and Griepentrog, 2001]. Mithen thinks that this explains why intense music practice during childhood may enable the maintenance of

perfect pitch into adulthood [Mithen, 2005: 78, 79]. He also argues that acquiring language leads to the *"unlearning"* of perfect pitch. People who can retain perfect pitch are those who practice music intensely during childhood or musical savants and autistic children whose cognitive impairments impede their language acquisition [ibid].

Volume depends on amplitude, which is the air pressure level of the sound wave. The most commonly used unit of measuring volume is the decibel (dB). The volume of the average voice during a conversation is 60 dB, that of quiet speech is around 40 dB, and that of shouting is around 75 dB [Karpf, 2006:41]. Volumes around 120 dB may create a sense of touch or movement and those louder may cause pain [ibid].

There is a distinction between volume and loudness; volume is measurable and it depends on the intensity of sound while loudness is subjectively perceptible and it can be affected by parameters other than intensity, such as the frequency of sound. For this reason, the Phon has been adopted as a unit for measuring perceived loudness [Hartmann, 1998].

Timbre is the quality that distinguishes a particular voice from any other voice of the same pitch and volume. Timbre is determined by the various overtones present in the sound, and their relative strengths. For instance, a musical instrument may be described in a number of ways including 'rich', or 'tinny', or 'hollow ', or 'fat', or 'thin', or 'dry' [National Center for Voice and Speech, 2005]. Overtones are the multiple secondary frequencies that exist in combination with the lowest frequency of the waveform. This lowest frequency is the fundamental frequency (F0). When these overtones are integer multiples of the fundamental frequency, they are called harmonics. When they are fractional multiples of the fundamental frequency, they are called partials.

Tempo, when related to speech rather than music, is the speed or rate of speech. The average tempo of an adult American or British speaker is around 120 to 150 words per minute (wpm) [Karpf, 2006:42].

Duration is how long the sound lasts. Like a wide pitch range, a wide durational range can be tied to the expression of affect [Van Leeuwen, 1999].

Many researchers have attempted to find a relationship between voice characteristics and emotions. Klaus Scherer, for instance, attempted to find the relationship between voice characteristics such as pitch, tempo and rhythm and the emotional expressions they convey. He found that a slow low-pitched voice conveys sadness, while a vocal expression that has a fast tempo and large pitch variations conveys happiness [Scherer, 1995:238].

Mehrabian conducted a number of experiments in order to find emotional correlates of the implicit prosodic characteristics of speech as perceived by listeners. Among these characteristics he found that the duration of a person's speech in a specific period of time indicates the level of dominance. Longer durations indicate more dominance. Louder voice and faster rate, however, indicate persuasiveness and influence while higher pitch and slower rate indicate submissiveness and passiveness [Mehrabian, 1981: 48, 49]. Monotonous voice, which has no variety in pitch, indicates that the person is less credible and less persuasive [Mehrabian, 1981: 152]. High-pitch level and slow speech rate indicate deceit and untruthfulness [Mehrabian, 1981: 153].

The following table summarizes the paralinguistic information conveyed by various voice characteristics based on studies and experiments undertaken by a number of researchers (Table 1).

Table 1: A table illustrating the paralinguistic information conveyed by different voice characteristics.

Vocal Characteristic		Paralinguistic Information	Example	Reference
Frequency which determines pitch		emotional affect	*"changes in pitch are an important cue for affective messages".*	[Slaney and McRoberts, 1998]
		gender	Men have lower pitch than women.	[Tidwell, 2003]
		size	pitch is related to size, and it is the size-related properties that lead to perceptions related to dominance and submissiveness	[Ohala 1984 and Bolinger 1964 quoted in Huron et al., 2000]
		age	*"the pitch, is the most important factor that conveys gender and also age information".*	[Ho Ching-Hsiang, 2001]
	Frequency modulation	social status/dominance	*"There's a hidden battle for dominance waged in almost every conversation--and the way we modulate the lower frequencies of our voices shows who's on top".*	[Schwartz 1996 quoted in Givens, 1998a]
	High pitch	submissiveness	*"higher vocal pitch is associated with submissiveness, whereas lower vocal pitch is associated with social dominance."*	[Ohala 1984 and Bolinger 1964 quoted in Huron et al., 2000]
			"Submissive humans, on the other hand, make high-pitched sounds to give the impression of being as small and non-threatening as possible".	[Karpf, 2006; 175]
		deference politeness lack of confidence	*"high or rising vocal pitch is associated with politeness, deference, submissiveness and lack of confidence".*	[Bolinger 1964 quoted in Huron et al., 2000]
		friendliness	*"low-pitched sounds are generally associated with aggressive signalling, whereas high-pitched sounds are generally associated with friendly, appeasing, or fearful signals".*	[Huron et al., 2000]
		tenseness helplessness nervousness	*"We associate high pitched voices with tenseness, helplessness, & nervousness".*	[Meade, 2002]

		positive emotional valence	*"High pitch is often correlated with positive emotional valence whereas low pitch is more likely to signal sad emotional valence".*	[Lattner et al., 2005]
		anger	*"…emotions such as anger, fear, and joy are all characterized by raised fundamental frequency (f0) and high intensity, whereas emotions such as sadness and boredom are expressed with low f0 and low intensity.""*	[Johnstone et al., 2005]
		fear		
		joy		
	Low pitch	boredom		
		sadness		
			"… low pitch is more likely to signal sad emotional valence".	[Lattner et al., 2005]
		dominance	*"Ohala has shown that higher vocal pitch is associated with submissiveness, whereas lower vocal pitch is associated with social dominance."*	[Ohala 1984 quoted in Huron et al., 2000]
		authority	*"low or falling vocal pitch is associated with authority, threat, aggression, and confidence".*	[Bolinger 1964 quoted in Huron et al., 2000]
		threat		
		aggression		
		confidence		
		strength	*"We associate low pitch voices with strength, sexiness and maturity."*	[Meade, 2002]
		sexiness		
		maturity		
			"A lower voice also connotes maturity".	[Corinthians quoted in Karpf, 2006]
spectral formants (F1, F2. . .) which determine timbre		identity	Men have lower formant frequencies than women.	[Lattner et al., 2005]
		body size	Formants are directly related to the size of the vocal tract and can therefore provide estimates of body size.	[Belin et al., 2004]
		vocal familiarity	The spectral formants determine the timbre of voice, so that the listener can recognize familiar voices and discriminate unfamiliar voices.	[Lattner et al., 2005]
		gender	The third and fourth formants particularly determine gender because they depend on the shape of the pharyngeal cavity, which is larger in males.	[Lattner et al., 2005]
Volume which	Loudness	anger	*"Increased volume is usually associated with*	[CCMS, 2003]

determines Loudness			*anger or strength and decreased volume with confidentiality".*	
		authority physical power/strength	*"Loudness indicates strength in Arabic cultures and softness indicates weakness; indicates confidence and authority to the Germans; indicates impoliteness to the Thais; indicates loss of control to the Japanese".*	[Tidwell, 2003]
			"the acoustic power might suggest the physical power of the individual or signal the individual's willingness to engage in physical confrontation."	[Huron et al., 2000]
		impoliteness	Loudness indicates impoliteness to the Thais.	[Tidwell, 2003]
		loss of control	Loudness indicates loss of control to the Japanese.	[Tidwell, 2003]
		urgency clarity	*"loudness may be associated with the urgency of the signal. That is, a loud vocalization may indicate a high desire to communicate or to communicate clearly".*	[Huron et al., 2000]
		hostility aggression	*"increased loudness is also likely correlated with hostility or aggression".*	[Huron et al., 2000]
	Softness	confidence shyness	*"A loud voice can be associated with confidence, enthusiasm, and self-assuredness […] A soft voice can reveal shyness, a lack of self confidence and a feeling of inferiority."*	[H2 Training and Consultancy, 2006]
			"Symptoms of shyness may include gaze aversion, a soft tone of voice, and/or hesitant or trembling speech."	[Davies, 2003]
			"A soft or inaudible voice may be associated with a psychological condition such as shyness […]".	[Encyclopedia of Nursing and Allied Health, 2007]
			"Speaking in a soft voice may mean shyness in some cultures or a matter of politeness in others. Speaking in a loud and brash manner might be hostile behavior in one culture or totally acceptable in another."	[American Public Works Association, 2001]

		confidentiality	*"increased volume is generally associated with anger and decreased volume with confidentiality".*	[CCMS, 2003]
		politeness	Soft voices are *"perceived as timid or polite".*	[Meade, 2002]
		weakness	*"Loudness indicates strength in Arabic cultures and softness indicates weakness."*	[Tidwell, 2003]
		timidity	*"it is more difficult to imagine being aggressive in the absence of loudness, and timid in the absence of quiescence."*	[Huron et al., 2000]
			Soft voices are *"perceived as timid or polite".*	[Meade, 2002]
		gender/ female	*"women tend to speak higher and more softly than men."*	[Tidwell, 2003]
Tempo	High speed	competence	*"When a speaker uses a faster rate they may be seen as more competent."*	[Meade, 2002]
		dignity	*"90-100 wpm suggests incapacity, dignity, or vanity".*	[Karpf, 2006: 42]
		incapacity		
		vanity		
		impatience	*"high speed (a clipped delivery) may be associated with impatience and low speed may be associated with pensiveness or uncertainty."*	[CCMS, 2003]
	Low speed	pensiveness		
		uncertainty		
intonation	Rising intonation	uncertainty	*"questions have rising intonation whereas declaratives have falling intonation."*	[Hirst and Di Cristo, 1998]
		interrogation		
	Falling intonation	declaration		

Table 1 shows that voice can convey much information about personality and feelings. One key issue that emerges in the table, and which will be pursued later in the context of my own work, is the possible relationship between voice and shyness.

b. Emotive Vocalizations

Expressing emotions is a vital component of our daily human interactions. A significant part of this emotional communication occurs non-verbally through the generation of emotive vocalizations as laughter, crying, and screaming. According

to Scherer *"Affect vocalizations are the closest we can get to the pure biological expression of emotion and one of the most rudimentary forms of communication."* [Friend and Bryant, 2000]. This probably suggests that vocal expression of emotions is as important as, if not more important than, other biological forms of expression. Unfortunately, *"vocal expression of emotion has not received as much attention as facial expression"* [Scherer 1986 quoted in Laukka, 2004].

It is believed that non-verbal expressions and emotive vocalizations existed before speech as a communicative and expressive channel [Ruch and Ekman, 2001]. Some researchers think that speech may have developed from non-verbal utterances such as laughter, crying, screaming etc. Mithen believes that these *"pre-linguistic"* utterances, which he refers to as *"Hmmmm"* (holistic, manipulative, multimodal, musical, and mimetic) and which are shared between music and language, represent a single precursor to what has now split into music and language [Mithen, 2005:27]. Lowey, who seems to agree with Mithen, uses Stansell in support of a connection between language and music:

> *"As evidence of this closeness, Stansell (2002) notes the process that occurs when someone begins to cry while speaking. Prosodic features of air control, pacing, tone, and tenor become more exaggerated and emotions break through in musical representations while language regresses into babbling."* [Loewy, 2004]

In some cultures certain vocalizations have emotional and psychological effects on the utterer. Some people, for instance, think that screaming is cathartic. In California, a purgative psychotherapeutic method called primal scream therapy was developed by Arthur Janov. A primal scream therapist encourages patients with emotional disorders to express their feelings by screaming in order (it is hoped) to re-live their birth experience.

In Taiwan, a laughter club has recently been opened to encourage people to laugh and have better health and a happier life. Members start the laughing sessions by uttering voices like ho-ho-ha-ha followed by practicing around forty-one kinds of laughter including crazy laughter and Lion laughter (Figure 7). Laughter yoga

was originated in Mumbai by Dr. Madan Kataria in 1995 [Kataria, 2002] and has since spread across the world.

Figure 7: Dr. Kataria leading a laughing session in Mumbai [Strubbe, 2003].

Darwin was one of the main scientists who tried to discern the nature of the reiterated sound that humans produce during laughter [Darwin, 1998: 206]. The laughter researcher, Robert Provine, made many discoveries during his study of the acoustic and rhythmic pattern of laughter. He noted that laughter consists of a series of vowel-like sounds each of which is 75 milliseconds long. These short sounds are repeated at regular intervals every 210 milliseconds [Provine, 1996]. He also found that the structure of laughter can either be of the "ha ha" type or the "ho ho" type, but it cannot combine both as in "ha ho ha ho". Provine also discovered that there is a gender difference in the extent of laughter and that women laugh much more than men [Provine, 1996]. He observed that a male comic can make listening audiences laugh more than a female comic can, but his evidence that this observation is cross-cultural is *"limited"* [Provine, 1996].

Provine also studied the placement of laughter in speech in what he referred to as the "*punctuation effect"* [Provine, 1996]. He noticed that the laughter is usually expressed like a vocal punctuation mark during a pause at the end of a phrase or utterance rather than occurring in the middle of an utterance.

Laughter is exploited by comedy producers who use an artificial canned laughter in their shows as a means of *"social bonding"* [Mithen, 2005: 81, 82]. The use of laughter and other emotive vocalizations as means of social bonding has also long ago been suggested by Darwin. He suggested that these vocalizations were employed throughout the animal kingdom *"as a call or as a charm by one sex for the other"* or *"as the means for a joyful meeting between the parents and their offspring, and between the attached members of the same social community."* [Darwin, 1998: 206]

The role of laughter and other vocalizations as a means of social bonding in the field of interactive media is incorporated in my own projects, which are discussed in the third chapter.

Another universal emotive vocalization is crying which usually occurs as an emotional response to pain, hunger, sickness, sadness, or even happiness. Crying is the first vocal expression and communication signal that a newborn generates. Its characteristics, including its pitch, tone, and dynamic strength are the main indicators of the baby's health [Loewy, 2004].

The different emotions expressed by babies' cries can be distinguished by the mother. The causes include hunger, anger, frustration, and pain. Pain crying, for instance, starts with a sudden cry that is followed by long cries and breathless pauses. The pain cry has been noted to be of 3 to 4 seconds at first, followed by 7 seconds of breath-holding before the next cry [Karpf, 2006: 95]. The hunger cry, on the other hand, is distinguished through its rhythmic nature and the duration of each cry which lasts for about 0.65 seconds [Karpf, 2006: 95].

A baby cry analyzer has recently been developed to help parents know why their baby is crying. Pedro Monagas, the developer of this sound-sensitive electronic device, claims that it can analyze the pitches and volume levels of crying with an accuracy of 98 percent [Monagas quoted in Bradbury-Carlin, 2007], and it can then categorize the reason into one of five causes: hunger, stress, sleepiness, boredom, annoyance.

Because the Japanese believe that crying is good for health, they annually hold a crying competition for babies. They refer to it as "crying sumo". In this competition, which has been going on for more than four-hundred years, two babies are carried facing each other while sumo wrestlers try to make them cry by shaking them gently [Karpf, 2006: 97], making scary faces, or shouting. The wailing champion is the one who cries first and an award is also given to the baby with the loudest cry.

Crying and laughing persist into adulthood, irrespective of verbal skills. The approach of encouraging these vocalizations is currently restricted mainly to laughter clubs, primal scream therapy and crying contests. In the third chapter I apply this approach to interactive media.

c. Vocal Segregates

Hesitation or difficulty in retrieving information during communication often results in the generation of vocal segregates. Vocal segregates are fillers, pauses, exclamations, and other hesitation utterances including 'um', 'ah', and 'uh'. Utterances such as "um" and "ah" are sometimes used to indicate comprehension, awareness, understanding, and agreement. 'Huh' is sometimes used to communicate uncertainty, while 'uh-huh' is sometimes uttered to indicate listening attentiveness, formality or consent. A pause is also considered a vocal segregate, and a prolonged pause during speech often indicates hesitation and lack of confidence.

Vocal segregates refer to sounds that do not easily fit into written units structured in sequence as part of a sentence that conforms to the syntactic rules of a language. The problem about identifying an utterance as a vocal segregate or even as a paralinguistic rather than a linguistic vocalization is that the boundary between the linguistic and the paralinguistic is blurred. In the light of this, it is not surprising that researchers have different views on the classification of vocalizations. These views will be problematized in the following sections.

d. "Holistic" Interjections

Mithen uses the term *"holistic"* to refer to primitive utterances that are not composed of segmented units [Mithen, 2005]. As mentioned in the previous section, he suggests that these utterances are part of what he calls *"Hmmmmm communication"* and which refers to the following:

> *"a peculiar proto-music/language that was holistic (not composed of segmented elements), manipulative (influencing emotional states and hence behavior of oneself and others), multimodal (using both sound and movement), musical (temporally controlled, rhythmic, and melodic), and mimetic (utilizing sound symbolism and gesture*". [Mithen 2005 quoted in Dissanayake, 2005]

For the purposes of this section, I will adapt the term *"holistic"* vocalization (or *"holistic"* interjection) to refer to a vocal expression that does not fall under any linguistic syntactic rule and that encodes a complete message with a fully comprehensive meaning in a short unit of vocalization that can stand on its own. Examples of holistic vocalizations, which can replace a whole sentence, include Hooray (I am happy), Wow (I am thrilled), Yummy (the food is delicious), Shhhhh (don't tell anyone), and Ouch (I am in pain).

Although these expressions consist only of one short utterance, they encode a complete message which if written or spoken in words would form a full sentence. According to Mithen, for instance, *"'YumYumyYummyYummm' is not a word but a holistic phrase; it initially means something like 'eat this lovely food, my darling'*." [Mithen, 2005: 204]

The main characteristic of many of these expressions is that they do not have generally accepted grammatical and orthographic/spelling rules, nor can they be found in all standard dictionaries as fundamental units of language. The fact that there isn't an agreed-upon spelling for holistic vocalizations emphasizes the precedence of the sound produced to express instant emotions over the meaning conveyed through the communicated message. These expressions or interjections are usually used to convey *"emotive or attitudinal content rather than referential meaning"* [Department of Linguistics and Modern English Language (LAMEL) at Lancaster University, 2002].

Kaplan who investigated the difference between 'ouch' and 'I feel pain', thinks that interjections *"are better analysed in terms of a Semantics of Use rather than (or as well as) a Semantics of Meaning"* [Kaplan 1998 quoted in Wharton, 2003: 182]. He suggests that the difference is based on whether the content is descriptive or expressive:

> *"[…] while 'I feel pain' has descriptive (truth-conditional/propositional) content, ouch has expressive (non-truth-conditional/non- propositional) content"* [Kaplan 1998 quoted in Wharton, 2003: 182].

One possible piece of evidence that holistic expressions may serve an expressive role more than a communicative role is the fact that a person may utter them as part of self-talk while playing a Nintendo game alone or while cursing a crashed computer. The analysis of the characteristics of these expressions when elicited by playing interactive games may facilitate the development of computers that recognize and react to these vocalizations as will be outlined in the third chapter.

In some countries, the mother makes an 'ssss' sound to encourage her baby to urinate. The 'ssss' sound in this case could also be considered a holistic utterance which is as if to say "go pee" if stated as a full-sentence command. This sound is perhaps used because it resembles the sound of flowing water or urination [Bauer, 2001].

Another important characteristic of these short expressions is that their brevity allows their loudness, rate, and other voice characteristics that are important to the meaning, to be easily controlled. Holistic vocalizations have the impression of urgency and quickness because they are usually used to express instant and sudden feelings. They all have an *"expressive and instinctive nature"* and they *"retain an element of naturalness and spontaneity"* [Wharton, 2003: 181]. This may explain why some of them are sometimes referred to as *"response cries"* [Wharton, 2003]. In light of this, these utterances are necessary to express instant feelings which can be expressed faster and louder if uttered paralinguistically than if uttered in words. When one is telling a friend a secret and someone immediately comes in, it will be faster to say 'shhh' than 'don't speak':

> *"Shh does not convey emotion: but it could be argued that its voiceless quality, together with the fact that it can be uttered continuously, make it a particularly suitable sound for urging someone to be quiet."* [Wharton, 2003: 207]

Uttering 'yo' to call Mark or to catch him before he leaves can be louder and faster than saying: 'Mark, come here'. Moreover, to express instant happiness, it is easier, faster, and more expressive to laugh or utter 'yay' than to say 'I am happy'. Loudness is also important here because 'yay' can be louder than 'I am happy' and thus can be more expressive of joy. One might speculate that people probably use holistic vocalizations rather than words, sometimes, because the speed and easiness that characterise the paralinguistic quality of these vocalizations allows less time to be given to expressing the emotion and more time to feeling it. Immediacy and ease of utterance are features that facilitated my own exploitation of these vocalizations rather than spoken commands as real-time controllers of my interactive work.

The fact that these expressions are short also allows for their reiteration and continuation either for emphasis, for more expressiveness, or even for continuous control of interactive media as in the case of my voice-controlled projects which are demonstrated in the third chapter. When one makes a mistake, it is easier and faster to utter 'oops, oops, oops' repetitively to strongly indicate that whatever

occurred was not deliberate than to utter a whole sentence that communicates the same meaning. Again, reiteration is probably proportional to the instinctiveness, expressiveness, and involuntariness of the action which induced the iterative vocal response. The involuntariness of natural laughter, for instance, could be the reason why laughter has a "*peculiar reiterated character*" as Darwin noted [Darwin, 1998: 206].

According to some linguists, even utterances like bother, damn, bloody, hell, shit (and other expletives), goodbye, yes, well, no, thanks could be considered interjections [Wharton, 2003:173] especially when used meaninglessly to express an attitude or emotion. They could also be referred to as phatic expressions or response cries [Wharton, 2003: 176] because their function is only to express rather than to describe and they only serve the purpose of maintaining a social conversation or expressing an instant emotion. The term *phatic,* which was first coined by Malinowski, refers to conversations that lack *"intrinsic meaning"* [Karpf, 2006; 185]. These include conversations such as *"what lovely weather today"* which are uttered for the sake of establishing social links and *"breaking the silence"* [Karpf, 2006; 185].

It is thought that unlike normal words of which phonemes are processed individually as sound units in the left hemisphere of the brain, expletives are processed as a whole unit in the right hemisphere [Ardó, 2001]. This suggests that the brain treats swear words as paralinguistic vocalizations. This is also suggested by the fact that aphasics and those suffering from Tourette Syndrome, some of whom lose the ability to generate speech, retain the ability to generate swear words [Ardó, 2001].

Where vocalizations are processed in the brain, however, is not the only factor that may determine the linguistic/paralinguistic classification of vocalizations. Volition also plays a significant role in relation to the linguistic/paralinguistic continuum. A number of other factors, which I concentrate on in the next section, may also be useful in attempts to classify vocalizations and the degree of volition involved in each type.

e. Classifications of Vocalizations

In the earlier parts of this thesis I have discussed how certain paralinguistic aspects of voice have been approached in various fields including the physiological, the artistic-performative, the cultural, the communicative, and the technological.

In the discussion so far, I have already indicated that the boundaries between the linguistic and the paralinguistic are not hard and fast. Following Schötz [Schötz, 2002] (Figure 8) the two may be broadly distinguished in this way: Language is spoken or written communication by words based on a conventional arbitrary code. Paralanguage is all other non-speech vocalizations that may or may not accompany speech. However not all classifications of vocalizations are commonly agreed upon in typological studies. Vocalizations have been divided into linguistic vocalizations and categories called paralinguistic, extra-linguistic, or non-linguistic [Schötz, 2002] (Table 2).

Table 2: Some classifications of vocalizations in literature [copied from Schötz, 2002].

	paralinguistic	extralinguistic	non-linguistic	(intra)linguistic
Carlson 2002		inhalation, exhalation, smacks, hesitation sounds		
Carlson & Granström 1997		attitudes, emotions		
Laver 1980	affective information	voice qualities identifying the speaker		
Lindblad 1992	emotions, attitudes, age, sex, dialect, sociolect, (health?)			
Marasek 1997	non-linguistic and non-verbal information; attitude, emotions, dialect, sociolect	physical &physiological features; age, sex, habitual factors		
Mixdorf 2002	speaker attitude, intention, dialect, sociolect		emotions and health	
Quast 2001	momentary changes; whispering, emotions	the speaker's basic state; physical, physiological (body & larynx size)		
Roach 1998	intentional; voice qualities (modal, falsetto, breathy voice etc.) and voice qualifications (non-linguistic vocal effects (laughing, sobbing, tremor etc.)		unintentional age, sex, health	
Traunmüller 2000, 2001	organic; age, sex, health, and expressive; emotion, attitude, adaptation to environment	perspectival; distance, direction, transmission channel		linguistic: dialect, sociolect, speaking style

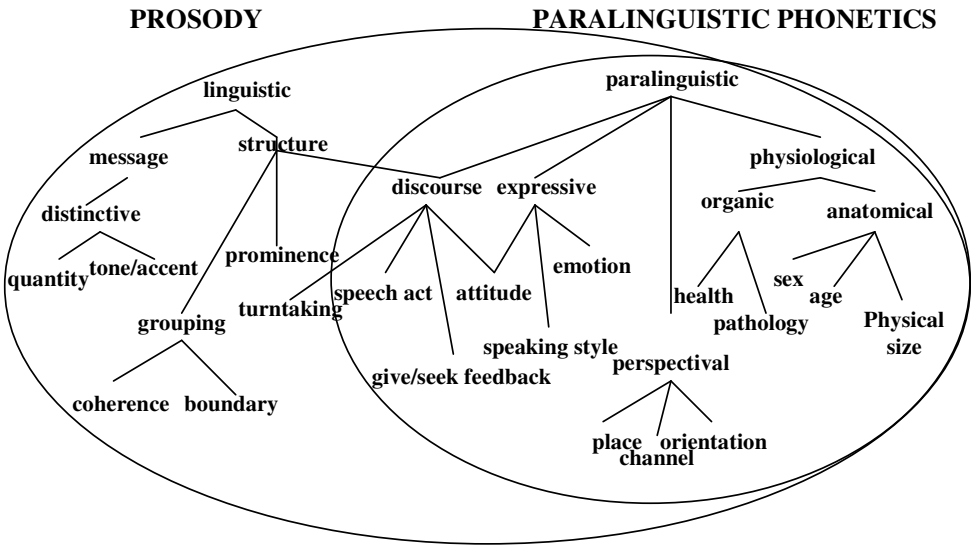


Figure 8: An impressionistic representation of paralinguistic and prosodic features [based on Schötz, 2002].

In the course of this research I have found it necessary to fix on certain definitions, which I clarify here before continuing.

I use the term *vocal paralanguage* to refer to any phenomena *"which generally fall outside the boundaries of phonology, morphology and lexical analysis. These phenomena are the voice qualities and tones which communicate expressive feelings, indicate the age, health and sex of a speaker, modify the meanings of words, and help to regulate interaction between speakers."* [Carey, 1980].

Vocal paralanguage will be divided into *paralinguistic vocalizations* (not necessarily linked to speech) and *prosody* (linked to speech). It includes vocal characteristics, emotive vocalizations, vocal segregates, and holistic interjections (Figure 9). Paralinguistic information that can be extracted from those vocalizations will be subdivided into two categories: extra-linguistic and non-linguistic. Extra-linguistic information will denote the emotional state of the communicator such as feelings and attitudes, and non-linguistic information will denote the physical and biological states such as gender, age, and health.

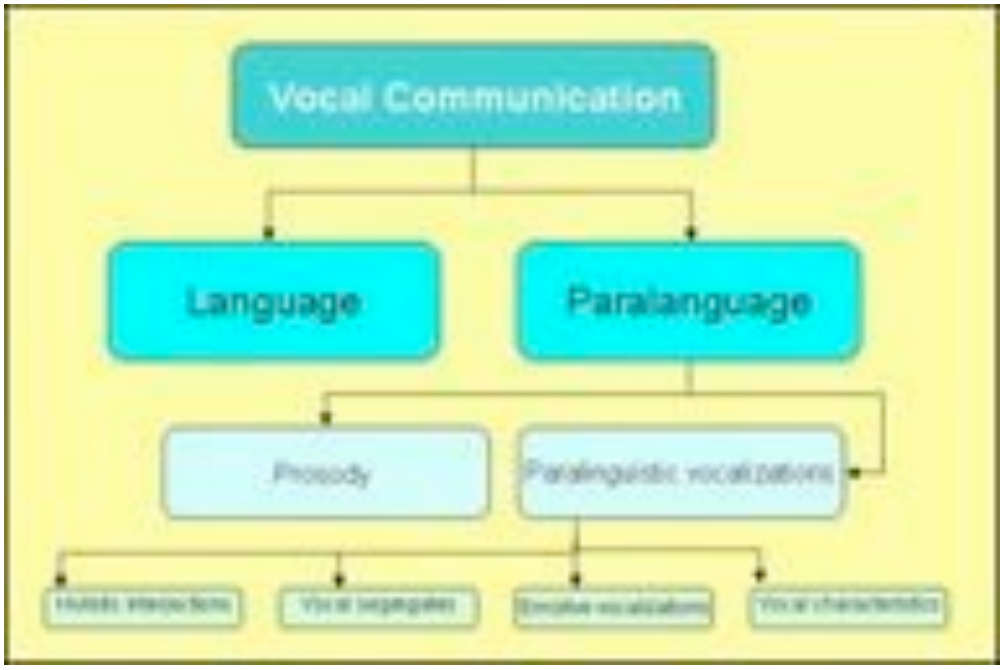


Figure 9: A diagram of the classification of vocal communication.

When accompanying words and strictly speech-related, vocal cues will be referred to as prosody. Prosody will be divided into linguistic prosody and paralinguistic prosody. In tonal languages, such as Chinese Mandarin, changing the tone or pitch of a word may change its lexical meaning. In such cases, shaping words' rhythmic, intonation and stress patterns for phrasing, accentuation and other lexical purposes will specifically be referred to as linguistic prosody. Shaping words' rhythmic, intonation, and stress patterns to convey emotions, attitudes, and other connotative non-lexical information will be referred to as paralinguistic prosody. A justification of these classifications is offered below.

The main factors that I have used to determine the classification of an utterance include: syntax dependence as opposed to syntax independence (whether the utterance can stand alone), linguistic non-productivity as opposed to linguistic *"productivity"* (whether the sound of the utterance can be reproduced in phonemes) [Wharton, 2003: 174], meaninglessness as opposed to meaningfulness (whether the utterance encodes a meaningful concept), volition

as opposed to spontaneity, naturalness, and instinct. The boundaries here, however, are very blurred and the predictive reliability of the criteria is quite low.

Voice characteristics are absolutely paralinguistic because they do not have a generally accepted written representation.

There is no doubt that a burp, a cough, an exhalation or inhalation, a sneeze, and yawn are all *"non-lexical"* or non-linguistic vocalizations [Sperberg-McQueen and Burnard, 2005]. Most of these vocalizations do not have a written equivalent or any kind of textual representation. It is only possible to refer to most of them by the name rather than by any linguistic phonemes that correspond to their sounds. In other words: They do not *"match the typical word phonology of English"* [Wharton, 2003: 199]. They also do not encode concepts as words do, nor are they *"rule-governed"* [Wharton, 2003: 197] as language is. Therefore, it may be reasonable to suggest that they are totally paralinguistic and are the second in the list of paralinguistic vocalizations. Laughter might also be considered among the above mentioned paralinguistic vocalizations, but since it can, to a limited extent, be represented textually using the phonemes 'Ha ha ha ha', I decided to place it after the above mentioned vocalizations in my paralinguistic scheme (Figure 10).

After laughter, vocal segregates are also considered paralinguistic. These may be represented textually, but only to a limited extent. The glottal stops, voice characteristics, and other vocal apparatus manipulations required to emit these vocalizations cannot be represented but only described textually.

There is no general agreement on how interjections, which I have referred to as holistic vocalizations, should be classified. Some researchers consider them *"semi-lexical"* or semi-linguistic [Sperberg-McQueen and Burnard, 2005]. Others think that they should be referred to as *"quasi-lexical vocalizations"* [Department of Linguistics and Modern English Language (LAMEL) at Lancaster University, 2002].

Some researchers, however, think that the above mentioned interjections cannot exactly be considered linguistic. They even call them *"vocal gestures"* because of the strong connection between them and gestures [Wharton, 2003: 197]. The interjection 'brrrr', for example, will most likely be uttered while shivering when one is cold.

Some researchers also think that utterances such as ouch and oops *"are not productive linguistically"* and cannot be considered linguistic while imprecations, when not used as interjections, are considered linguistic because of their *"productivity"* [Wharton, 2003: 174].

Among the valiant attempts to classify vocalizations are those of Wharton [Wharton, 2003:175] who suggests that interjections are *"marginal to language"* because they are syntax-independent and they only signify feelings.
Ameka suggests that interjections should be divided into primary and secondary interjections. Primary interjections are utterances, such as 'oops' and 'ouch', which can only be used independently as interjections and cannot be *"movable between word-classes"* [Ameka 1992 quoted in Wharton, 2003: 175]. Secondary interjections are utterances, such as 'hell' and 'shit' which can either be used syntactically or independently *"to express a mental attitude or state"* [Ameka 1992 quoted in Wharton, 2003: 175].

Encouraged by Wharton's view that *"interjections seem to share with para-linguistic or non-linguistic behaviours the property of being partly natural and partly coded"* [Wharton, 2003: 183] I decided to place most interjections in the middle of my linguistic/paralinguistic tentative sketch (Figure 10). My classification also depends on whether the interjection is *"linguistically productive"* or not*: "ugh differs from yuk in that the former ends in a velar fricative that is not linguistically productive in English"*. [Wharton, 2003: 199]

The last form of utterance to be considered is onomatopoeic words. Since these words are "*linguistically productive"* [ibid] and can be placed syntactically within a sentence (ex. He hiccupped during the test), I have placed them at the

very end of the linguistic boundary in my classification diagram before the paralinguistic division starts. This emphasizes Muller's statement that *"language begins where interjections end* [Muller 1836 quoted in Wharton, 2003: 175] and reflects my attempt to modify it into the following: language begins where interjections end and where onomatopoeic words start. Onomatopoeia is further discussed below.

The second controversial issue about vocal communication is the attempt to classify it into either voluntary communication or involuntary communication. Voluntary communication is that which is intended by the utterer [Srinivasan and Lim, 2000: 3]. Involuntary communication is that which is "*not under the volition of the sender"* but that may include information that the utterer is either aware of or not aware of [Srinivasan and Lim, 2000: 3].

Elements of vocal communication which can be involuntary include hesitation, latency in response, speech rate, pitch, and speech errors or slips of the tongue. Most people, regardless of their age, may generate involuntary linguistic and paralinguistic vocalizations during game play or sports games. These may include imprecations, response cries, or even grunts as in the case of the tennis player Jimmy Connors who grunted loudly every time he hit the ball during the Wimbledon tennis championship in 1981. When asked about his grunting, he claimed that it was involuntary and that he had no control over it [Wharton, 2003: 184].

Some emotive vocalizations such as laughter can also sometimes be involuntary as well as other expressive vocalizations of pleasure that are associated with having a relaxing message, eating a delicious meal, or engaging in a sexual activity [Beeman, 1998]. Interjections are also sometimes considered involuntary because as suggested earlier, they could be the result of the involuntary gestures or facial expressions that they accompany.

Some involuntary vocalizations may be caused by disorders such as Tourette Syndrome, which was mentioned earlier. People who suffer from Tourette

Syndrome, a neurological disorder, may utter involuntary vocalizations called vocal tics [Tourette Syndrome Association, 2005]. These include production of noises such as stuttering, throat clearing, tongue clicking, barking sounds, and cursing [Hartman, 2004]. Extreme and dramatic vocal tics that may be uttered by people with Tourette Syndrome include coprolalia (uttering swear words) or echolalia (repeating vocal utterances of others) [Tourette Syndrome Association, 2005].

To clarify my own thinking, I have roughly mapped the role of volition in relation to the linguistic/paralinguistic continuum.

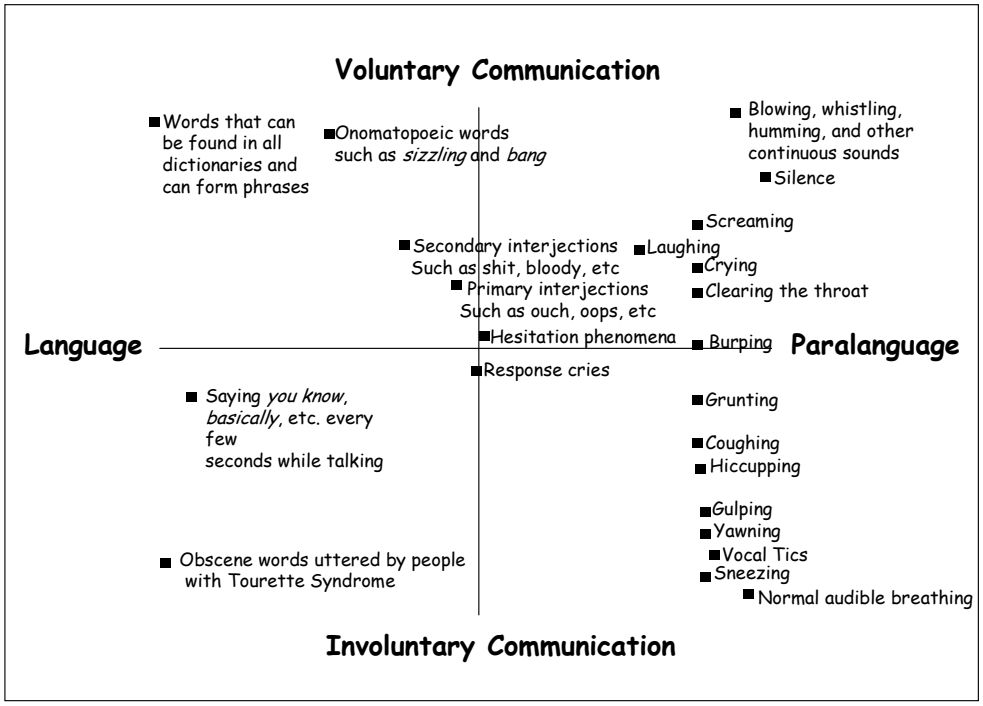


Figure 10: A tentative sketch that I drew to organize my thoughts about the classification of vocalizations based on their linguistic status and the degree of volition involved in producing them.

2.4 Communicative Aspects

Many people generate a variety of vocal imitations of sounds to convey a closer mental image of their descriptions. Some of these imitations are paralinguistic or

in non-speech form (vocal mimesis), and others are linguistic or in word form (onomatopoeia) and of course they overlap. This section will address both forms of communication as well as parent-infant communication and will glance at how implementing these forms of communication in the field of interactive media may have contributed to the development of certain paralinguistic voice-controlled applications.

a. Vocal Mimesis

Charles Darwin believed that human predecessors used to attract each other by uttering rhythmic cries before obtaining the linguistic skills that would allow them to express their feelings in words [Darwin, 1871:880]. Some linguists believe that language evolved from these music-like expressions that early humans made either to attract the opposite sex or to communicate about forthcoming threats or caught preys while hunting [Mithen, 2005:176]. Hunting activities and religious rituals involved mimicking animal sounds and movements to indicate what has been seen while hunting or to warn against a forthcoming danger [Mithen, 2005: 168]

Dr. Jerome Lewis, a researcher at the London School of Economics undertook ethnographic studies of the Mbendjele, Congo-Brazzaville forest hunters. He found that these people are known for vocal mimicry. Instead of describing his story of trying to hunt a gorilla in the forest to his tribe, when an Mbendjelic hunter comes back to the village after a hunting trip, he produces mimicked noises of the gorilla [Lewis, 2002]. To get monkeys to come down of high trees, the Mbendjele mimic monkey calls to deceive them into believing that one of their children has fallen from the tree [Lewis, 2002].

Nowadays, non-speech vocal imitations of airplane, truck, car and animal noises are very common among children. Many vocal imitations are also used as sound effects and voice-overs in almost all cartoons and animations. Percy Edwards, a famous mimic of animals, is known to have mimicked all kinds of animal and bird sounds for cartoons, radio, film, and TV from the 1930s to the 1980s.

Recently Marcus Coates produced Dawn Chorus; a film that features several British singers mimicking birdsong. Coates' aim was to explore the relationship between the human voice and birdsong [BALTIC Centre for Contemporary Art, 2007].

Another vocal imitation pattern that is common between adults as well as children is musical whistling, or the use of whistling as an instrument that reproduces the melody of a certain song. Many people, for instance, whistle the tune of their favorite songs while working, drawing, cooking, or washing. Query by Humming (QBH) applications which are used to retrieve a piece of music and search for a particular song, are based on vocal imitation of the melody of the required song. The user imitates or reproduces part of the tune by humming to a system that compares the hummed input with an audio database of songs, and then returns a list of similar songs (See for example http://www.musipedia.org/whistle.0.html).

QBH applications are not the only example of the convergence between vocal mimesis and interactive media. Kelly Dobson, a researcher at Massachusetts Institute of Technology (MIT), undertook a study of vocal imitation of machine sounds in an attempt to investigate its therapeutic outcomes. She carried out an experiment to detect the similarities between the sound characteristics of various machines (blender, drill, hoover, coffee maker, sewing machine) and vocal imitations of their sounds. She observed that the most imitated characteristics of machine-generated sounds were their *"pitch, roughness, energy, and transients"* [Dobson et al., 2005]. Dobson and others have exploited vocal mimicry in interactive projects that are discussed in the third chapter.

The software company, Elmorex Ltd, developed a *"voice-to melody"* application called *Rring* [Elmorex Ltd, 2000]. This software enables the user to create a personalized ring tone by dialling a certain number and humming or singing into the phone which is used as a microphone in this case. The software then converts the voice signal into a ringing tone of similar acoustic characteristics which the user can download into the mobile phone.

Vocal imitation may also be of great significance to studies in automatic speech recognition. Breidegard and Balkenius developed a speech recognition system that can learn speech sounds by listening to and imitating speech input [Breidegard and Balkenius, 2003]. Studies and experiments on child language acquisition aided the developers in improving the system.

Vocal imitation also plays an important role in language acquisition and in social development. In the 1990's, Giacomo Rizzolatti's and Michael Arbib's discovery of mirror neurons played a key role in explaining imitative behavior in humans, especially the imitation of oral expressions that aid in the acquisition of language [Rizzolatti and Arbib, 1998]. Mirror neurons are also thought by some researchers to be responsible for the observation and execution of hand gestures [Corballis, 2003]. Corballis argued that these mirror neurons show that speech evolved from gesture and not from vocalizations [ibid]. Although many researchers have investigated gesturing as an accompaniment of speech, there are also some intriguing gestural accompaniments of non-speech vocalizations as I observed during the experiment that I discuss in the fourth chapter.

Before discussing more details of infants' acquisition of language in section C of this chapter, the next section explores how vocal mimicry has probably contributed to the development of language. The fact that mimicry was a primary means of communication has caused some researchers to believe that miming animal calls and the sounds of nature is what led to the formation of onomatopoeic words [Mithen, 2005: 169] as will be discussed in the following section.

b. Onomatopoeia

Onomatopoeia is the formation of a word which sounds like the action or object it represents. Examples of onomatopoeia include the words: sizzling, bang, smash, splash, crash, miaw, buzz, hiccup, beep, and ding dong.

Not all onomatopoeic words result from a vocal imitation of the sound of the signified object. Some words are correlated with the physical characteristics of the object they represent. It is thought that some forms of onomatopoeia existed a long time ago when early humans not only imitated the vocal cries of the animals they wanted to refer to but also used utterances made of vowels that represented the physical features of these animals. Berlin's comparative studies of the names used in Peru and those used in Malaysia for fish led him to infer that there is an association between the size of fish and the vowels used in their names; the vowel 'i' was more likely to be present in the names of small fish while the vowels 'e', 'a', 'o', and 'u' were more likely to be used in the names of larger fish [Berlin, 2005].

Jespersen, who proposed the existence of sound synaesthesia in the 1920's, believed that trying to orally represent the shape and size of some objects led to the formation of words that may easily be linked to the objects they represent. He suggested that there is an association between the size of the oral cavity while uttering some words and the size of the physical object represented; the sound 'i' is mostly associated with small objects, while 'u', 'o', and 'a' are associated with bigger objects [Jespersen quoted in Berlin, 2005]. He believed that the reason for these associations is that these sounds are generated by the tongue and lips which mimic the size of the represented object. Jespersen's claim was supported by a simple experiment that the linguist Edward Sapir undertook. He made up two words, 'mil' and 'mal', which differed in their middle vowel. He told the participants that these were the names of tables and asked them which name indicated the larger table and which name indicated the smaller table [Lowrey and Shrum, 2007]. Almost all participants chose 'mil' to refer to the small table and 'mal' to refer to the large table [Lowrey and Shrum, 2007].

Wolfgang Kohler conducted a similar experiment in 1927. He displayed two illustrations in front of his test subjects and asked them which of the two represented the sound 'maluma' and which 'takete' [Levin and Lieberman, 2004:1]. One of the shapes had smooth curves while the other had star-like sharp edges (Figure 11). Almost all test subjects chose 'maluma' to refer to the round

shape and ‘takete’ to refer to the angular shape, indicating the possibility of *“synaesthetic mappings between shape and sound”* [Levin and Lieberman, 2004:1]. One suggested explanation is that the round shape was similar to the sound ‘maluma’ in its lack of *“abrupt discontinuities”* and that the angular shape was similar to the sound ‘takete’ in its *“rapid changes”* [de Götzen, 2004].

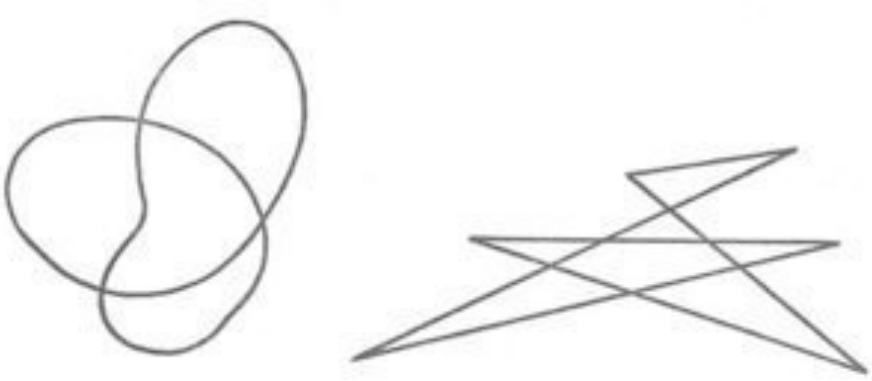


Figure 11: Visual representations of the sound ‘Maluma’ and ‘Takete’ by Wolfgang Kohler. [copied from Levin and Lieberman, 2004]

The role of onomatopoeia in the development of language is still the subject of debate. Some commentators agree that onomatopoeia has significantly contributed to the development of many linguistic utterances [Limber, 1982] [Allott, 1989]. With this in mind, onomatopoeic words of a foreign language might be easier to understand than other more complicated words that do not have a strong relationship between their sounds and meanings. A more common form of communication that is claimed to be universal, however, is infant-directed communication or parentese which will be explored in the next section. Some researchers suggest that the use of infant-directed communication for training speech recognition systems might enhance their automatic learning and discrimination of phonetic categories [Kirchhoff and Schimmel, 2005]. Some developers have used infant-directed speech to train robots that can recognize approval, disapproval, praise, or prohibition [Breazeal and Aryananda, 2002]. Because of the potential role of the exaggerated prosody that adults generally direct towards infants in developing and training systems that can recognize affective intent, it is important to cast a short glance on this communicative aspect of voice.

c. Parentese

Parentese, motherese, fatherese, baby-talk, or IDS (infant-directed speech) are all terms used to refer to the music-like style of speech used by parents to communicate with their babies. Almost every person who talks to a baby uses a high-pitched voice that is characterized by a lengthened melodic quality and repetition. Anne Fernald, a psychologist in Stanford, conducted studies about mother-infant interaction across different cultures. She found that the long pauses and word stretches that characterize infant-directed speech are *"functions that seem to cross national boundaries"* [Fernald, 1992];

> *"mothers of all nations address their babies with universal melodies: short, sharp staccato for warning ('Nein! Nein!'); rising and then falling pitch for praise ('BRA- vo!'); a long, smooth, low frequency for comfort ('Oooh, pobrecito!'), and a high, rising melody for calling attention to objects ('Where's the buzz-a-BEE?')."* [ibid].

Thus Fernald finds that parentese involves paralinguistic features that are common among almost all cultures. These features include high pitch, slow tempo, wide pitch range, extended intonation and exaggerated inflection.

Some studies have shown that infants are more sensitive and responsive to parentese than to normal speech, and that they are much more reactive to vocal expressions than to facial expressions [Mithen, 2005: 70]. Infant-directed speech is thought to be more effective than stroking in calming infants [Mithen, 2005: 70]. Other studies of IDS found that infants are generally more responsive to singing than to speaking and that they spend more time watching audio-visual recordings of their mothers while they sing than while they talk [Trehub and Nakata, 2002].

Dunbar thinks that IDS was formed as a way to attend to the babies when physically distant from them in what he refers to as *"vocal grooming"* [1993]. He suggests that physical grooming evolved into *"vocal grooming"* as a result of early human ancestors' inability to groom each other because of their increase in

group size and the amount of time that grooming would require at the expense of the time dedicated for food search [Dunbar, 1993] .

Because Infant-directed speech is highly affective, spontaneous, and universal, studies on the prosodic features of IDS may augment the development of computers that recognize and respond to human emotional expressions as will be explained in the third chapter. Malcolm Slaney and Gerald McRoberts, for instance, developed *Baby Ears*; a recognition system for affective vocalizations that is based on experiments that involved gathering acoustic data from parents talking to their infants [Slaney and McRoberts, 1998]. The exaggerated prosodic features of IDS *"are providing cues to phonetic prototypes that may be useful for both machine and human learning"* [Ostendorf et al., 2003].

The prosodic character of infant-directed speech is thought to play a role in facilitating the acquisition of language. The placement of pauses helps the infant perceive the beginning and end of sentences and to split and identify individual words and phrases [Nelson et al., 1989].

Infant-directed speech not only plays a major role in language acquisition, but also has significant psychological and physiological effects on the infant's growth by facilitating feeding and sleeping. Experiments carried out by Jayne Standley showed that singing lullabies significantly improved infants' sucking abilities and resulted in their weight gain [Standley, 2000].

Many other studies about the psychological effects of singing on infants have been undertaken, and many also have been conducted about the general effects of singing and music on adult listeners. However, there is hardly any research about the psychological and physiological effects of singing on the singer. The very few studies that I found related to this subject will be explored in a later chapter.

While parentese is considered cross-cultural by some researchers, the next section explores paralinguistic communication which varies by culture.

2.5 Cultural Aspects

In most, if not all, cultures in the world, the verbal content of speech is not the only form of communication. A large part of communication involves non-verbal forms of communication which may be vocal, facial or gestural. They help listeners recognize the emotional state of the communicator and the content and connotation of speech. Many remarkable uses of vocal paralanguage exist within various cultures for different communication purposes.

In Arab countries, the zaghrooda (ululation) is a high-pitched utterance generated by moving the tongue between the lips repetitively to express joy during happy occasions. To direct the sound of the zaghrooda, a hand is usually shaped like a U and placed around the lips.

In La Gomera, one of the Canary Islands in Spain, El Silbo is used as a means to communicate between villages. El Silbo is a whistling language that is based on four consonant sounds (ch, y, g, k) and two vowels (a, i) that allow the whistler to form around 4000 different words. The language has a whistled equivalent for each phoneme used in the local spoken language. The stability of the pitch range determines the equivalent vowels while pitch transitions between these vowels determine the equivalent consonants. The duration of the whistle and its pitch level convey the stress of the whistled speech.

Figure 12: La Gomera's inhabitants using El Silbo to communicate with each other [Useher, 2004].

This language is used by the inhabitants of the island to communicate across its hills and valleys. In order to whistle, one hand's index finger is placed in the mouth to produce pitch variations and the other hand is shaped like a U to direct the sound (Figure 12).

El Silbo inspired Bohlen and Rinker [2004] to develop the *Universal Whistling Machine*, a device that whistles when its built-in video camera tracks the presence of a person in its surrounding area (Figure 13). Once a passer-by leaves the field of view of the camera, the device starts whistling to seek that person's attention. When the passer-by whistles, a noise-reducing microphone connected to the device captures the sound. An FFT- based (Fast Fourier Transform) pitch tracker and signal sampling [Bohlen and Rinker, 2004] are used to analyze the input signal and to exclude non-whistled signals. In order to imitate input whistles, the device then generates synthetic whistles which are similar to yet different from the input signal in some acoustic aspects. These aspects may include transformations in pitch or tempo, or even reversing the input signal.

Figure 13: *The Universal Whistling Machine* developed by Marc Bohlen and JT Rinker [Technovelgy.com, 2004] [Fruhlinger, 2004].

Whistling has also been effectively used as a cross-cultural form of input by Adam Sporka and Sri Kurniawan. They implemented whistling as an alternative controller of a mouse pointer in what they refer to as a "*Whistling User Interface*" [Sporka et al., 2005]. This interface allows the user to whistle in order to control the cursor orthogonally; either horizontally or vertically (Figure 14).

Figure 14: A *Whistling User Interface* that allows for whistling to control a mouse pointer [Sporka et al., 2005].

When the initial pitch of the whistle is lower than a specified threshold, the pointer moves horizontally. When the initial pitch is higher than the specified threshold, the pointer moves vertically. Shifting the pitch from low to high moves the pointer to the right, and shifting it from high to low moves it to the left. A short whistle causes the pointer to click.

Another application that employs whistling is *WhisRaider*; a pocket PC game in which the pitch of the user's whistle steers the spaceship horizontally and also allows the player to shoot space invaders. The higher the pitch of the whistle, the more to the right the spaceship moves. *WhisRaider* is available at www.phonature.com.

There are many other forms of non-speech sounds and vocalizations that are specific to certain cultures but which, if exploited effectively in the field of interactive media, might be used as cross-cultural means of human-computer

interaction. Before moving into a systematic discussion of several other voice-controlled applications, the next section investigates voice signal processing techniques and presents some of the most significant voice-related technologies.

2.6 Technological Aspects

a. Voice Recording

In 1877, Edison built the first machine that recorded and reproduced the human voice. In 1887, Berliner's development of the gramophone as an improved version of the phonograph led to the development of the modern phonograph. Since then, the advent of voice recorders paved the way for many creative implementations and inventions.

Many sound poets exploited the development of voice recording techniques. One of these poets was Dufrêne who recorded *"cri-rhythmes"*. These were poems that consisted of a wide variety of vocalizations, cries, shrieks, and ululations. In 1898, Poulsen invented magnetic recording. Later, German scientists developed magnetic recording into a technique that involves tape cassettes. Tape recorders allowed for editing, cutting, arranging, rearranging, speeding up, slowing down, and adding effects to the vocal piece after the real-time performance [McCaffery, 1978]. Recording vocal effects such as canned laughter, cries, and other noises and reusing them would not have been possible without the advent of voice recording.

Nowadays, digital recording has overtaken analog recording and made possible perfect copying of recordings. Digital technology has made it easier to edit, analyze, duplicate, loop, and perform various processing algorithms on the recorded voice. It has also made it possible to filter, manipulate and add various effects to the live voice input into a computer's microphone.

c. Voice Analysis

The invention of the microphone by Emile Berliner in 1877 and the development of its commercial form later on by Thomas Edison was crucial to the development of various voice analysis mechanisms. Among these mechanisms is the spectrograph which captures sound waves and generates a spectrogram, a graphical representation of the sound signal that plots the frequency against time. The intensity at a particular time is represented by brightness in a colored spectrogram or darkness in a grey-scale spectrogram. The spectrograph is thus a visual output derived from a sound input. Artistic applications of this principle are discussed later in the third chapter.

One of the methods of calculating spectrograms from the voice signal involves the use of a Fourier Transform; an algorithm named after the French mathematician Jean Baptiste Joseph Fourier. A Fourier transform converts the signal from a time domain into a frequency domain displayed in terms of sinusoidal functions. It breaks up a complex signal into individual frequency components and their amplitudes. By using this conversion a variety of digital signal processing techniques can be applied to achieve different purposes. Unwanted frequency components, for example, can then be eliminated in order to decrease hissing or background noise. The signal can be filtered by cutting off higher or lower frequencies. Many variants of the Fourier Transform exist including Fast Fourier Transform, Short-Time Fourier Transform, Continuous Fourier Transform, and Discrete Fourier Transform. Some of these algorithms, especially the Discrete and Fast Fourier Transforms play a major role in voice signal processing and the recognition, manipulation, synthesis, and compression of voice. The Fast Fourier Transform is the main technique underlying all my voice-controlled projects, which are presented in the third chapter.

c. Voice Recognition

It is important to highlight the difference between voice recognition and speech recognition. Although these two terms are often used synonymously, it is best to distinguish their meaning. Voice recognition is the computer's ability to analyze the voice signal in order to identify a person from another. Speech recognition, which is further discussed in the next section, is the computer's ability to analyze the speech signal in order to identify a word from another. Voice recognition software is biometric and is usually used to verify the voice of a certain speaker and prevent unauthenticated use of a system. Speech recognition software, which can usually be trained, can be used either by a certain speaker (speaker-dependent) or by any speaker (speaker-independent).

Voice recognition systems are also referred to as voice authentication or voice verification systems. Often used for security purposes, they extract the acoustic parameters of the speaker's voice or speech. They then compare the speaker's 'voice print' with one that has been previously enrolled and stored. Such systems have recently been improved to the extent that they cannot be tricked by the use of a pre-recorded response by the speaker in what is commonly referred to as spoofing. They have been equipped with anti-spoofing features that involve generating a number of random phrases that the speaker is required to say in order to be verified.

d. Speech Recognition

Speech recognition is the conversion of a speech signal captured by a microphone into words, which then potentially allow meaning to be inferred by a computer. Speech recognition programs are sometimes trained to match the voice input to a digitized voice sample and are programmed to then execute a certain command. Voice signals can also be processed using Fourier Transform algorithms which match the characteristics of the voice-input with the characteristics of the previously stored samples. Applications for speech

recognition include information retrieval systems, robotics, computer-based speech commands, speech dialling, and telephony.

Several factors affect the quality of recognition. These include background noise, the type of microphone used, and the type of software used to convert the voice signal into computer commands. The difficulty in speech recognition lies in the difference in the vocal characteristics, speaking rate, and the variety of pronunciations produced by different people. Other recognition difficulties may arise from changes in the type of microphone used and the relative position of the microphone to the speaker and to ambient noise.

Speech recognition systems are differentiated based on a number of features including: speaking mode, enrolment, and vocabulary. The speaking mode is determined by whether the system recognizes continuous speech or discrete speech. Enrolment is determined by whether the system is speaker-dependent or speaker-independent. A speaker dependent application will only recognize a certain speaker's speech after s/he enrols by providing some speech samples. A speaker-independent application will recognize any speaker's speech signal without enrolment. The vocabulary list also differs from one system to another depending on the number of words the system is programmed to recognize. The larger the vocabulary, the more the chances of it containing similar-sounding words, and the greater the probability of confusion.

Existing computer-based speech recognition engines and dictation software include *Via Voice* by IBM, *Dragon Naturally Speaking* by ScanSoft, and *Microsoft Speech Application Programming Interface* (SAPI). Available open source speech recognition engines include *Sphinx* by The Sphinx Group at Carnegie Mellon University and *Julius* at Kyoto University and sponsored by IPA (Information Technology Promotion Industry) [Kawahara et al., 2000]. Many other speech-related processing technologies exist, including speech synthesis, speech compression and decompression, and non-verbal speech recognition. Speech synthesis, which is sometimes called text-to-speech (TTS), is the inverse of speech recognition and involves artificial generation of speech.

Speech compression and decompression are important in the transmission and storage of the speech signal for telecommunications purposes.

Non-verbal speech recognition, on the other hand, involves the recognition of non-verbal speech features. Quast [2002] is exploring programs that can measure and recognize nonverbal content. His projects aim to train a pattern recognition system to recognize *"non-verbal speech"* by which he means the non-verbal information in the speech signal. According to Quast, word recognition, which is the conventional method of speech recognition, has been investigated for a while now and progress in improving the accuracy of this technique is slow. He suggests the employment of multiple communication channels in speech recognition, such as paralinguistic voice as well as speech.

e. Recognition of Non-Speech Aspects of Voice

There have been some efforts to incorporate more forms of non-verbal communication in user interfaces in order to make them more human-oriented. Nevertheless, research in affective computing and computer-based recognition of emotions through vocal expressions is still in its infancy. There have recently been some attempts to develop systems which recognize and respond to human emotional states. Most of these attempts use facial expressions as means of emotion recognition. However, there are some studies which indicate that the human is better able to decode emotions from the voice than from the face [Kappas, 2003].

Although voice might be considered the "*neglected child"* of emotion research [Kappas, 2003], there is a growing body of research aimed at exploring the acoustic features that vary with the emotional state in human voice. These studies examine the relationship between the characteristics of vocal expression and emotional state. The following table (Table 3) illustrates the characteristics of voice that Xiaoqing believes generally to differentiate between the expression of anger, joy, and sadness [2003].

Table 3: The emotional states inferred from various voice characteristics [Xiaoqing, 2003].

Feeling	Loudness	Pitch	Timbre	Rate	Enunciation
Anger	Loud	High	Blaring	Fast	clipped
Joy	Loud	High	Moderately Blaring	Fast	Somewhat Clipped
Sadness	Soft	Low	Resonant	Slow	Slurred

The table indicates the possible relationship between vocal characteristics and emotional state of the utterer. These characteristics can be detected and analyzed separately using various signal analysis mechanisms. However, the different emotive vocalizations that these parameters *collectively* convey are not easily distinguished by acoustic feature analysis. According to studies by HP Laboratories, anger and happiness have similar prosodic features and can often be confused with one another [Yacoub et al., 2003]. A study by Klaus Scherer revealed that the fundamental frequency (F0) range had the most powerful effect on emotion recognition [Scherer, 1996]. Narrow F0 indicated sadness, while wide F0 indicated high arousal and strong negative emotions such as annoyance and anger [Scherer, 1996]. The study showed that these effects are continuous and that the relationship between the size of F0 range and the strength of emotion expression is linear. High intensity was related to aggressive affects. Long-durational voice segments (slow tempo) were associated with sadness, while short voice segments were linked to joy [Scherer, 1996].

A group at MIT Media lab are developing and exploring new application areas of affective computing [Picard, 2000]. Madan, for instance, is developing the *Jerk-O-Meter,* an application that runs on a mobile phone or via a voice over internet protocol (VOIP) application [Madan, undated]. The system analyzes paralinguistic features of speech in real-time and gives feedback on the user's interactions with the mobile. For example, when the speaker is speaking loudly

and annoying surrounding people, it sends a text message through the mobile to inform the speaker that s/he is being a *"jerk"* [ibid].

The selection of the appropriate acoustic features that should be extracted from a voice signal and the selection of the processing algorithm have a major role in improving the mechanisms of affect recognition. Scherer [1996] points out that there should be more interchange between physiologically and acoustically oriented voice scientists and psychologists studying vocal emotion expression. The selection and definition of the acoustic parameters that are directly pertinent to affective states is still in its early stages [Scherer, 1996]. Moreover, the development of high-quality emotional synthesis requires a complete model of the emotion process, including effects on the nervous system and their accompanying acoustic changes. Of course, such a model does not yet exist [Johnstone, 1996]. A highly accurate and consistent method of recognising, analyzing, and evaluating emotional states in users' emotive vocalizations is still one of the major objectives in human-computer interaction.

Although a lot of affective computing-related research focuses on the recognition of emotive aspects, there are other forms of non-speech vocalizations which also deserve more attention. These include vocal segregates such as 'ahhh' and 'mmm' and interjections such as 'oh' and 'ouch'. They could be easier than emotive vocalizations to identify because they are closer to language than emotive vocalizations (Figure 10). In other words, some of these segregates can more or less be written and uttered in a word-like structure. They can therefore be recognized, to a certain extent, by speech recognition software despite the fact that they are arguably considered non-speech utterances.

To investigate this theory, I aim to develop *TOT* (Tip of the Tongue) as an attempt to solve the 'tip of the tongue' experience. A person undergoing this experience tries to remember a certain word while feeling that it is right at the tip of his/her tongue. Most of the times, the person manages to retrieve some letters of the word, the way it sounds, or its mental representation. It is claimed that people can remember the first letter of the forgotten word around 50 percent to

70 percent of the time, and they can also often remember the number of syllables that the forgotten word contains [Herr, 2002].

For this irritating reason and in order to investigate the possibility of using non-speech recognition as a complement to speech recognition, I propose the idea of *TOT*. Unlike the *Jerk-O-Meter* which gives its user feedback on his or her interactions during phone conversations, *TOT* gives the user feedback and assistance during his/her face to face interactions. It is designed to be integrated into a portable pocket device, such as a PDA, that mainly consists of a touch screen with an onscreen keyboard as well as a headset/microphone.

TOT listens to its user during a conversation. When it recognizes vocal segregates such as 'mmm' and 'ahh', it assumes that there is a forgotten word at the 'tip of its user's tongue'. It then listens to the user's attempts to remember similar words and plays an audible list of related words through the user's headphones while also displaying them on the screen. When the user utters 'shhh', the audio playback is stopped. A more advanced version of the device might be connected to a glove that reminds the user of the name of the person s/he shakes hands with. When the user utters the interjection 'Oh', the system assumes that the user has forgotten the name of someone s/he is just about to shake hands with. It immediately scans the other person's thumb through the gloves, compares the fingerprint with a previously saved one, and reminds the user of the name. The integration of non speech and speech recognition in such a device might be useful to people suffering from nominal aphasia. This exploitation of non-speech recognition of vocal segregates and interjections might also be useful for complementing the user's memory and augmenting his/her social communication skills. After all, several elements of voice such as *"rhythm, repetition, antithesis, alliteration and assonance, epigrams, proverbs, and formulaic expressions"* have already been exploited as mnemonic techniques [Karpf, 2006; 199].

f. Electronic Manipulation of Voice

Many voice-manipulation programs exist today. Some of them are as simple as toys, while others are much more complicated installations. One voice-changing toy is *Hasbro Darth Vader Voice Changer* which was developed in 2004. This toy allows children to change their voice into one similar to the character in the Star Wars movies.

There are also telephone voice modifiers that are designed to be used as disguisers. These may change the apparent gender and age of the speaker while talking over the phone. Some people use them for entertaining purposes while others use them seriously, for example to pretend that they are someone else when they get a call that they are too busy to respond to. Some voice changers, such as *The Pretender Voice Changer*, have extra features such as allowing the user to play the voice of a crying baby, a barking dog, or even a ringing bell in order to end a phone conversation [Milestone International, 2006].

Nowadays, there are even more advanced electronic forms of voice manipulation. A unique and humorous interactive installation that employs this technique is *Headspin* [Someth;ng (sic), 2005]. This installation involves a washing machine that users are prompted to set to spin, put their heads in, and shout (Figure 15). It allows users to select from a variety of washing programs, and depending on the washing program and the cycle speed selected, the user's voice characteristics are manipulated and the visual effects are displayed.

Figure 15: *Headspin*; a voice-manipulative washing machine [Someth;ng (sic), 2005].

Voice not only can be manipulated by electronic tools, but can itself manipulate various interactive modalities including acoustic, visual, and physical computer-based media as will be explored in the next chapter. In the rest of this thesis I will seek to show that there are untapped elements of voice that may play a useful role in the field of interactive media.

I outlined in this chapter the elements that have been investigated and/or applied in various fields of knowledge. Voice was placed within the context of relevant physiological, artistic-performative, paralinguistic, communicative, cultural, and technological issues and topics.

Its physiological aspects were highlighted mainly to investigate the factors that may underlie or affect the delivery and control of voice which is exploited as an input in the work that will be discussed in the rest of this thesis. Its artistic-performative aspects were described generally to facilitate a later discussion about the role of expressive paralinguistic voice input in transforming users into performers. Its paralinguistic aspects were studied in order to provide an overview of the range of sounds that are considered paralinguistic and that can be voluntarily controlled and hence allow for volitional paralinguistic control of interactive media. Its communicative aspects were outlined in order to highlight the significance of vocal forms of communication such as vocal mimesis, onomatopoeia, and parent-infant communication in augmenting studies of voice-related technologies. Its cultural aspects were reviewed in an attempt to explore the vocalizations that are specific to certain cultures but that, if exploited in the field of interactive media, can be used as cross-cultural means of human-computer interaction. Finally, its technological aspects and advancements were traced in an attempt to set the wider context for some of the relevant technological aspects that are involved in the development of the projects discussed next.

3. Paralinguistic Vocal Control of Interactive Media

3. Paralinguistic Vocal Control of Interactive Media

Throughout the earlier parts of this thesis, I have repeatedly touched on the potential use of the paralinguistic aspects of voice as input to interactive systems. I now confront directly the possible applications of paralinguistic vocal input. In particular, I underline the pragmatic approaches that developers, including myself, have taken and that may be used to inform any attempt to develop non-speech voice-controlled systems.

The investigation of new input mechanisms is a widespread endeavor in the field of human-computer interaction. Many alternative input mechanisms are being developed in an attempt to augment expressivity, engagement, immersion and naturalness, and to enrich the diversity of interactions. Mechanisms include 3-D locators, gesture trackers, haptic input devices, foot-operated devices, eye trackers, and speech recognition. As noted earlier, speech recognition is often confined to the analysis of the verbal aspects of voice.

Finding practical implementations of paralinguistic vocal control may permit users to use a wider variety of expressive hands-free input techniques and enrich the vocabulary of interaction. Moreover, this technique may be used for therapeutic purposes by asthmatic and vocally-impaired users, and it may possibly be used as a training tool by vocalists and singers. When used by players as an input mechanism, vocal expressions – as I will show – can be an entertaining output to an audience and may act as engaging sound effects in voice-controlled interactive performances.

The purpose of this chapter is to suggest ways in which vocal paralanguage can be usefully exploited in interactive media, and to explore characteristics of voice that can be employed in controlling interactive applications. The chapter discusses existing audio-visual and voice-visual systems and suggests new forms of artwork aimed at the use of paralinguistic vocalizations for expressive interaction. It also discusses existing audio-physical systems and proposes their

expansion into *voice-physical* artwork aimed at the use of paralinguistic vocalizations to physically control real objects.

First, it may be useful to note the small amount of research which has been done in relation to users' paralinguistic vocal responses to interactive media experiences, since clearly such responses are capable of being harnessed as input.

Tom Johnstone and his colleagues are currently investigating the use of computer games to elicit a range of emotions, which are studied in terms of their subjective, physiological, and expressive manifestations [Johnstone, 1996] [Johnstone et al., 2005]. They are specifically interested in the study of the psychological and physiological mechanisms, which cause a range of changes to the acoustic speech signal during game-play. They conducted an experiment that involved analyzing thirty-three participants' speech characteristics and measuring their physiological reactions while they played a spaceship game called *XQuest*. They reported that the fundamental frequency and the energy of speech became higher following *"obstructive events"* such as the destruction of the ship and lower following *"conducive events"* such as the completion of a game level [Johnstone et al., 2005:516]. They suggested that the reason could be that physiological arousal, as indicated by measurements of skin conductance, was comparatively high following the destruction of the ship.

Vocal expression of emotions elicited during human-computer interaction may be better explored if complemented by the analysis of non-verbal vocal expressions. The investigation of these paralinguistic expressions presents a number of benefits. Because these expressions are not linguistic, they can potentially be used to compare vocal responses to interactive media across different cultures [Belin et al., 2005]. Moreover, they are *"primitive expressions of emotions"* [ibid] and are probably better conveyers of the natural emotions and instantaneous responses that usually result from the frustration, stress, anxiety, or excitement of interacting with interactive systems.

The exploitation of emotional responses in reaction to interactive media may enable computer users to select a mode within the interface, in which the computer becomes sensitive to the user's emotions as reflected in non-verbal utterances. This concept does not seem to be far away from already being implemented. Computer scientists at Monash University in Australia have already developed a system, *SoundHunters*, which allows the user to log on to a computer by laughing. *SoundHunters* logs a user onto networked computers through detecting the sound of the laughing user's footsteps in order to determine his/her location and activate the nearest computer [Nowak, 2003].

Further work on paralinguistic vocal *responses* to interactive media is needed in order to improve and enrich the studies on paralinguistic vocal *input* to interactive media that will be investigated in the rest of this chapter. According to Johnstone, computer games may induce *"stronger"* and *"more varied emotions"* than other induction techniques (such as those listed in [Scherer, 1986]) [Johnstone, 1996]. They allow for a degree of immersion that would result in the elicitation of strong and natural vocal reactions [Johnstone, 1996]. The analysis of the characteristics of these expressions may facilitate the development of computers that recognize and react to these vocalizations and their characteristics. This is explored next.

3.1 Voice-Visualization

a. Audio-Visual Applications

A significant component of my own projects is visual outcomes which correspond in some way to vocal input. This section of the thesis therefore explores mappings between visuals and sound. It first provides a brief outline of historically-related work and then describes significant contemporary related studies and applications. It then explores the various audio-visual composites used in existing multimedia works.

When Newton analyzed the color properties of sunlight in the 17th century, he believed that there was a natural synergy with music [Allchin, 2002]. This led him to assign the colors of the spectrum to the seven notes of the musical scale [Collopy, 2000: 357] by way of what is commonly known as the 'color music wheel'. Since then, there have been many attempts to forge a link between the musical and visual domains. Others who came after Newton used his work as a basis for inventing 'color organs'. These musical instruments were built to display modulated colored light in some kind of fluid fashion allegedly comparable to music [Moritz, 1997]. In addition, other methods of visualizing sound were devised for scientific purposes. These included oscilloscopes, spectrometers, and strobe lights. According to Roads:

> *"One method involved modulating a bunsen burner with sound and observing the effect on the flames…Rudolf Koenig built precision instruments for generating sound images that he called manometric flames […] More direct images of sound waveforms appeared in the mid-nineteenth century. The Wheatstone Kaleidaphone (1827) projected vibrating motions onto a screen"* [Roads, 1996:500].

While earlier work in audio-visualization was principally scientific in its motivation (see for example the discussion about spectrograms in section 2.6), lately there has been significant research interest in audio-visualization for entertainment and performance purposes. Several developers have attempted to develop audio-visual applications and performances that translate musical cues into graphical feedback or vice versa. Some of them employ novel mapping techniques between visual properties and sound characteristics. Some focus on voice-visualization to establish meaningful correlations between vocal input and visual output.

Sound visualization involves developing mappings between graphical parameters (color, height, position, size, shape, etc.) and sound characteristics (pitch, volume, timbre, duration, etc.). One basic example of audio-visual applications is *Apple iTunes* that includes the display of visual effects associated with various characteristics of the music. Other similar applications include *Windows Media Player*, *Sonique*, and *WinAmp*.

An interesting issue is the basis on which visual output parameters are mapped to auditory characteristics. Lipscomb and his colleagues' enquiries among software developers revealed that the designer's selection of the visual parameters tended to be arbitrary [Lipscomb and Kim, 2004:72]. They conducted an experimental multimedia investigation to find meaningful visual *"correlates"* for auditory parameters [ibid]. They assumed that there would be *"preferred"* audio-visual combinations. They found that pitch was perceptually matched with vertical location, loudness with size, and timbre with shape. Both pitch and loudness were also found to be best matched with color, while duration did not particularly match any of the characteristics [Lipscomb and Kim, 2004:73].

These findings should be treated with caution because cultural, environmental, and age differences may affect the perception of audio-visual composites. Moreover, such mappings may also depend on the kind of application in which they are implemented and on the level of interactivity required. Programs such as *iTunes*, for instance, do not require a high level of interaction and the visual output in these programs is mainly incorporated for decorative purposes. In such simple audio-visual applications, the developer may exploit the inherent tendency of the viewer's brain to coordinate the two sense modalities [Bregman, 1990]. On the other hand, in applications in which the user is not just a passive viewer and/or listener but an active interacting agent, the audio-visual mappings determine the way the user interacts with the interface. In the voice-visual applications that will be investigated in the next section, for instance, forming *"perceptually meaningful"* [Lipscomb and Kim, 2004] mappings is crucial.

d. Voice-Visual Applications

Quite a few developers have focused on voice-visualization 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 round a table and use their voices to control visual data displayed through data glasses (Figure 16).

Figure 16: *Hidden Worlds*; an audio-visual installation by Levin and Lieberman [2002].

Voices are transformed into *"worm-like figures"* which are visible in three-dimensional form to the players wearing goggles, and their two-dimensional shadows that are projected on the table are visible to everyone. The position of each figure is determined by the position of each user. The duration of the figure's appearance is mapped to the duration of the vocalization, while the figure's diameter is mapped to volume. Pitch determines the figure's *"flocking behaviour"*.

Igarashi and Hughes also investigated novel voice-visual mapping techniques. They mapped the speed of a scroll bar to the pitch of a user's voice so that the higher the pitch, the faster the speed of scrolling. They also applied this technique to a "*speed dependent automatic zooming interface*" [Igarashi and Hughes, 2001:156] as well as to voice-controlled television settings. They

developed an interface that explored the complementary use of speech and non-speech vocalizations where the viewer utters *"volume up"* to specify the feature to be controlled, and then utters *"ahhh"* continuously to control the volume level. The longer the duration of the *"ahhh"* utterance, the louder the volume.

In The University of Washington, Susumu Harada and his colleagues developed *The Vocal Joystick*, a system that allows users with motor impairments to use their voice characteristics and vowel sounds (Figure 17) to control an on-screen mouse pointer [Harada et al., 2006]. Loudness is used to control the speed of the cursor while different vowels are used to control its direction ([ae] for up, [uw] for down, [aa] for right, and [iy] for left [International Phonetic Alphabet Symbols]).

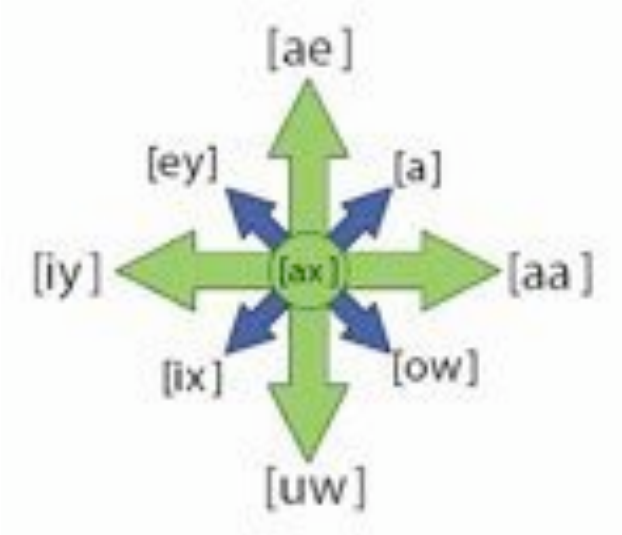


Figure 17: *Vocal Joystick*; a system that allows for the control of the direction of a mouse pointer by uttering different vowel sounds [Bilmes et al.,2005] .

The system was implemented in a game that allows the user to vocally control a virtual fish and direct it towards targets. It was also implemented to control a simulated robotic arm [Malkin and Brandi, 2007], to browse web pages, to navigate maps, and to draw. Although the use of vowels as an input is a novel technique, it introduces cognition-related latency due to the difficulty of remembering the vowel that corresponds to each direction. It seems that the developers have tried to simplify this task by putting the vowels on a kind of perceptual spectrum.

Paralinguistic voice input has not only been used in practical applications but it has also been implemented in entertaining and educational games. Pertuu Hämäläinen developed the *Hedgehog game* which helps players learn how to control the pitch of their voices while using it to control a hedgehog [Hämäläinen et al., 2004].

Figure 18: The pitch of the player's voice controls the vertical position of the hedgehog and maintains its path on the melody-controlled pathway of *The Hedgehog* game [Hämäläinen et al., 2004].

A melody with which the player is encouraged to sing along in synchrony is played. The path that the hedgehog must follow twists according to the pitch of the played-back melody (Figure 18). The vertical position of the hedgehog is determined by the pitch of the player's voice; if it matches that of the melody, the hedgehog remains on the right path.

Fiesta Connexxion is an online voice-controlled racing game for the Ford Motor Company [2006]. The game allows the player to drive a virtual car by vocally

imitating the sound of a car engine. The noteworthy feature of the game is that it maps a *"voice meter"* to the player's voice (Figure 19) [Ford Motor Company, 2006].

Figure 19: *Fiesta Connexxion*; a voice-controlled racing game promoting Ford Motor Company [Ford Motor Company, 2006]

Another voice-controlled game is the collaborative multiplayer game *Organum* (Figure 20). 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, y and z 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.

Figure 20: *Organum*; a collaborative voice-controlled multiplayer game [Niemeyer et al, 2005].

Like *Organum*, *SpitSplat* maps players' vocal characteristics to movement along axes. *SpitSplat* is a voice-controlled game that I developed with two other postgraduates at the Lansdown Centre for Electronic Arts (Appendix B). It was my first attempt to investigate the use of vocal paralanguage as control interface. Voice is used to 'splat' the appropriate color onto several moving targets (Figure 21). Players aim the 'color splat' at moving targets by altering certain qualities of their vocal expression including pitch, loudness and duration. The position of the voice-controlled splat along the x-axis is mapped to the pitch of the player's voice, and its position along the y-axis is mapped to the loudness of the player's voice. So, the louder the voice, the further up the splat moves, and the higher the pitch, the further to the right it moves. Thus, a high-pitched and loud vocalization directs the 'color splat' towards the top right corner of the stage while a low-pitched and soft vocalization directs it towards the bottom left corner (Figure 22).

Figure 21: A screenshot of *SpitSplat*; a game controlled by the pitch and amplitude of the player's voice.

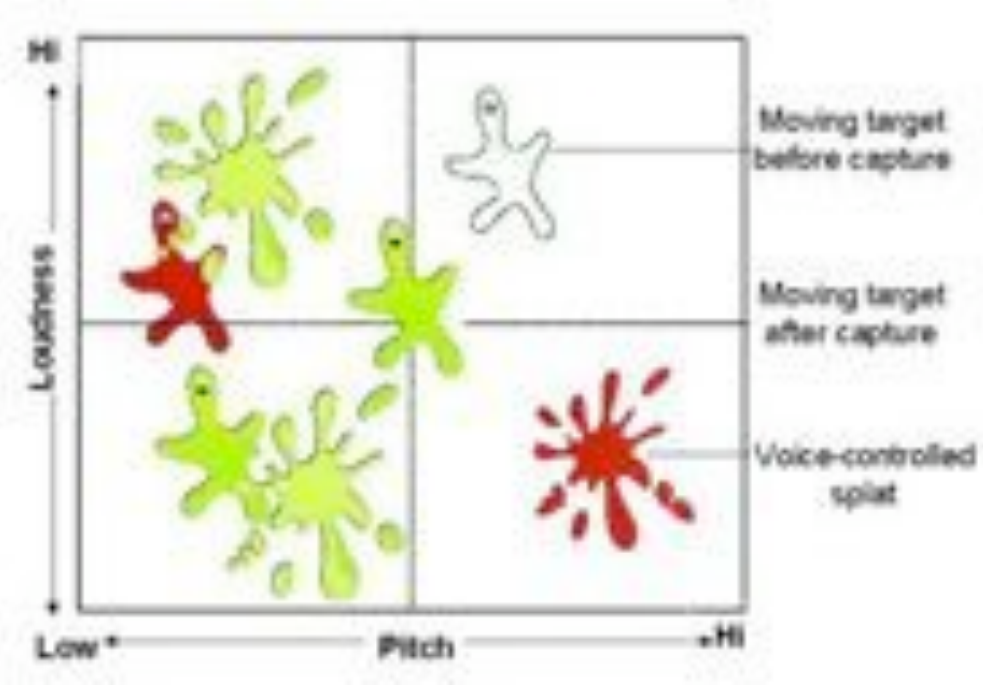


Figure 22: Pitch controls the horizontal movement of the 'splat' while Amplitude controls its vertical movement.

Prior to studying the issue of mappings in greater depth for the present Ph.D., I assumed that these mappings would be *"perceptually meaningful"* and *"readily apparent"* because on a piano keyboard, high pitches are on the right side, while low pitches are on the left side [Lipscomb and Kim, 2004]. Moreover, in many computers, the volume control slider is controlled by moving the handle up to make the sound louder or down to make it softer.

Finally, an influential artwork, *Messa Di Voce*, uses a novel technique that makes speech visually appear to emerge from the speaker's mouth [Levin and Lieberman, 2003]. In one of its sections, *Pitch Paint*, performers use the pitch of their voices to paint on a projected screen (Figure 23). In another, *Jaap's Solo*, the vocalist Jaap Blonk uses his voice to emit bubbles which appear on the screen. A computer vision technique allows him to use his shadow to interact with the bubbles and move them away from him (Figure 24).

Figure 23: *Pitch Paint* from *Messa Di Voce*; an audio-visual installation by Levin and Lieberman [2003].

Figure 24: *Jaap's Solo* from *Messa Di Voce*; an audio-visual installation by Levin and Lieberman [2003].

The voice-visual mappings implemented in *Messa Di Voce* were influenced by Wolfgang Kohler's photosynthetic experiments discussed in chapter 2. The results of these experiments led the developers to infer that there are possible synaesthetic mappings between shape and sound [Levin and Lieberman, 2004] which they explored in *Messa Di Voce*. This artwork demonstrates how such voice-visual performances may potentially introduce a dramatic approach to multimedia applications, and encourage expressive vocal improvisations. This realization led to the development of *Sing Pong*, a voice-controlled game to be discussed in section 3.3.

3.2 Vocal Telekinesis

Watching a table vibrate below the speakers while testing my sound-based projects has always fascinated me. The sound output from those speakers did not only move the surface of the table but also directed my thoughts towards a way to program sound to physically move objects in the real world.

Most people who conducted the 'sound-vibrations experiment' with rice at school know that sound does actually cause motion; it causes objects to vibrate at the frequency of the waveform. Each object has a set of natural frequencies at which it may vibrate. While the stereo is playing music, the air around the stereo vibrates. If one of the sound frequencies generated from the stereo matches the natural frequency of a nearby object, resonance will occur.

There are claims that some sopranos are able to shatter glass with their voices. When the singer's voice is loud enough and when its pitch matches the resonance frequency of the glass, the glass may shatter. Another phenomenon that can be caused by resonance is the collapse of a bridge. A recent example was the footbridge across the Thames built to celebrate the millennium. The Millennium Bridge started swaying when crowds walked across it. Dampers were installed but the bridge is still known to irreverent Londoners as "The Wobbly Bridge" and probably will be for ever.

Within the field of interactive media, sound can also be programmed to "cause" any kind of movement. In many of the few audio-physical applications which exist today, physical movement is what controls sound and not the opposite. Some of the developers of these applications map various aspects of sound to the physical movement of the user, while others map sound parameters and other sonic elements to the physical movement of real objects as will be explored in the next section.

a. Audio-Physical Applications

Among the projects that map sound parameters to the position of real objects is the performance instrument *Audiopad*, which allows for live electronic music composition (Figure 36). Developed by Patten and his colleagues at the MIT [2002], this instrument's interface is projected on a tabletop surface and is controlled by pucks.

The spatial position of the pucks on the table determines certain properties of the sound. Each puck is associated with the number of samples that the performer intends to control. *Audiopad* detects the position and orientation of the pucks and maps the volume and other parameters of the music to these parameters [Patten et al., 2002].

Figure 36: *Audiopad*; a performance instrument that allows for live electronic music composition [Patten et al., 2002].

The performer can associate each puck with a different track. Each puck resonates at a different frequency on the table. Hence, radio sensors are used to determine the position of each puck. Rotating a puck controls the volume of the

associated track. Changing the position of a puck changes the *"effect settings"* of the associated track.

One of the remarkable audio-physical projects that map sound to the physical movement of the user is *SoundSlam* in which the user punches a bag at certain locations to trigger and control pre-recorded audio files and create song compositions [Kirschner et al., 2003]. The bag contains pressure sensors, a global acceleration sensor, and a sound processor unit [Kirschner et al., 2003]. It sends out midi commands to the computer which triggers sound as a response to the physical input.

A similar project that exploits physical movement as an input to control music is *MvM* (Music Via Motion). Developed by a team led by Kia Ng at the Interdisciplinary Centre for Scientific Research in Music, *MvM* is a system that uses infrared light to capture the user's three-dimensional movement and turns it into music [Ng et al., 2000] [Ng, 2004]. Reflective balls are attached to the clothes the user wears and infrared light is projected onto these balls which are monitored by several cameras [Wakefield, 2004]. Through these cameras, the computer recognizes the position of the balls and transforms the instructions into music [Wakefield, 2004]. Although *MvM* was initially developed for dancing, the developers propose that the technique could also be used to scroll a webpage [Wakefield, 2004].

Another audio-physical installation that plays music in response to the physical input of the user is *Digiwall*. Developed at the Interactive Institute in Sweden, the installation is a climbing wall which acts as a musical instrument [Liljedahl et al., 2005]. The grips of the wall, which act as the keys on a piano keyboard, are equipped with sensors that detect the climber's position on the wall and play music accordingly. The installation employs a variety of interaction models. In one of these models, *Music Memory*, each climbing hold triggers a melodic fragment when touched. The climber's task is to find the pairs of holds that trigger the same fragment.

A similar concept is implemented in *Sonic City* which was developed as a collaborative project between the Interactive Institute and the Victoria Institute in Sweden. *Sonic City* maps music to the user's physical movement through the city and to environmental factors such as light intensity, pollution, and presence of metals [Gaye and Holmquist, 2004]. The user, who is thereby composing music by walking, can hear the music in real-time through headphones. The user wears a jacket with built in sensors that sense the user's physical movement and the environmental factors.

One last noteworthy audio-physical device is *Balance Booster* which is worn on the belt and connected to stereo headphones to help people with balance disorders maintain balance. The device contains sensors that detect when the user sways outside a pre-specified vertical zone [Dozza et al., 2005]. Different tones are used to indicate the direction of the user's sway. The further out of the *"safe zone"* the user sways, the louder the sound output through the user's headphones.

All the above mentioned systems map sound to physical movement. They thereby reflect a significant step towards moving the computer and its user away from the screen. Focusing on sound output, or even sound input, as a primary means of interaction may allow the use of other modes of interaction simultaneously. It facilitates the avoidance of using the traditional mouse and keyboard. Using the acoustic mode as the main mode of interaction may reduce visual interaction with the screen, but it may also open up possibilities for visual interaction with objects other than, and away from, the screen. William Gaver explains this in terms of *attention directionality*:-

> *"Visual objects exist in space but over time, while sound exists in time, but over space […] In order to take advantage of visual information, one must look in the appropriate direction. Sounds may be heard from all around: One does not have to face a source of sound to listen to it. This implies that sound can convey information to users despite their orientation, while visual information depends on users' directed attention"* [Gaver, 1989a].

Unlike the installations discussed in this section, which map sound to physical movement, the next section explores attempts to map physical movement to sound or even voice.

b. *Voice-Physical* Applications

The current availability of a wide variety of input mechanisms reflects the potential of the human body as a rich source of input. Recently, voices, fingers, hands, and eyes have all been explored and employed in different multimedia applications as an input. This has, to a certain extent, transformed part of the computer monitor into a mirror that reflects the user sitting in front of it while also displaying the visuals processed "behind" it. These techniques, however, are not generally radical enough to break the barrier between the user and the computer. The widespread reliance on the visuals, and hence the desktop screen holds many users back from the perception of an amorphous computer that any device may embody. Only recently has more attention been driven towards what O'Sullivan and Igoe refer to as *"Physical Computing"* or using computers to affect the physical world:

> *"When asked to draw a computer, most people will draw the same elements: screen, keyboard, and mouse. When we think 'computer,' this is the image that comes to mind. In order to fully explore the possibilities of computing, you have to get away from that stereotype of computers. You have to think about computing rather than computers. Computers should take whatever physical form suits our needs for computing"* [O'Sullivan and Igoe, 2004].

With the increasing ubiquity of computer vision and video tracking, physical movement has almost always been associated with and expected from the user rather than from the computer. When expected from the computer, physical movement is often linked with android (humanoid robot) behavior either in reaction to a remote control or to spoken commands.

In this section, I aim to reflect my interest in programming computers to compensate for users' deficiencies rather than merely to imitate their abilities. Voice-visual applications, for instance, may counterbalance our inability to explicitly measure voice characteristics and produce visuals in reaction to them. My objective is to exploit the computer's capacity to respond multimodally to voice not only by visuals but also by movements, odors, or any kind of physical output. Through reversing the role of sound in an installation by making it a controlling factor rather than a controllable element, I hope to foster the concept of *voice-physical* interaction in the field of interactive media. *Voice-physical* installations are likely to prove a fruitful expansion of the possibilities already shown to be inherent in voice-visual and other existing forms of voice-controlled installations. Such an expansion seems to have a natural synergy with the move towards *"The Invisible Computer"* described by Norman [Norman, 1999].

In 1878, Edison built the *voice-engine*, or *phonomotor*, which converted voice-induced vibrations, acting on a diaphragm, into motion which can drive some secondary device. He thought that one might *"talk a hole through a board"* [Edison quoted in Dyer, 1929]. While experimenting with the telephone, Edison made a paper toy that resembled a man sawing wood. When one recited loudly into a funnel, a ratchet wheel connected to the diaphragm would move and cause a pulley connected to it to rotate. The pulley which was connected by a cord to the paper toy caused the toy to move; *"Hence, if one shouted: 'Mary had a little lamb,' etc., the paper man would start sawing wood"* [Edison quoted in Dyer, 1929].

In the late 1960s Edward Ihnatowicz developed two sculptures, *SAM* (Sound Activated Mobile) and *The Senster*, both of which responded to sound (Figure 37). *SAM* was an electro-hydraulic sculpture consisting of a flower-like form mounted on a *"vertebrate-like neck"* [Zivanovic, 2000]. When *SAM* detected a voice or any kind of noise, its structure was inclined towards the source. *Senster* was a more advanced sound-activated hydraulically-operated sculpture. Around 15 feet long and looking like a lobster's claw, it responded to a soft voice by turning towards it, and to a loud voice by shying away [ibid]. Two input

mechanisms were employed to achieve this: four microphones to detect sounds and their directions, and two radar transceivers to detect movements. The computer differentiated between background public noise and foreground voices aimed at interacting with the *Senster* by comparing the sounds detected by the microphones.

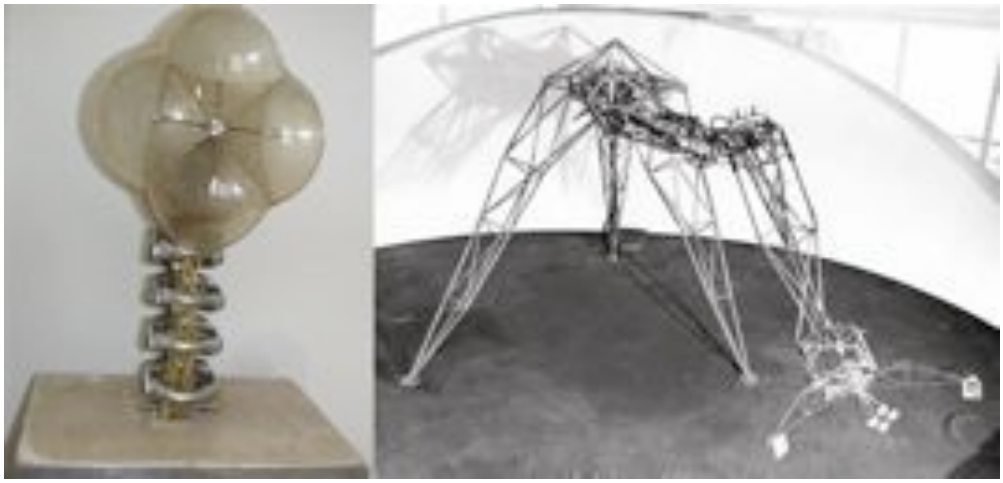


Figure 37: *SAM* (left) and *Senster* (right) are sound-activated hydraulically-operated sculptures that were developed by Edward Ihnatowicz in the 1960s [Zivanovic, 2000].

During the 1980's toys such as the dancing flower and the dancing Coke can were launched (Figure 38). These toys contain a sound chip which responds to sound and makes the toy 'dance'.

Figure 38: The dancing Coke can and the dancing daisies were developed during mid 1980's.

In 2000, Kelly Dobson developed *Blendie*; a blender that responds to a *"blender-like"* voice [Dobson, 2003] (Figure 39). The blender contains motion control hardware and sound analysis software that track the pitch of voice and matches its motion to it. The higher the pitch of the voices generated the faster the blender spins.

Figure 39: *Blendie*; a vocally-interactive blender developed by Kelly Dobson [Dobson, 2003] [Correa, 2004].

In the same year, the Danish artists Kjell Yngve Petersen and Karin Sørensen developed *Smiles in Motion* [Sørensen and Petersen, 2000]. This installation features a set of two chairs that vibrate in reaction to visitors' speech and non-speech voices (Figure 40). These vibrations become the two visitors' means of communication while sitting on the chairs. The louder the voice of the sender, the more the chair of the receiver vibrates. By transforming voice into movement that is caused by hidden motors in the seats, the chairs are designed to link the two visitors and allow them to understand each other's expressions.

In 2001, *His Master's Voice (HMV)* was developed by Kirschner and others [Kirschner et al., 2001]. *HMV* is a robotic board game in which players use their voices to move ball robots. The balls are programmed to move only when a certain pitch is hummed (Figure 41).

Figure 40: *Smiles in Motion*; a set of chairs that vibrate in response to visitors' voices [Sørensen and Petersen, 2000].

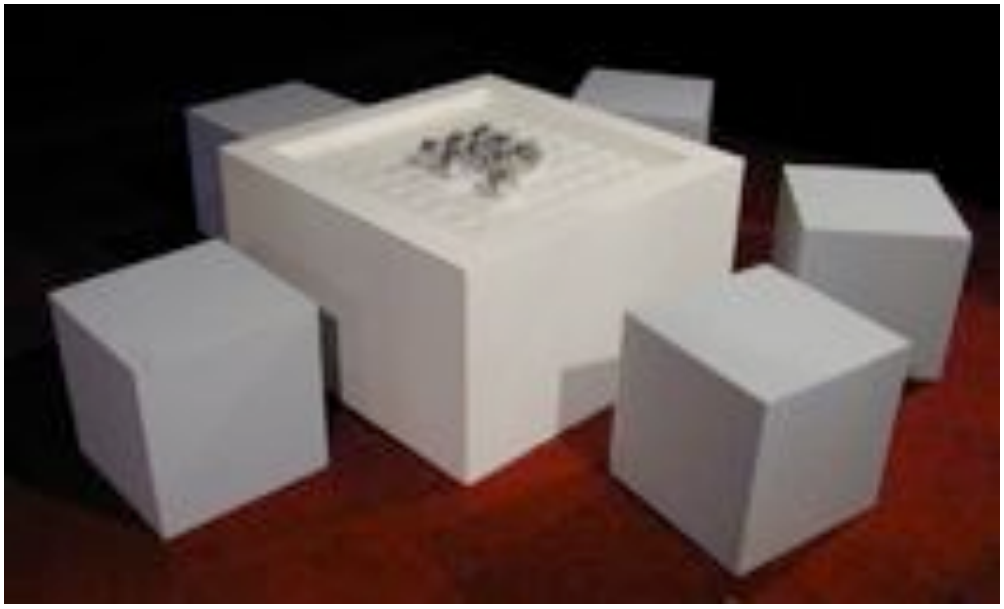


Figure 41: *HMV*; a voice-controlled robotic board game [Kirschner et al., 2001].

In 2002, Borland and others developed *Front*; a project that features two voice-controlled suits made of inflatable plastic air sacs: *"Each suit has two systems of inflatable air sacs, one for aggressive, and the other, defensive response"* [Borland et al., 2002]. A microphone in each suit detects each user's voice, and the volume of the voice controls the amount of inflation (Figure 42).

Figure 42 : *Front*; voice-controlled inflatable suits [Borland et al., 2003].

In 2004, Alexandre Armand and Bram Dauw developed *Commotion*; a two-player car-racing game in which the volume of each player's voice is mapped to the acceleration of an electric car [Armand and Dauw, 2004]. The louder the voice, the faster the car moves. To control the car, each player wears a helmet that vibrates when the player's voice causes the car to go off track (Figure 43).

Figure 43 : *Commotion*; voice-controlled racing cars [Armand and Dauw, 2004].

In 2005, Daan Roosegaarde developed *4d Pixel*; an interactive wall that represents a dynamic surface consisting of physical pixels which react to voice [Roosegaarde, 2005]. When activated by voice, these pixels move to display patterns or text (Figure 44).

Figure 44 : *4d Pixel*; a voice-controlled interactive wall [Roosegaarde, 2005].

All the above mentioned projects investigate the possibility of representing voice physically by detaching it from the body that produced it and transforming it into a visible and concrete element. I was interested in developing these ideas further. For that reason, my voice-visual work progressed towards the development of a number of *voice-physical* installations, which are described in the next section.

3.3 Implementations and Evaluations of Voice-Visualization and Vocal Telekinesis

a. *Sing Pong*: a Voice-Visual Game

Sing Pong is a computer projected game based on the use of the characteristics of paralinguistic utterances to control the traditional video game *Pong*, originally created by Ralph H. Baer in 1966. The original video game consists of two on-screen paddles and a ball. Each paddle is controlled by a player, who moves it up and down using a keyboard or a joystick. Hitting the ball with a paddle moves it to the other side of the playing field where the other player hits it back.

Figure 25: An illustration of the voice-controlled game *Sing Pong.*

Sing Pong, however, which I developed with another postgraduate at the Lansdown Centre for Electronic Arts during my MA studies, enables players to control the paddles using their voices (Figure 25).

The game is displayed on a projection screen, and the two players stand in the light beam of the projector (Figure 29). The height of the paddle is mapped to the volume of the player's voice (Figure 26). The louder the volume, the taller the paddle becomes. The position of the vocal paddle corresponds to the position of each player's head (Figure 27). The position of the head's shadow on the projection is tracked using a web-cam to determine the original position of the head in the real environment. Hence, the paddle appears to emerge from the player's mouth (Appendix B).

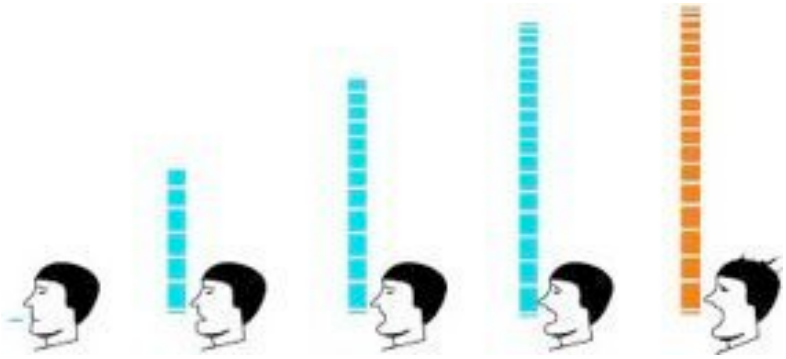


Figure 26: The height of the paddle is mapped to the volume of a player's voice.

Figure 27: The position of the paddle is mapped to the position of the player's shadow.

Sing Pong is also playable by players in wheelchairs. They may move the wheelchair forward and backward to control the position of the paddle, while using their voices to control the height of the paddle (Figure 28).

Figure 28: *Sing Pong* is playable by players in wheelchairs.

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 initializes 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 initializes 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.

The main hardware requirements for the development of *Sing Pong* included a web camera, a USB external sound card, a projector, two wireless microphones, and a fast computer (Figure 29). The programming tool used to develop the game was Macromedia Director and Lingo. Two Xtras (external software modules) for Director, Fast Fourier Transform Xtra (FFT) and

TrackThemColorsPro Xtra (TTC), were used for voice analysis and simple computer vision respectively.

TrackThemColorsPro is used in *Sing Pong* to track players' movements by detecting their shadows. It detects objects based on their color and brightness and calculates the position of these objects. The first step in programming the game involved mapping the position of a paddle to the position of a player's shadow. An invisible rectangle was mapped to the position and size of each shadow. In order to make the paddle appear as if emerging from a player's mouth, the paddle was programmed in a way that positioned it a certain distance below the top and from the edge of the bounding rectangle (Figure 30). The game was initialized by players positioning their shadows in the middle of the screen.

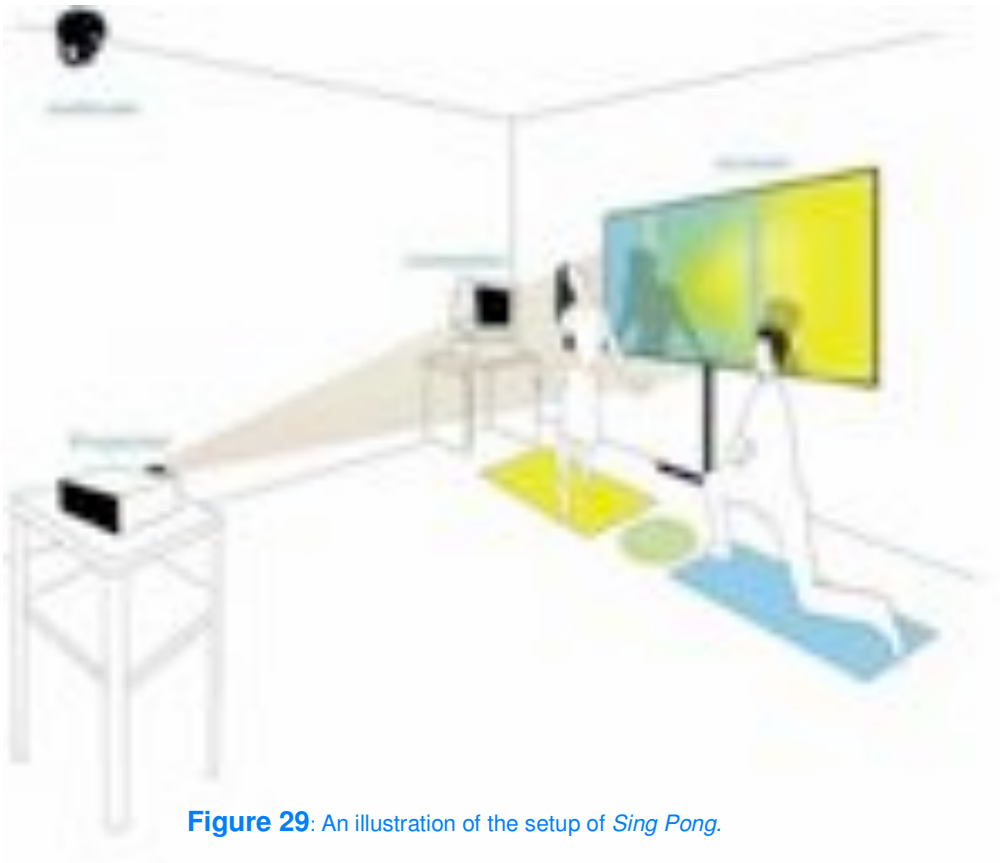


Figure 29: An illustration of the setup of *Sing Pong*.

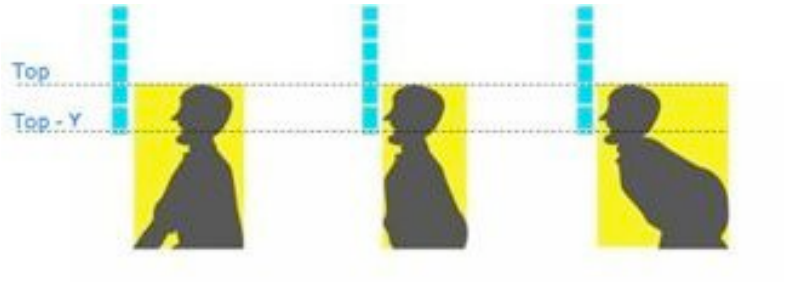


Figure 30: The paddle is positioned a certain distance below the top of the bounding rectangle.

While testing the game some interesting problems arose from the use of the invisible bounding rectangle. These problems included the fact that some players used their hands to raise the height of the bounding box or to extend its width in order to control the position of the paddle (Figure 31). This "problem" was actually a solution for short players. As so often, a potential bug was turned by users into a feature.

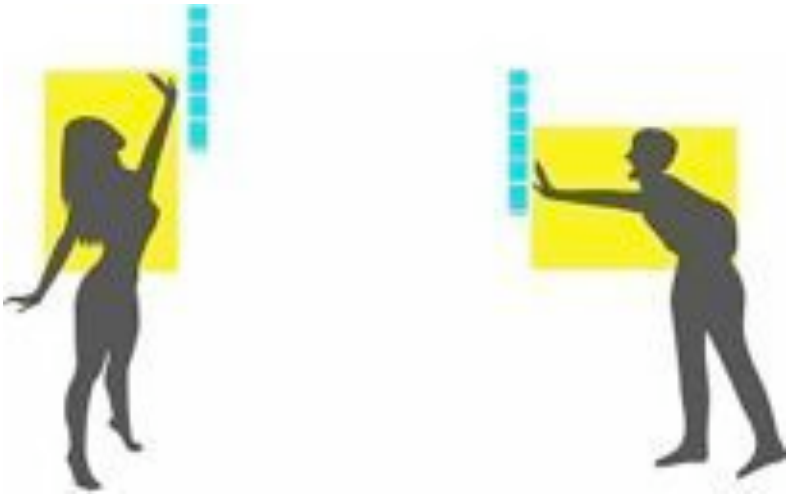


Figure 31: Players using their hands to raise the height of the bounding box or to extend its width.

Another problem was the inconsistency in the position of the paddle. Since the width of the player's shadow determined the position of the paddle, the gender and weight of the player influenced the position of the paddle (Figure 32).

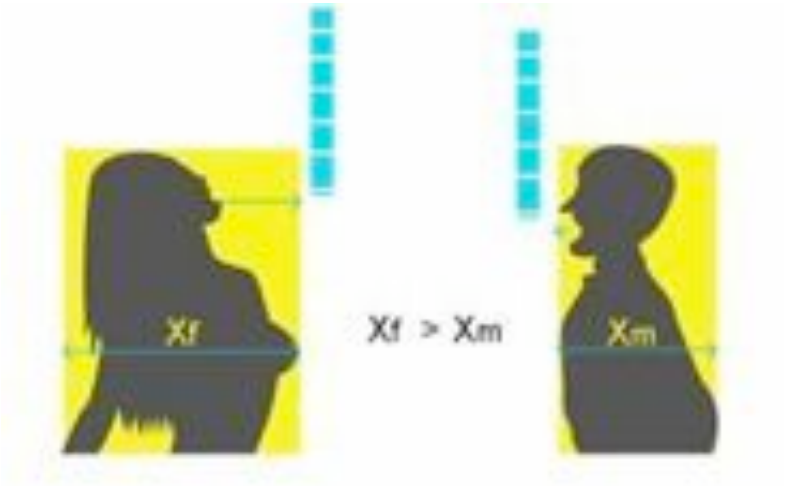


Figure 32: The paddle is positioned further away from the mouth when the player is an adult female.

Problems to be overcome included latency, which was a main problem in an early prototype of the game, until the code was made more efficient. Lighting also affected the darkness of the shadows and therefore the response of TrackThemColorsPro. For example, if one of the players was standing close to the window and sun light made the player's shadow lighter, the tolerance in tracking black blobs (on the lighter side of the screen) had to be reduced.

All the problems discussed above are related to video-tracking. There were other problems encountered that were related to the detection and signal processing of the players' voices.

asFFT, an Xtra for Macromedia Director, is used in *Sing Pong* to analyze player's vocal input signals. It employs the Fast Fourier Transform (FFT) algorithm and is written by Antoine Schmitt [Schmitt, 2003]. The Fast Fourier Transform is used to capture the frequency data of the user's voice in near-real time. Hence, it adds voice analysis capabilities to Macromedia Director and allows the program to respond to live voice input.

In order to enable asFFT to detect the players' vocal input as two separate input signals, an external USB sound card is used. This allows asFFT to recognize the

built-in sound card and the external one as two separate devices and to analyze their inputs separately and simultaneously.

While detecting voice input, one of the issues of concern was the possibility of having acoustic feedback into the microphones from the speakers or interference from the other player's voice. This problem was largely obviated through the minimization of sound effects used in the game and by increasing the minimum amplitude level to which the paddle was mapped. Moreover, the microphones were attached to headsets and were therefore very close to each player's mouth.

Another interesting, though predicted, problem was that some players used the brute-force technique of screaming loudly to maintain the paddle at the full height of the screen. This was avoided by specifying a maximum allowed amplitude, at which the paddle turns red: the ball bursts when it hits a red paddle (Figure 33). In this way solving a problem (again) spawned a new feature in the game. This rule induced players to demonstrate higher level of control over their vocal skills and to maintain voices which were neither very loud nor very soft.

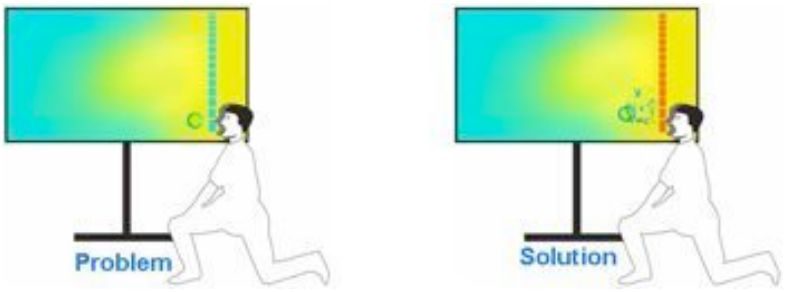


Figure 33: Specifying a maximum amplitude level in order not to allow players to keep the paddle as high as the screen.

The exhibition of the game in a gallery in London led to further unexpected informal observations and speculations (Figure 34). From a spectator's point of view, the playing space and the players seemed to be an integral part of the game. Their shadows were images, their voices were sound effects, their movements were animations, and their interactions were physical events. Moreover, any movement within the physical space allocated for playing and any interference from the audience could alter the play, and include them in the

interface. Some spectators disturbed 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 child who enjoyed projecting his shadow to interfere while his dad and sister played. *Sing Pong* confirmed that it is often unwise to consign audience members to the role of 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 could significantly alter the game. Thus, the actual interface was no longer determined merely by what the player saw within the screen but by what the camera saw within its field of view.

The game required a dark space, and this was possibly one of the factors which encouraged people to play it in front of others, and to vocalize in a relatively uninhibited way. The microphone which is normally used to amplify and exhibit the voice was in such a voice-visual installation used to transform the voice into visuals, and thus perhaps conceal it aurally by displaying it visually. This possibly shifted the players' focus from their voices as voices to their voices as visuals (Figure 35). I speculated that they did not think of their voices as an audible element of which they are the source as much as they thought 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, it seemed to me that players' vocal engagement with the visuals in *Sing Pong* diverted their attention and anxiety from thinking about what spectators thought of them towards thinking of play. They did not face the audience as singers but each other as contenders. This is possibly what eliminated shy player's shyness. On the other hand, certain people (perhaps the more outgoing ones) appeared to consider such a game an opportunity to release their energy, express their emotions, and impress or grab others' attention. All of this led me to infer that paralinguistic vocal control of interactive media may play a significant role in disinhibiting players as will be further explored in the fifth chapter. It also

stimulated my interest in pursuing an experiment (discussed in the fourth chapter) that investigates shyness as one of the factors that may affect users' vocal and behavioral interaction patterns with a non-speech voice-controlled system.

Vocal paralinguistic control is a new style of interaction between the human and the machine. It enables users to interact in an expressive manner by executing vocal expressions. In *Sing Pong*, the interaction sometimes led to unexpected dramatic excitement and creative improvisations. Some people, especially those who seemed to be shy, chose to whistle rather than generate "ahhh, ooh" voices. Others started jumping in order to raise the height of the paddle. Many players enjoyed shouting and attracting other visitors' attention. Some players placed two fingers in front of the projector in an attempt to control a paddle with each finger. Such improvisations are difficult to imagine in a speech recognition-based game that restricts players to limited languages and accents. Paralinguistic vocal control proved to be a transcultural mechanism enabling the development of performance-centered applications and allowing a high level of engagement and immersion by participants, regardless of their languages. It 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 users' hands and bodies 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 multimodal means of interaction, and extends the scope of interaction vocally, modally, and spatially. 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 but also engage spectators as second-level users. In *Sing Pong*, both players and audience are entertainers and entertained.

Figure 34: Two performers playing *Sing Pong* during an exhibition in London (2004).

Figure 35: Performers using their voices to make their paddles taller and using their hands to raise the position of their paddles.

It seemed that players' pleasure in interacting with *Sing Pong* could – in part – be due to the invisibility of the computer which is usually visible, and to the visibility of the voice which is usually invisible. Moreover, the invisibility of the computer has, perhaps, injected a note of magic into the game-play. For that reason, the work discussed next 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 color), 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 refer to as *Vocal Telekinesis*.

b. *sssSnake*: a *Voice-Physical* Game

Over the years, the perception of voice mainly as a communication tool has led to the development of many techniques and devices such as record players, telephones, and radios. The main aim was to preserve, reproduce, or transfer voice as it is; an audible means by which an expression is communicated. *sssSnake*, however, is an attempt to transform voice into a visible means by which a real object is physically altered.

I have not managed to *"talk a hole through a board"* yet [Edison quoted in Dyer, 1929]. However, this section explains how I was able, in a sense, to shout a coin out of the screen in *sssSnake*. The project is one of three comprising the practical component of my Ph.D.

sssSnake is a two-player *voice-physical* version of the classic *Snake* game. The game consists of a table on which a real coin is placed and a virtual snake is projected (Figure 45). Four microphones, one on each side of the table, are used to detect the voices. One player controls the snake by uttering 'ssss', while the other player moves the coin away from the snake by uttering 'aahh'.

The position of a player round the table determines the direction of the coin's or snake's path. The snake moves towards the player uttering 'ssss', and the coin moves away from the player uttering 'aahh'. Through prompting players to run round the table, the game encourages physical participation as well as vocal activity. A video documentation of *sssSnake* is available (see Practical Projects p.208).

Figure 45: *sssSnake*: a *voice-physical* version of the classic *Snake* game (2005).

The 'ssss' and 'aahh' voices are not differentiated through speech recognition but rather through the detection of frequency differences between the high-pitched sibilant 'ssss' and the low-pitched 'aahh' (Figure 46). Thus, the game is an attempt to exploit non-speech voice characteristics and explore new uses and implementations of sound theory in the field of interactive media.

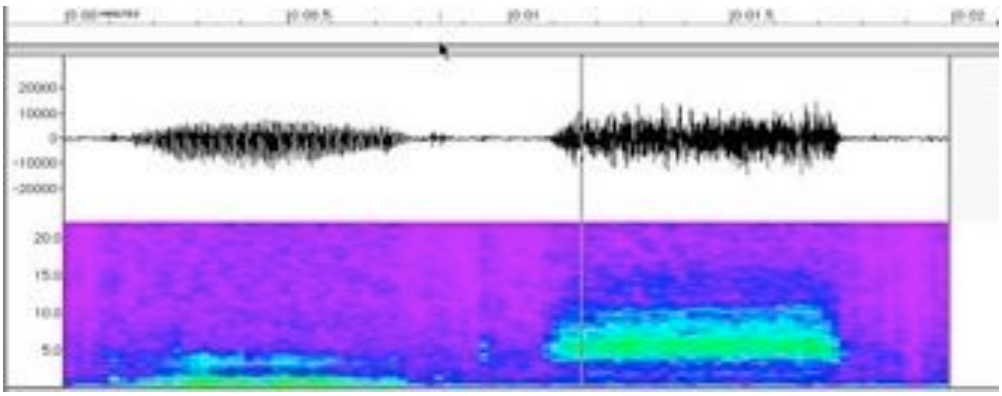


Figure 46: The difference in frequency between the low-pitched 'ahh' (left) and the high-pitched 'ssss' (right).

sssSnake was programmed in Lingo/Macromedia Director, using the asFFT Xtra discussed earlier. Four microphones are used to detect voice input. These microphones are plugged into the USB ports of the computer through USB audio adapters by different manufacturers. The reason for using different adapters is to overcome a limitation of asFFT which causes it to detect several adapters by the same manufacturer as one.

In order to avoid offering extreme physicality as a viable gameplay strategy, it was decided to design the game in such a way that it is never in both players' interests to share the same microphone. The coin was therefore programmed to move away from the player uttering 'aahh', and the snake to move toward the player uttering 'ssss' (Table 4). Thus, sharing a microphone is to the advantage of the player controlling the snake and to the disadvantage of the other. This decision was taken into account in the algorithm described below.

Being on the perimeter of the game table, the microphones were close to one another. As a result, detecting the microphone nearest to the source of a particular sound was a major challenge. When two sounds are uttered simultaneously, the simple expedient of detecting the microphone registering the loudest volume fails. The voiced 'aahh', which requires opening the mouth and involves the vibration of the vocal folds, is always louder than the unvoiced sibilant 'ssss' which is interdental and only requires a partial opening of the air

stream and involves no vibration. Thus the microphone that detects the loudest volume is activated by the player uttering 'aahh'. One cannot however assume that the microphone that detects the second loudest volume is activated by the player uttering 'ssss'. The volume detected at this microphone could also be caused by the player uttering 'ahh'.

The following steps are taken to analyze vocal input:

1.Using the asFFt Xtra, frequency data is collected for each microphone and separated into two components corresponding to the frequency spectra of the 'ahh' and 'sss' sounds.

2. The microphone with the most powerful signal in the high frequency spectrum is identified.

3. If the power of this spectrum exceeds a specified threshold, this microphone is identified as being the closest to a player uttering 'sss' and removed from further consideration (Table 4).

4. The microphone (one of three or four, depending on the result of step 3) giving the most powerful signal in the low frequency spectrum is identified.

5. If the power of this spectrum exceeds a specified threshold, this microphone is identified as being the closest to a player uttering 'ahh'.

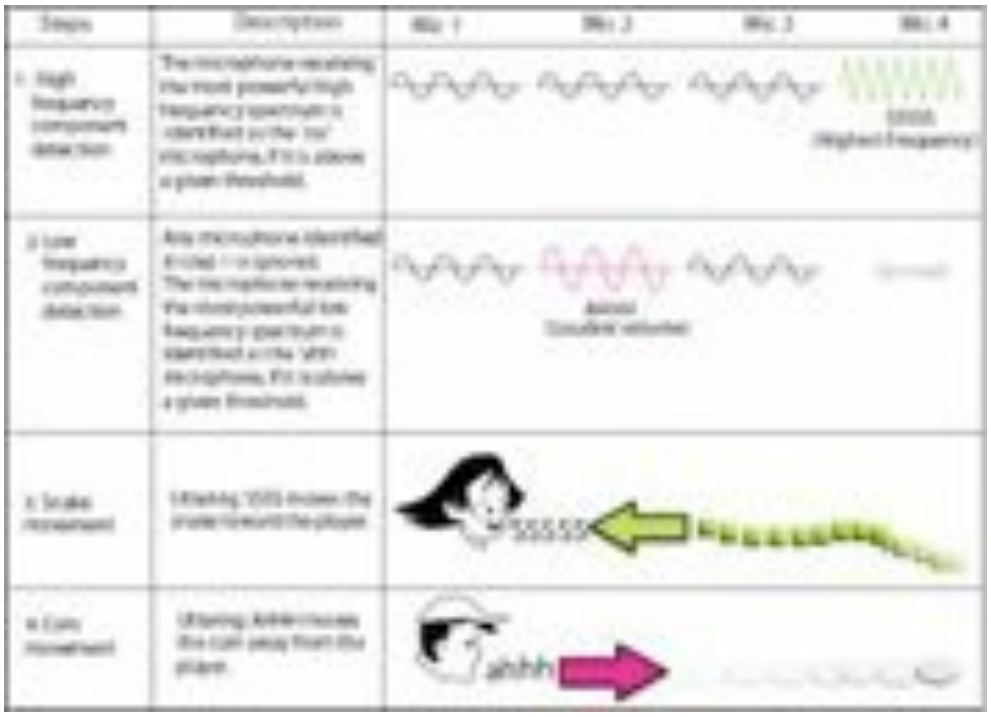
This algorithm was developed after considerable experimentation. In addition, significant work went into calibration, namely:

1. Identifying the most suitable ranges for the two frequency spectra.

2. Identifying the most suitable threshold for each spectrum.

It may be noticed that the algorithm privileges the detection of the 'sss' sound at the expense of the 'ahh' sound when both players are nearest to the same microphone. Paradoxically, this *favors* the player uttering 'ahh' as it is never in his/her interest to be at the same microphone as the other one, as explained earlier.

Table 4: The steps involved in determining the microphone and utterance used in *sssSnake*. In the example given, both players are vocalizing.



The next step involved using Hewlett-Packard Graphics Language (HPGL) to program a plotter to move its head in response to the 'aahh'. A magnet was attached to the head, and the plotter was hidden below the table. Hence, the virtual coin, which was initially used to test the screen-based version, has been replaced by a real coin moving away from a snake projected on the surface of the table.

The completed screen-based version of the game was exhibited and user-tested in an exhibition in London (September, 2005). The game attracted people from different age groups, and allowed children to play with their friends, siblings, parents, or even grandparents (Figure 47). Several players seemed to react positively to the fact that *sssSnake*, unlike many other games, requires a great amount of physical movement (Figure 48). They also suggested increasing the width of the table in order to prompt players to run larger distances. The running and bumping which the game involved, as well as the tendency of some players

to forget which side of the table they should be at, brought much laughter into the play. From an entertainment perspective, this kind of interactivity made the game fun for players to play and for spectators to watch. From an artistic-performative perspective, such a *voice-physical* game may direct an interactive system to performative extremes and allow it to be experienced as a physically and emotionally holistic engagement, with the whole body as a source of expressive input and performance. From a social perspective, the game brought many people together. Players who did not know each other played together (Figure 49), and viewers who had not been introduced laughed and joked together (Figure 50).

Some players who considered that they had become experts in the game taught new visitors how to play. The game also encouraged seemingly shy people to overcome their shyness; watching others play expressively seemed to entice them to try it. The game involved lots of touching, pulling, and pushing. Some players of opposite sex apparently liked the fact that the game increases the physical proximity, reducing interpersonal distance. *sssSnake* encourages the multidisciplinary investigation of the proxemics [Hall, 1966] and social aspects of computer-mediated communication.

Although many spectators attempted to get involved in the game and transform it into a four-player game, the game worked better with two players only. Having four players allows movement of the coin or snake along one axis only. It also hinders physical movement round the table.

Most of the players found the coin more difficult than the snake to control. The coin required a smooth and continuous 'aahh' or 'ooohh' to move, while any tone of 'ssss' seemed to move the snake easily. This was probably due to the ambient noise and the sound of movement round the table which were low-pitched enough to affect the calibration for the 'aahh'. On the other hand, not many surrounding sounds or background voices were high-pitched enough to affect the 'ssss' calibration. Although the difficulty of moving the coin is

fundamentally a flaw, the continuous tone and steady pitch required to move it turned out to be a good chance to train the vocal cords.

Alternatively, the accuracy of detection of the high-pitched 'ssss', made it tricky to move the snake using other utterances. This obliged players to only use 'ssss' to move the snake. One of the players, for instance, had a lisp and tried enthusiastically to produce a perfect 'ssss' to move the snake. The snake's movement, in reaction to her voice, assured her that her 'ssss' was not an 'thththh'. Accordingly, it is not entirely inconceivable that the game may turn out to be therapeutic to players who suffer from speech impairments.

Figure 48: children playing *sssSnake* in an exhibition in London (2005).

sssSnake prompted both shy and outgoing people to play it. I observed that, on the whole, apparently shy players chose to control the snake using the voiceless 'ssss' and apparently outgoing players did not mind shouting 'aahh' to move the coin. For all players, especially shy ones, the game and its presentation seemed to involve some kind of a psychological preparation process or what could be referred to as *vocal disinhibition*. Before trying the game or even entering the room at which it was exhibited, almost all visitors who read the instructions

smiled while reading the part which indicated that they would be using ‘aahh’ ‘oohh’ ‘ssss’ voices to play. Perhaps this smile was a manifestation of their mental image of themselves making these voices.

Figure 49: Players who did not know each other played *sssSnake* together (2005).

Figure 50: Viewers who had not been introduced laughed and joked together (2005).

When I explained the game to visitors, many of them repeated after me when I uttered 'ssss' or 'aahh' even before they started playing. It seemed that they were trying to vocally "visualize" and practice the game-play. After leaving the room, some players kept on uttering 'ssss' and laughing even outside the context of the game: the production of non-speech vocalizations seemed enjoyable in its own right.

c. *Blowtter*: a Voice-Controlled Plotter

Blowtter, the second practical project submitted for the Ph.D., is a voice-controlled plotter that allows a user to blow into a *mic-board* in order to draw (Figure 51). The *mic-board* is a small square board that consists of four microphones, one on each of its four corners (Figure 52). Blowing into the top microphone moves the head of the plotter up, blowing into the bottom microphone moves it down, blowing into the left microphone moves it left and blowing into the right microphone moves it right – as if pushing it. The notion of blowing fits neatly with the action of pushing the head. Blowing is used as an input in order to facilitate the directness and continuity required for drawing which is not easily achieved by using a spoken command repetitively such as saying "move..move…move" or "move, 10". Igarashi and Hughes further explain the advantage of this technique:

> *"[…] one can say 'Volume up, ahhhhhh', and the volume of a TV set continues to increases while the 'ahhh' continues. The advantage of this technique compared with traditional approach of saying 'Volume up twenty' or something is that the user can continuously observe the immediate feedback during the interaction. One can also use voiceless, breathed sound."* [Igarashi and Hughes, 2001: 155]

Where appropriate, speech commands are also used in *Blowtter*. For example, "start plotter" and "stop plotter". Saying "up" raises the head and allows for moving it without drawing. Saying "down" moves the pen into the paper. Saying the number of the pen causes it to be selected. The objective behind the employment of speech-recognition is to investigate the implications of its

integration with non-speech voice input. A video documentation of *Blowtter* is available (see Practical Projects p.208).

The advantage of using blowing lies in the possibility of targeting it with one microphone without interference through nearby microphones. The act of blowing itself tends to involve unconsciously placing the mouth very closely to the blown-into object (Figure 53). Thus, each microphone, rather than sharing airborne sounds with the other microphones, actually generates its own (therefore local) sound by acting like the mouth-hole on a flute. Blowing is perceptually voiceless despite the fact that the microphones may detect it and the computer may measure its voice characteristics. Unlike voiced sounds which involve the vibration of the vocal folds while being generated, blowing and other unvoiced sounds only involve air passing through the larynx without causing the vocal folds to vibrate. Furthermore, the range of amplitude values that voiced sounds can produce is wider and louder than the range of amplitude values generated by blowing. This fact is exploited in specifying a higher threshold for speech than for blowing as will be explained in detail later.

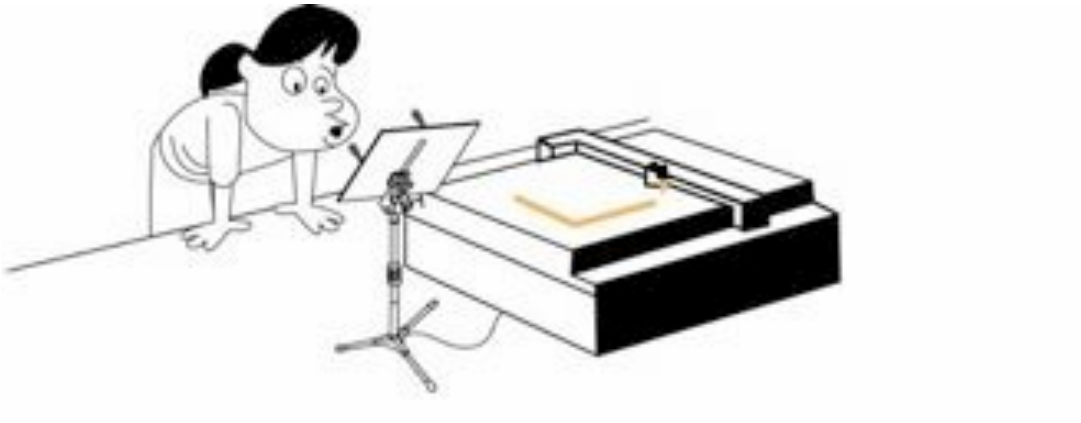


Figure 51: *Blowtter*: a voice-controlled plotter that allows a perhaps disabled user to blow into a *mic-board* in order to draw.

Figure 52: A user blowing into the *mic-board* (right) in order to draw.

Figure 53: A user trying to write the word "BLOWTTER" by blowing.

The main hardware components of *Blowtter* include a DXY Roland plotter, four USB audio adapters, four microphones, and a fast personal computer. The application was again programmed in Macromedia Director/Lingo. Three Xtras were used: asFFT, Chant, and Direct Communication. Some of the techniques used in this project are similar to those in *sssSnake*.

The vocal signal is analyzed using the asFFT Xtra. By using the Xtra to measure and compare the different volume levels at each microphone, the software was enabled to recognize which microphone of the four the user is blowing into. The microphone at which the maximum amplitude is detected determines the direction for moving the pen. To avoid the plotting head reacting to any voice input other than blowing, an amplitude threshold is specified. When a volume above the threshold is detected, the software assumes that the user is speaking rather than blowing and the speech recognition Xtra, Chant, is activated. Furthermore, to avoid the plotter head reacting to soft ambient noises, another lower threshold is specified. Any noises below it are completely ignored. The Hewlett-Packard Graphics Language (HPGL) is again used to control the plotter. In order to insure accuracy, a pitch threshold is also specified. When a pitch (here identified as the most powerful part of the frequency spectrum) above the pitch threshold and a volume below the volume threshold are detected, the software assumes that the user is blowing because blowing is high-pitched in comparison to most speech sounds.

Chant Speech Kit, a speech-recognition Xtra for Director, is used to recognize the commands spoken by the user. The Xtra supports a number of speech recognition applications, in this case Microsoft SAPI. Chant Speech Kit consists of a command, grammar, and dictation vocabulary list. For *Blowtter*, however, only the command list is enabled and customized to include only the commands that control the plotter. The dictation vocabulary is disabled, hence detection errors are minimized by programming the Xtra to capture only the commands in the customized command list.

Direct Communication Xtra allows Director to communicate with an external device either through the serial or the parallel port.

Blowing as a paralinguistic activity has also been utilized innovatively in a very few interactive works. The *Yacht* game by Nintendo DS, for instance, allows the player to blow into the microphone to propel a boat and steer it in the sea (Figure 54). The force of breath appears to determine the speed of the yacht. Another

Nintendo game, *Candles*, also allows the player to blow into the microphone to blow out candles; the larger the size of the candle and its flame, the harder the player must blow (Figure 55). In both cases, users probably believe that the force of the air-stream is the direct determining factor, whereas it is really the volume of the sound created by blowing on the microphone.

Figure 54: The *Yacht* game by Nintendo DS allows the player to blow into the microphone to propel a boat and steer it in the sea [Okimoto, 2005].

Blowing "Windows" by Matsumura at The Royal College of Art, London, also employs blowing [Matsumura, 2005]. It allows the user to blow into one end of a hose of which the other end is directed towards a computer screen in order to control, move, and rearrange desktop icons (Figure 56). The duct contains a wireless microphone that detects the blowing and measures its intensity. This intensity as well as the size of the file represented by the icon determines the speed at which the icon moves. The hose also contains tilt switches that detect the angle at which the hose is held and determine which side of the desktop to rearrange.

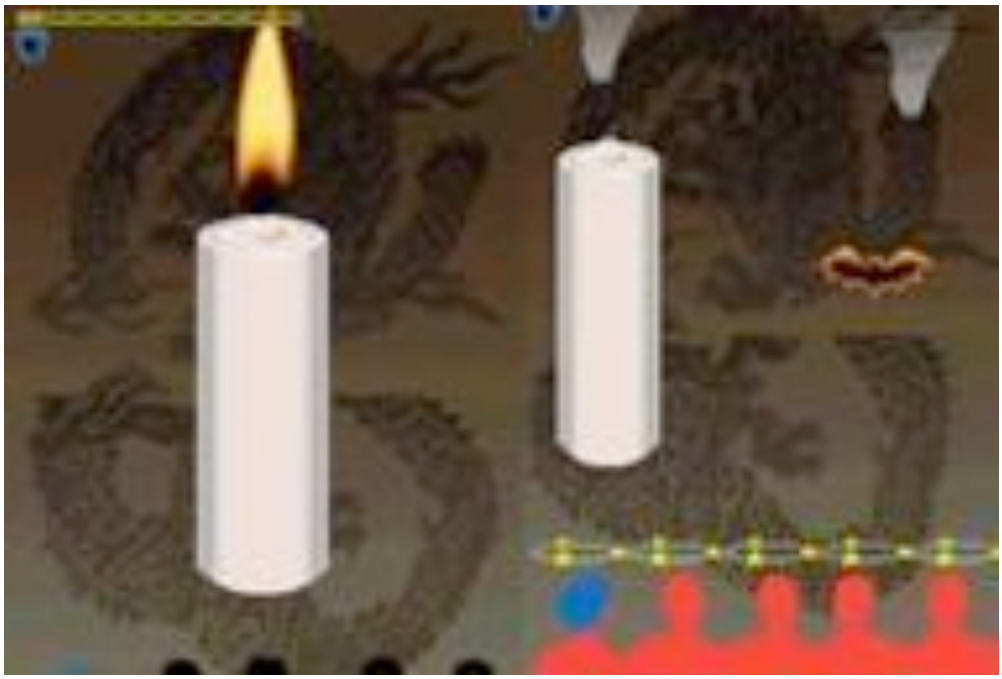


Figure 55: *Candles* by Nintendo DS allows the player to blow into the microphone to blow out candles [Okimoto, 2005].

Another remarkable application of blowing is *Kirifuki* which allows for interaction with visual desktop objects by inhalation and exhalation (Figure 57). The system consists of a breath microphone switch containing a Polhemus sensor that detects the orientation of the user's head, a projector that projects the desktop on a desk, and a magnetic gyroscope [Iga and Higuchi, 2002]. The microphone differentiates between breathing in and blowing out using the differences between the acoustic signals. Iga and Higuchi found that the randomness of the exhalation signals is less than that of the inhalation signals, and by measuring and comparing the local peak numbers the system ingeniously recognizes the sound input as either an inhalation or an exhalation [Iga and Higuchi, 2002]. The mouse pointer is mapped to the orientation of the head which is tracked using the Polhemus sensor. The implementation of this technique allows for a variety of interaction mechanisms. The technique can be applied to make the user's breath diffuse the icons around the mouse pointer, while an inhalation may assemble these icons again. It also allows an inhalation to cut a virtual object, while an exhalation pastes it again (Figure 58). If the object being manipulated is three-dimensional, an inhalation deflates it while an

exhalation inflates it again. The system may also be used for drawing where an exhalation sprays the drawn shape while an inhalation erases it. Another application of the technique involved inhaling to reduce the amount of coffee in a virtual cup, and exhaling to fill it up again (Figure 59).

Figure 56: A user blowing into a hose to rearrange desktop icons in *Blowing "Windows"* [Matsumura, 2005].

Figure 57: A user interacting with desktop objects by inhaling and exhaling into a microphone in *Kirifuki* [Iga and Higuchi, 2002].

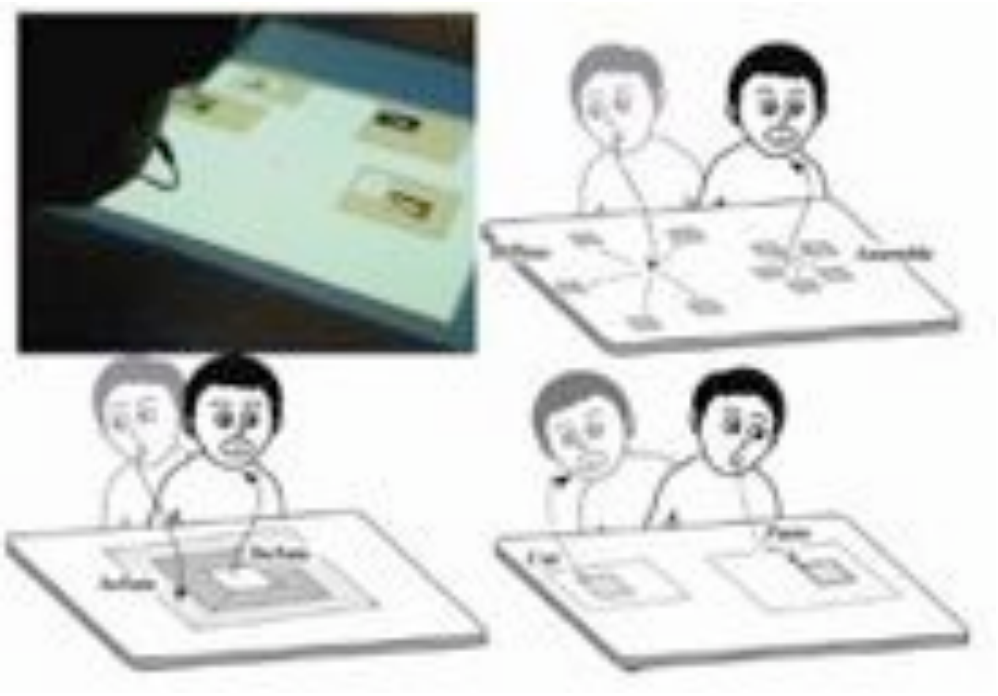


Figure 58: *Kirifuki* allows for interaction with visual desktop objects by inhalation and exhalation [Iga and Higuchi, 2002].

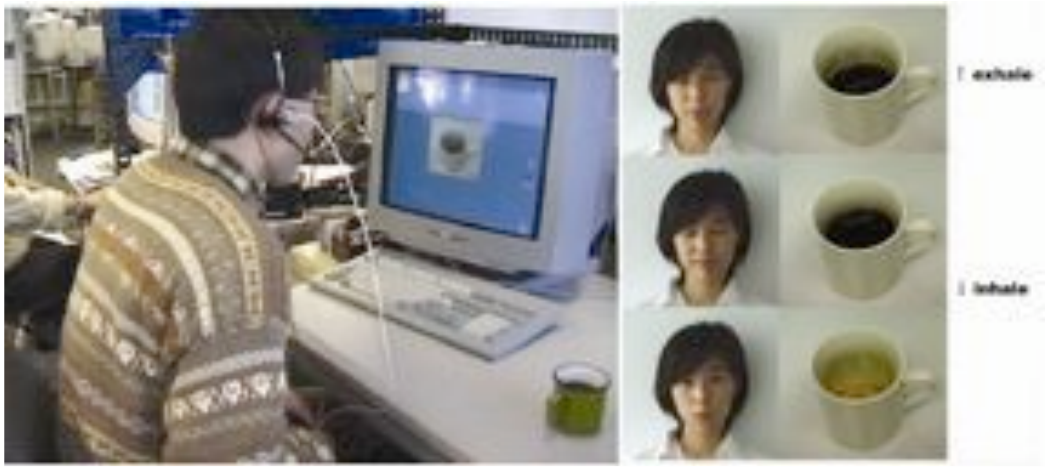


Figure 59: *Kirifuki* allows for interaction with a virtual coffee cup by inhalation and exhalation [Iga and Higuchi, 2002].

Figure 60 : *Blow up*; breath-controlled fans [Snibbe, 2005].

One last noteworthy installation is *Blow Up* (Figure 60). Developed by Scott Snibbe, this installation allows the sender to blow into a group of twelve small impellers that control an array of twelve large fans which replicate and magnify the speed and movement of the small impellers on the receiving end [Snibbe, 2005]. The array of small impellers is placed in one side of the gallery, while the fans they control are placed on another side. Through the development of this installation, Snibbe seems to emphasize that the physical interaction of our voices and breaths with surrounding media has a significant influence on our inference about the existence of these body activities.

In addition to the exploitation of blowing as an input source, *Blowtter* involves a utilization of the un-voiced-ness of blowing to propose a new use for the microphone. The fact that blowing is un-voiced makes it possible to place the four microphones used to control *Blowtter* in a very close position to each other on the *mic-board*. The size of the board is around 10x10 cm and the distance between one microphone and the other is around 5 cm (Figure 52). This minimizes the need to move from a microphone to another where only the user's

face is expected to move slightly to direct the air stream into one of the microphones. For such an application designed especially for disabled users, minimizing movement is very necessary. The *mic-board* may thus be used by disabled users as an alternative to a joystick in some modified versions of existing games. It may also conceivably be used as a cheap alternative – in some systems – to sip-and-puff controllers which are comparatively expensive, technically complicated, and may pose the risk of spreading infections from a user to another. Further implementations may also involve increasing the numbers of microphones. Hence, several microphones may be placed on a circular rather than a square board or even forming a matrix of microphones. Blowing into a microphone on the right side of the board and then into a microphone on the left side would draw a straight line from right to left. Moving the head slightly in a certain pattern while blowing sequentially into one microphone after another would draw the pattern or shape that the user's head composes while moving. This would increase the opportunity of allowing smoothly graduated control.

I carried out a qualitative evaluation of *Blowtter* with five disabled children at Bahrain Al-Shamil Rehabilitation Center. Their disabilities included cerebral palsy, growth impairment, motor impairments, and Down's syndrome. Owing to these disabilities all of the participants had varying degrees of difficulty in using their hands to draw.

The children were individually presented with *Blowtter* in the presence of their instructor, Sahar Sabri. They were first asked to hold the pen and draw a flower like the one that Sahar drew in front of them. This task was performed in order to roughly indicate the extent of difficulty in drawing by hand. Next, they were asked to blow into the *mic-board* to draw.

The first participant, a seven-year-old male (moderate cerebral palsy and growth impairment), could not understand that he was expected to blow into the *mic-board*. The instructor suggested that this could be because the *mic-board* was an unfamiliar device to the child. She suggested allowing the child to see a natural

physical reaction to blowing first, cut several slits into the side of a paper, positioned the paper directly in front of the *mic-board*, and asked the participant to blow at it and observe the effect of the air-stream (Figure 61). This turned out to be an effective technique as it facilitated the participants' understanding of what they were expected to do and encouraged them to blow. It was clear, however, that the first participant had difficulties in blowing strongly, which reflects a need for incorporating a technique that allows the user to adjust – and in this case decrease – the volume threshold for the detection of blowing.

Figure 61 : The instructor tried to encourage the child to blow by allowing him to observe the effect of the air-stream on a piece of paper with some slits cut into it.

Unlike the first participant who did not seem to entirely perceive the relationship between blowing and the movement of the plotter head, the second participant, an eight-year-old male (Down's syndrome and low vision), laughed every time he blew and saw the plotter head moving. This encouraged him to keep on blowing and drawing. One intriguing observation was that when the instructor asked him to draw a chicken, this participant started imitating chicken sounds

and clucking. It might be fruitful to provide a collection of generic graphics that can be vocally 'clipped-out' (as in clip art) and automatically plotted when the user imitates the sound of the animal or even says the name of the object to be plotted.

The third participant, a ten-year-old female (mental retardation and speech impairment), also realized the association between blowing and the movement of the pen. This caused her to blow harder while pointing at the pen to show her instructor that the pen was moving. The instructor remarked: "*We usually put balloons or tennis balls in water and ask children with breathing difficulties to blow them in order to encourage the children to blow. This device seems more effective for asthmatic users*."

The fourth participant, a nine-year-old male (Down's syndrome), tried to move the pen using his hand when the instructor instructed him to move the pen, but after the instructor demonstrated how he should blow to draw, he cupped his hands in front of his mouth and used them to direct the air-stream toward the *mic-board*.

The fifth participant, a nine-year-old male (growth impairment) did not seem to perceive the relationship between blowing into the *mic-board* and drawing. His attention was directed toward blowing into the *mic-board* only and not toward watching the effect on the plotter at the same time. This led the instructor to remark that the location of the *mic-board* (which was on the right side of the plotter) and its height (which was almost equal to the height of the plotter) may affect the user's perception of the relationship between the *mic-board* and the plotter. Because the *mic-board* was higher than the drawing area and not positioned directly in front of it, some participants did not immediately establish a connection between the *mic-board* and the plotter. Because the pen holder is on the left side of the plotter head while the *mic-board* was on the right side of the plotter, the participants who were on the right side of the plotter (due to the location of the *mic-board*) seemed to have difficulties in seeing the drawn line, which was semi-covered by the plotter head if viewed from their location.

This prompted me to realize that I had already encountered this problem without fully recognizing it when I tested *Blowtter* with two normal participants who volunteered to test it at Middlesex University. One of these participants, a nineteen-year-old female, kept moving her head to the left to see if the pen was plotting after every time she blew. This could simply have been obviated by positioning the *mic-board* on the left side of the plotter, or even better, in front of it.

This adult participant as well as the second participant, a twenty-six-year-old male, both managed to write the word 'Blowtter' by blowing. The only blowing-related-limitation that I noticed was that it was tiring after a while and the participants had to take a break every few minutes because they ran out of breath. Of course, it would have been even more tiring if the user had to blow to move the head all the way to the location of a pen in order to select it, rather than utter the number of the pen. This is where the use of speech-recognition seemed sensibly to complement the paralinguistic voice input. Unlike blowing, which had human-related limitations, speech-recognition revealed technology-related limitations. During the development stage, I noticed that the engine confused the numbers 'one', 'two', etc. I therefore decided to program the software to recognize the numbers 'first', 'second', etc. instead. Although this increased the recognition accuracy, the only three commands that the engine could always accurately distinguish were 'second', 'seventh', and 'eighth'. I also had to add homophones and similar-sounding words to increase accuracy; saying 'it', 'ate', 'eight', or 'eighth', for example, would all cause the plotter to select the eighth pen. Another limitation was that the recognition engine had to be trained by each participant for at least fifteen minutes. Even after being trained, it still was not accurate in distinguishing between all the numbers. Moreover, the engine could not recognize any of the numbers that the second normal participant uttered with a Portuguese accent. As for the disabled participants, they all only spoke Arabic, which was often incomprehensible even to me due to their speech impairments. It would have been almost impossible for the system to recognize their speech because the calibration algorithms of the engine I used (Microsoft SAPI) assume "normal" speech. If *Blowtter* only depended on spoken commands, I would not

have been able to test it on those five disabled users, and probably not on non-English-speaking users. This is where blowing, a form of paralinguistic input, seemed to have complemented speech-recognition and overcome some of its weaknesses.

On the basis of my evaluation of *Blowtter* and comparing it with my previous evaluations of *Sing Pong* and *sssSnake*, it seems more appropriate to employ paralinguistic input alone to control a game or an entertaining work than to control a practical and serviceable application. Interaction with a game demands less accuracy than with a functional application. A game usually offers a chance to try again after losing. Error and loss are meaningful components of the action in most games but surely not part of any practical application. For this reason, a game is especially likely to benefit from exclusive use of paralinguistic input, whereas in the case of a practical everyday application, this kind of input is likely to be of greatest benefit when used with a complementary input mechanism.

One advantage of using blowing in *Blowtter* rather than using voiced sounds as in *sssSnake* is that it ensures accuracy by minimizing microphone interference. The underlying technique is shared in both applications, but the extent to which a technical inaccuracy is perceptible and tolerable by the user is not. Because of a limitation of asFFT when used with multiple microphones, two microphones could wrongly detect the same amplitude level even when the player was much closer to one microphone than the other while making loud sounds in *sssSnake*. Occasionally, this complicated the amplitude comparison operation and caused the coin to momentarily deviate. That was, nonetheless, hardly noticeable by the players who were, anyway, running and laughing. For a disabled user sitting still and using *Blowtter* to draw accurately, however, such deviation is intolerable and unquestionably perceptible. Therefore, the un-voiced-ness of blowing and the inherent notion of having to bring the mouth very close to the 'blown' object, which in this case is one of the microphones, make it possible to obviate this problem.

Another important aspect to keep in mind while developing a voice-controlled work is its context. Since fun and laughter are often part of an entertaining system, making non-speech sounds to control a game is reasonable and –as a bonus– attention-grabbing to passers-by. On the other hand, a user interacting with a practical voice-controlled interface will most probably not want an audience watching and laughing. As *Blowtter* is not designed as a game, this is another reason behind choosing blowing rather than vocalizing to control it.

The use of *Blowtter* as a drawing tool for physically impaired users could have significant potential. Non-speech voices can provide this user segment with a new dimension for exploring their artistic talents or even a means of communicating thoughts, ideas and concepts. Exploring novel aspects of voice-controlled applications might empower the physically impaired in other ways too.

d. *Expressmas Tree*: a Voice-Controlled Christmas Tree

When thinking of Christmas decorations, one would usually visualize brightly colored ornaments and sparkling lights. Vibrating vocal cords are usually far from one's imagination. This section, however, presents *Expressmas Tree*: the vocal decoration of a Christmas tree. It is the third and final practical project of the Ph.D..

Expressmas Tree is an interactive voice-physical installation with real bulbs arranged in a zigzag on a real Christmas tree. Generating a continuous voice stream allows users to sequentially switch the bulbs on from the bottom of the tree to the top (Figure 62). Longer vocalizations switch more bulbs on. *Expressmas Tree* is a game in which every few seconds, a random bulb starts flashing. The objective is to generate a continuous voice stream and succeed in stopping upon reaching the flashing bulb. This causes all the bulbs of the same color as the flashing bulb to light. The successful targeting of all flashing bulbs within a specified time-limit results in lighting up the whole tree and "winning".

A video documentation of *Expressmas Tree* is available (see Practical Projects p.208).

The main hardware components included 52 MES light bulbs (12 volts, 150 milliamps), 5 microcontrollers (Basic Stamp 2) (Figure 63), 52 resistors (1 k), 52 transistors (BC441/2N5320), 5 breadboards, regulated AC adaptor switched to 12 volts, a wireless microphone, a serial cable, a fast personal computer, and a synthetic Christmas tree (Figure 64).

The application was programmed in Pbasic and Macromedia Director/Lingo. Two Xtras (external software modules) for Macromedia Director were used as before: asFFT and Serial Xtra.

Because one Basic Stamp 2 chip contains only sixteen I/O pins, four chips were used to control the fifty-two bulbs and one extra chip was used to control these four chips and to communicate with the computer. Thus, one of the five Basic Stamp chips was used as a 'master' stamp and the other four were used as 'slaves'. Each of the slaves was connected to thirteen bulbs, thus allowing the master to control each slave and hence each bulb separately. Three of the sixteen I/O pins in each chip were used for computer-chip communication in the case of the master chip and for chip-chip communication in the case of the slave chips.

This work explores an element of voice that has not been exhaustively exploited in multimedia applications: its duration. This parameter, which is calculated through the difference between the onset and offset of voice, is what determines how high on the tree is the bulb that is finally lit. The exploitation of voice duration in an interactive work might open new doors towards exploring voice-timing perception and temporal coordination. It may also contribute to a better understanding of the psychophysics of vision and voice interaction as well as *"bimodal sensory integration"* [Grauwinkel and Fagel, 2006], an interaction that involves the stimulation of two sense modalities. It may also play a role in the treatment of some voice disorders such as spastic dysphonia; *"a communicative*

disorder characterized by difficulties in voice control, specifically voice initiation, termination, and maintenance" [Izdebski et al., 1978].

Figure 62: *Expressmas Tree*: a voice-controlled Christmas tree.

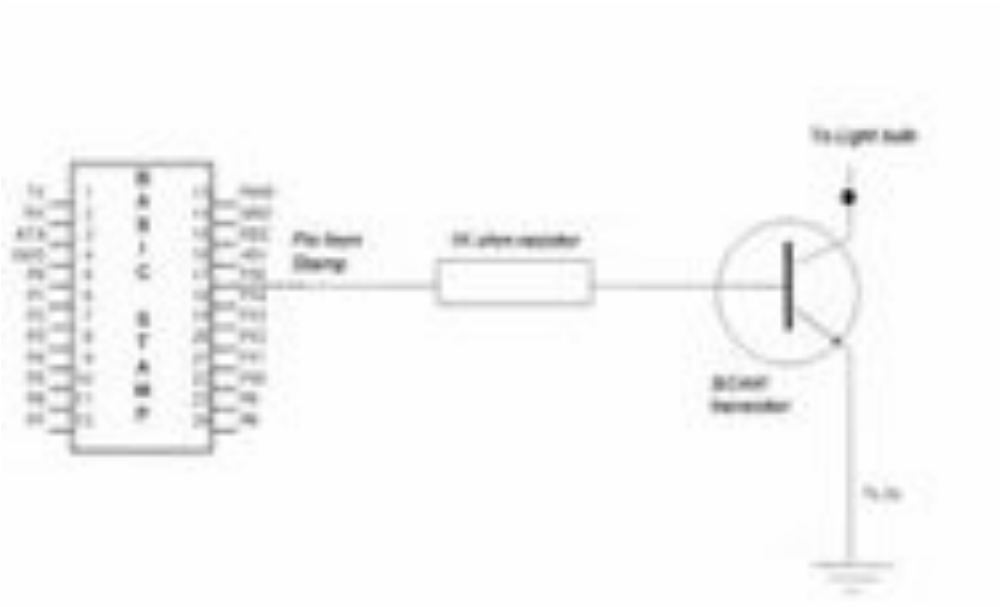
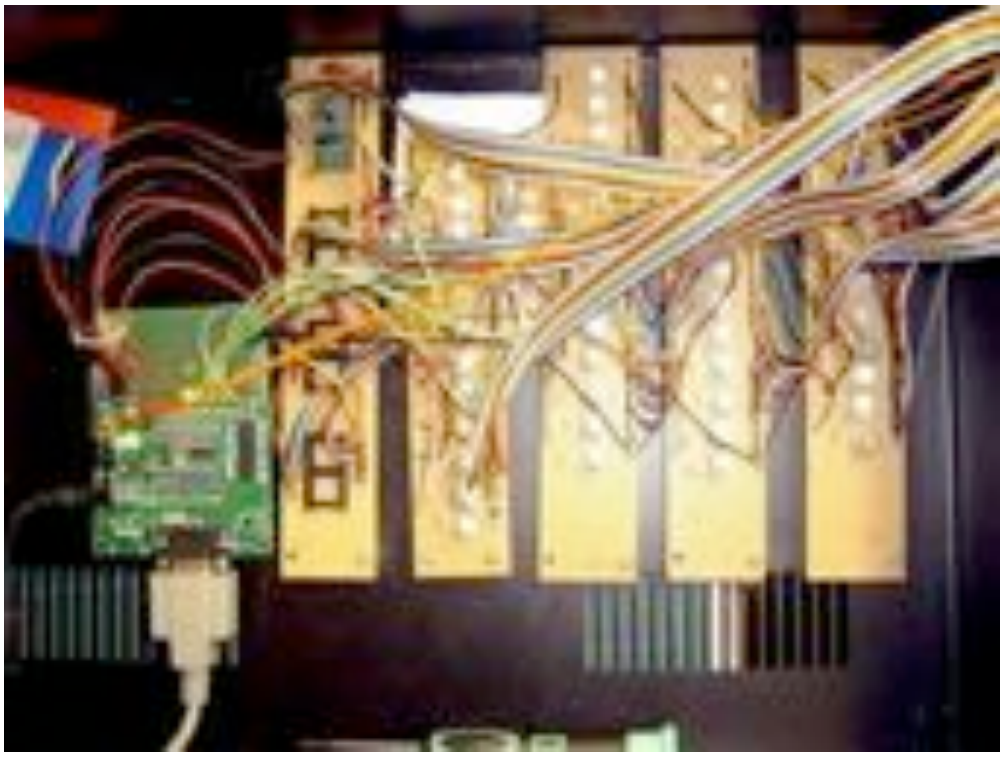


Figure 63: A schematic representing the basic elements and structure for connecting a bulb to the circuit used in *Expressmas Tree.*



Although my work is mainly an attempt to exploit non-speech voice, the employment of *Vocal Telekinesis* in the development of speech-controlled *voice-physical* applications may also be useful. A functional application would be a *voice-physical* avatar or robot that is pre-programmed to recognize speech and physically display its equivalent in sign language. This robot may be portable or may also be connected to or placed next to a television for the hard of hearing.

While the previous chapter attempted to put together knowledge from a large number of domains in the service of my theoretical research on paralinguistic vocal control, this chapter attempted to put together knowledge from the field of interactive media in the service of my practical work.

The chapter offered a comprehensive review of existing audio-visual and audio-physical work in order to explore the correlations that developers assumed and

the techniques that they have adopted to create mappings between sounds and visuals. It then advocated the progression of audio-visual applications into voice-visual installations, and the evolution of audio-physical applications into voice-physical performances. It therefore introduced the terms *voice-visual* and *voice-physical* to the field of interactive media in an attempt to enrich the repertoire of interaction. It provided a review of existing *voice-visual* work that employs voice as an input and screen-based visuals as an output (in what I refer to as *voice-visualization*) and of existing *voice-physical* work that employs voice as an input and non-screen-based physical events as an output (in what I refer to as *vocal telekinesis*).

It finally discussed my own implementations and evaluations of *voice-visualization* by presenting *Sing Pong* and of *vocal telekinesis* by presenting *sssSnake, Blowtter*, and *Expressmas Tree*. Evaluating these projects revealed some of the benefits of paralinguistic vocal control. These included the possibility of maintaining a near-real-time and continuous response, the potential of cross-cultural use, the relatively low cost and the technical ease of using the microphone as an input device, and the possibility of use by people with motor impairments especially if their impairments are accompanied by speech impairments.

The progression from *voice-visual* applications to *voice-physical* installations revealed further benefits. Like most non-conventional interfaces, the use of voice as a physical tool pushes certain boundaries of human-computer interaction and allows interaction with forms of output beyond conventional screen-based two-dimensional and three-dimensional visuals. It allows the visual interface to extend beyond the computer screen and to become an integral part of the physical world. It may hence free up parts of the body other than the vocal apparatus and allow their use as complementary input techniques. In the case of entertaining and performative applications such as *sssSnake*, this allows for a higher degree of expressiveness and richer modes of input. In the case of functional and practical applications such as *Blowtter*, this allows for a higher degree of control, adroitness, and naturalness through the complementary use of

body parts (such as feet) with which the motor-impaired user is already skilled and familiar.

Another potentially enriching benefit is the dimension of magic and ambiguity that the visibility of the voice and the invisibility of the devices in such projects add to the interactive experience. This dimension, especially when unconstrained by restrictive instructions, allows creative users to make a virtue out of technical, physiological, and other limitations by supplementing them with their own interpretations, improvisations, and graceful interactions. The next chapter shows how impelling users to interpret interactive systems for themselves may unveil or even conjure unforeseen and inspiring patterns of interaction.

4. Preferences and Patterns in Paralinguistic Voice Input

4. Preferences and Patterns in Paralinguistic Voice input

The practical projects undertaken during my Ph.D. have been playful and open-ended, exploiting paralinguistic aspects of voice through a variety of informal methods. However, I became interested – as the work progressed – in the particular voice-input-related aspects of the users' behavior, and decided to look at this rather more formally.

This chapter investigates the factors that affect users' preferences of non-speech sound input and determine their vocal and other interaction patterns with a non-speech voice-controlled system. It attempts to throw light on shyness as a psychological determinant and on vocal endurance as a physiological factor. It hypothesizes that there are certain types of non-speech sounds, such as whistling, that shy users are more prone to resort to as an input. It also hypothesizes that there are some non-speech sounds which are more suitable for interactions that involve prolonged or continuous vocal control. To examine the validity of these hypotheses, I exploited *Expressmas tree* in an experimental approach to investigate the factors that may affect users' preferences and interaction patterns during non-speech voice control, and by which the developer's choice of non-speech input to a voice-controlled system should be determined.

As no other studies appear to exist in the precise paralinguistic vocal control area addressed by this research, the chapter comprises a number of experiments that explore the preferences and patterns of interaction with non-speech voice-controlled media. In the first section, I discuss the experimental designs, procedures, and results. In the second section, I present the findings and their implications in an attempt to lay the ground for future research on this topic. The eventual aim is to help other developers of such systems in their input selection process, to enable them to avoid vocal input patterns that may be considered undesirably awkward, and to favor patterns that are serendipitously *"graceful"* [Wiberg, 2006]. In the last section, I discuss the conclusions and suggest directions for future research.

The project that propelled this investigation was *sssSnake* which was discussed in the last chapter. While user-testing *sssSnake*, players who were apparently shy seemed to prefer to control the snake using the voiceless 'sss' and outgoing players preferred shouting 'aahh' to move the coin. A noticeably shy player asked: "Can I whistle?". This question, as well as previous observations, led to the hypothesis that shy users prefer whistling. This prompted the enquiry about the factors that influence users' preferences and patterns of interaction with a non-speech voice-controlled system, and which developers might need to consider while selecting the form of non-speech sound input to employ.

In addition to shyness, other factors are expected to affect the preferences and patterns of interaction. These may include age, gender, cultural background, social context, and physiological limitations. There are other aspects to bear in mind. I, for instance, prefer uttering 'mmm' while testing my projects because I noticed that 'mmm' is less tiring to generate for a prolonged period than a whistle. This seems to correspond with the following finding by Sporka and Kurniawan during a user study of their Whistling User Interface [Sporka et al., 2005];

> *"The participants indicated that humming or singing was less tiring than whistling. However, from a technical point of view, whistling produces purer sound, and therefore is more precise, especially in melodic mode."* [Sporka et al., 2005]

The next section presents the experiments and their results.

4.1 Experiments and Results

a. First Experimental Design and Setting

The first experiment involved observing, writing field-notes, and studying video and voice recordings of players while they interacted with *Expressmas Tree* as a game during its exhibition in a canteen at Middlesex University.

Experimental Procedures:

Four female students and seven male students volunteered to participate in this experiment. Their ages ranged from nineteen to twenty-eight years. The experiment was conducted with one participant at a time while passers-by were watching. Each participant was given a wireless microphone and the following written instruction: "use your voice and target the flashing bulb before the time runs out". This introduction was deliberately couched in vague terms. The participants' interaction patterns and their preferred non-speech sound were observed and video-recorded.

Participants were then given a questionnaire to record their age, gender, nationality, previous use of a voice-controlled application, why they stopped playing, whether playing the game made them feel embarrassed or uncomfortable, and which sound they preferred using and why. Finally they filled in a 13-item version of the Revised Cheek and Buss Shyness Scale (RCBS) (Appendix A) [Cheek, 1983]. The aim was to find possible correlations between shyness levels, gender, and preferences and patterns of interaction.

Results:

As the tree was set up to look as indistinguishable as possible from a conventional Christmas tree, passers-by had to be informed that it was interactive. Those who were with friends were more likely to come and explore the installation. The presence of friends seemed to encourage shy people to start playing (Figure 66). Some outgoing players seemed to enjoy making noises in order to cause their friends and passers-by to laugh rather than as part of their immersion in the gameplay. Other than the interaction between the player and the tree, the game-play again introduced a secondary level of interaction: that between the player and the friends or even the passers-by. Many friends and passers-by were eager to help and guide players by either pointing at the flashing bulb or by yelling "stop!" when the player's voice reaches the targeted bulb. One of the players (P6) tried persistently to convince his friends to play the game

(Table 5). When he stopped playing and handed the microphone back to me, he said that he would have continued playing if his friends had joined in. Another male player (P3) stated *"my friends weren't playing so I didn't want to do it again"* in the questionnaire. This could indicate embarrassment; especially that this participant was rated as "somewhat shy" on the shyness scale, and wrote that playing the game made him feel a bit embarrassed and a bit uncomfortable.

Four of the eleven participants wrote that they stopped because they *"ran out of breath"* (P 1, 2, 4, and 10). One participant wrote that he stopped because he was *"embarrassed"* (P5). Most of the rest stopped for no particular reason while a few stopped for various other reasons including losing the game. Losing could be a general reason for ceasing to play any game, but running out of breath and embarrassment seem to be particularly associated with stopping in a voice-controlled game such as *Expressmas Tree*.

Table 5: The profile of participants in experiment 1(scoring over 49= very shy, between 34 and 49 = somewhat shy, below 34 = not particularly shy) [Cheek, 1983].

Participants	Preferred Vocalization	Shyness Score (0-65) 65=very shy	Shyness Level	Participant Profile			Have you used Voice-Controlled applications before?	Did playing the game make you feel embarrassed?	Did playing the game make you feel uncomfortable?
				Gender	Age	Given Nationality			
1	Ahh	29	Not partic.	F	27	Greek	Y	N	N
2	mmm	35	Somewhat	F	28	Malaysian	N	N	N
3	ooh	41	Somewhat	M	22	British	N	a bit	a bit
4	ooh	41	Somewhat	M	20	Zimbabwean	N	a bit	a bit
5	Ahh	42	Somewhat	M	19	British	N	Y	Y
6	Ahh	30	Not partic.	M	19	Cypriot	Y	N	N
7	Ahh	38	Somewhat	M	20	British	N	Y	N
8	mmm	37	Somewhat	F	20	Polish	N	a bit	Y
9	Ahh	21	Not partic.	F	20	British	Y	N	N
10	Ahh	37	Somewhat	M	24	British	N	Y	Y
11	mmm	44	Somewhat	M	22	British	N	N	Y

The interaction patterns of many participants consisted of various vocal expressions, including unexpected vocalizations such as 'bababa, mamama, dududu, lulululu', 'eeh', 'zzzz', 'oui, oui, oui', 'ooon, ooon', 'aou, aou', talking

to the tree and even barking at it. None of the eleven participants preferred whistling, blowing or uttering 'sss'. This is presumably due to their awareness that their voices can better control the installation when they are louder than the ambient public noise. It may also possibly be related to the *"Lombard effect"* which is the tendency of people to change the characteristics of their voices in noisy environments [Lane and Tranel, 1971]. Six of the participants preferred 'ahh', while three preferred 'mmm', and two preferred 'ooh'. Most (four) of the six who preferred 'ahh' were males while most (two) of the three who preferred 'mmm' were females. All those who preferred 'ooh' were males (Figure 65). Out of the eleven participants, participant 9 was the least shy (shyness score = 21) and participant 11 was the most shy (shyness score = 44). The least shy participant preferred 'ahh' and the most shy participant preferred 'mmm' (Table 5). This, however, can only be a rough indication since the number was too small for me to draw statistically valid conclusions.

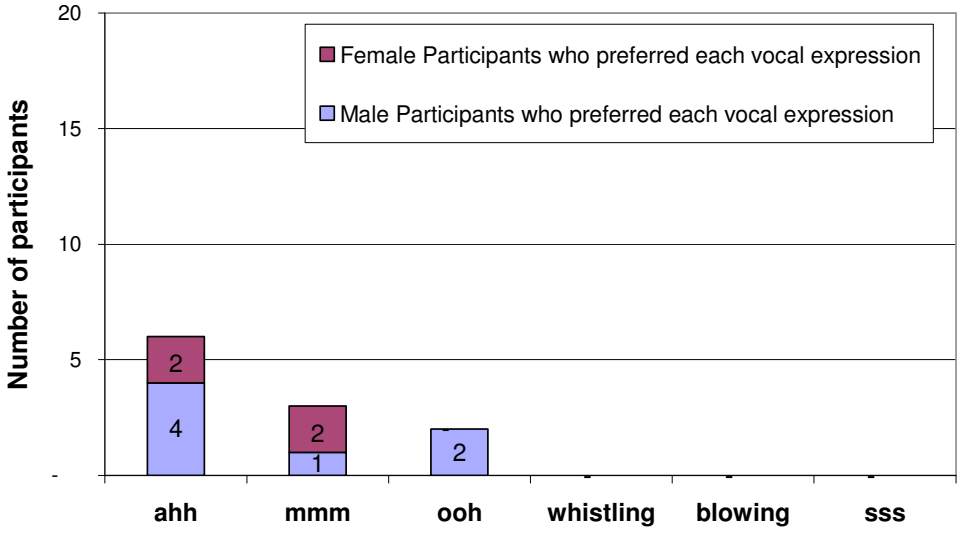


Figure 65 : Correlating the preferences and genders of participants in experiment 1.
The sound indicated is the one ranked as most preferred by each participant.
Sounds are arranged on the abscissa from the most preferred (left) to the least preferred (right) according to the combined results of both experiments (Figure 78).

Figure 66 : A participant playing *Expressmas Tree* while her friend enjoys watching her play.

Figure 67 : A participant uttering 'aah' to control *Expressmas Tree.*

b. Second Experimental Design and Setting

The second experiment involved observing, writing field-notes, as well as studying video-recordings and voice-recordings of players while they interacted with a simplified version of *Expressmas Tree* in a closed room (Figure 67).

Experimental Procedures:

Thirty-seven subjects (eighteen females and nineteen males) were recruited to participate in this experiment. A fee of GBP10 was paid. Their ages ranged from ten to sixty-two years. One participant stopped in the middle of the experiment because she was *"embarrassed"* to use her voice in a quiet room. She said that she would feel more comfortable with using her voice if the tree was exhibited in a public space and if her friends were around. Her preliminary results were discarded without further statistical analysis.

The simplified version of the game that the thirty-six participants were presented with singly was the same tree but without the flashing bulbs which the full version of the game employs. In other words, it only allowed the participant to light up the sequence of bulbs consecutively from the bottom of the tree to the top. The participants were presented with this simplified version in an effort to eliminate confounding factors such as engagement in gameplay and striving to understand the rules. The experiment was conducted in a closed office. Each participant was given a wireless microphone and a note with the following instruction: "See what you can do with this tree". This introduction was deliberately couched in very vague terms. After one minute, the participant was given a note with the instruction: "use your voice and aim to light the highest bulb on the tree". After another minute, the participant was given a note with the instruction: "make non-speech sounds and whenever you want to stop, say 'I've finished' ". The participant was then asked to type an answer to the question: "why did you stop?". The instructions were written rather than verbal in an effort to ensure that consistency would be maintained.

Each participant underwent a vocal endurance test, in which s/he was asked to light up the highest bulb possible by continuously generating each of the following six vocal expressions: whistling, blowing, 'ahhh', 'mmm', 'ssss', and 'oooh'. These were the six types most commonly observed during evaluations of previous work. However, the participants were also asked to generate the sound that they thought would best allow them to target the highest bulb. Every six participants performed the six sounds in a different order, so as to ensure that each sound was in turn tested initially without being affected by the vocal exhaustion resulting from previously generated sounds. The duration of the continuous generation of each type of sound was recorded along with the duration of silence after the vocalization. As most participants mentioned that they *"ran out of breath"* and were observed taking deep breaths after vocalizing, the duration of silence after the vocalization may indicate the extent of vocal exhaustion caused by that particular sound. After the vocal endurance test, the participant was asked to rank the six vocal expressions based on preference (1 for the most preferred and 6 for the least preferred), and to state the reason behind choosing the first preference. Finally each participant filled in the same questionnaire used in the first experiment including the Revised Cheek and Buss Shyness Scale. In order to insure that the participants vocalize as long as they could and would not stop as a result of reaching the highest bulb, the speed at which the bulbs consecutively lit was reduced. Moreover, the top bulb was deliberately disabled in an attempt to trick the participants who managed to light the highest bulbs into believing that that last bulb could light up if they vocalized longer.

Results:

When given the instruction "See what you can do with this tree", most of the participants did not vocalize to interact with the tree, despite the fact that they were already wearing the microphones. They thought that they were expected to redecorate it and therefore their initial attempts to interact with it were tactile and involved holding tinsel or the baubles in an effort to rearrange them (Figure 68). One participant responded: "I can take my snaps with the tree. I can have it in

my garden". Another said: "I could light it up. I could put an angel on the top. I could put presents round the bottom". The conventional use of the tree for decorative purposes seemed to have overshadowed its interactive application, despite the presence of the microphone and the computer. Only ten out of the thirty-six participants clearly realized it was interactive; some of them thought that it involved video tracking and moved backward and forward to interact with it, some thought that they could interact with it by clapping, but only six realized at first that it was voice-controlled.

When given the instruction "use your voice and aim to light the highest bulb on the tree", thirteen participants' initial vocal attempt was paralinguistic, twenty-one participants' initial vocal attempt was linguistic, and two participants were not sure of what to do (Figure 69). The fact that the initial vocal attempt of the majority of participants was linguistic reflects the ingrained assumption, by many, that the verbal rather than the nonverbal aspects are employed when voice is used as an input.

Figure 68 : When given the instruction "See what you can do with this tree", most of the participants attempted to redecorate it (despite the presence of the microphone and computer).

Table 6: The profile of participants in experiment 2
Each group contains six participants instructed to utter the same order of vocalizations.
That order is not represented in the diagram.

Participants	Preferred Vocalization	Shyness Score (0-65)	Participant 's Profile		
			Gender	Age	Nationality
1	mmm	36	F	19	Bahraini
2	ahh	39	M	62	British
3	ahh	40	M	25	Taiwanese
4	whistling	36	M	25	British
5	ahh	26	M	22	Portuguese
6	ahh	40	F	53	British
7	whistling	44	F	26	British
8	ahh	36	M	12	British
9	ahh	40	F	55	British
10	blowing	34	F	24	Chinese
11	mmm	28	F	18	Italian
12	ahh	36	M	25	German
13	mmm	31	M	21	Asian/British
14	ahh	42	F	29	British
15	ahh	40	M	45	British
16	mmm	38	M	19	British/Cypriot
17	whistling	34	M	18	Greek
18	ahh	42	M	21	British
19	whistling	37	M	13	British
20	mmm	42	M	55	British
21	ooh	35	F	10	British
22	ooh	34	M	12	British
23	ooh	27	F	25	Greek
24	ahh	40	F	19	British
25	ahh	43	F	25	British
26	mmm	38	F	26	British
27	mmm	39	M	17	British
28	mmm	35	F	20	British
29	whistling	34	M	50	American
30	sss	35	F	21	British
31	blowing	35	M	24	Indian
32	mmm	35	F	25	Lebanese
33	mmm	34	M	27	British
34	mmm	45	F	19	British
35	ahh	38	M	20	Colombian
36	ooh	41	F	21	British

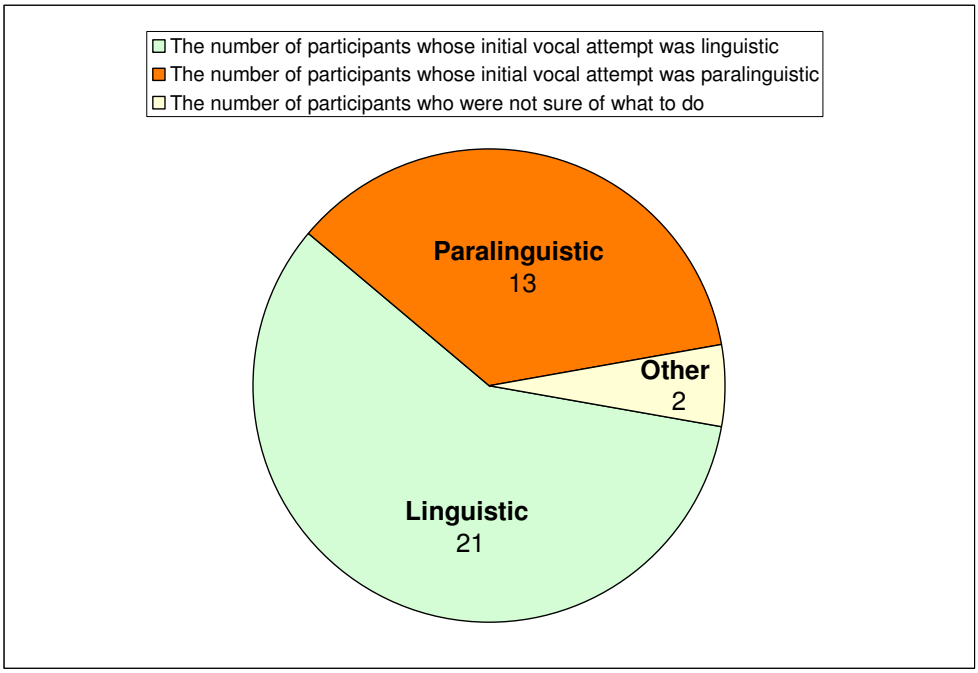


Figure 69 : Categorizing the initial vocal attempt by the participants into linguistic Vs. Paralinguistic.

The most frequent linguistic attempt was saying "hello". One of the participants (P29) sat very close to the tree and repetitively said "hello" to separately address each bulb while moving his mouth consecutively from one bulb to the next (Figure 70). Other linguistic attempts included saying "hi" and "beep", uttering commands, thinking aloud, and talking to the tree.

Examples of the commands used included the following: *"Turn lights on, turn turn"*, *"light up"*, *"lights on the top of the tree, turn on"*, "*On, off, light, Christmas, tree, bulb, on"*.

Examples of phrases uttered while trying to switch the bulbs by thinking aloud included: *"Perhaps if I speak more loudly or more softly the bulbs will go higher"*. Another example was the following: *"Does it work with volume?"* (while raising the volume of her voice), *"or high pitch?"* (while raising the pitch of her voice).

Examples of phrases uttered while talking to the tree included: *"hello there Christmas tree. How are you today? You are not very bright? You are not very cheerful? Come on. Light up a bit*".

Figure 70 : Participants sitting very close to the tree and vocalizing.

Some participants' conception of voice-control seemed to be very attached to the verbal elements of voice to the extent that, despite my instruction to generate a continuous voice, they chose to generate it in the form of continuous speech. One of the participants (P20), for instance, succeeded in lighting up the bulbs while quickly saying: *"you've got to keep going and going and going and going and going and going..more and more and more and saying saying saying this that the other A B C D E F G H I J K L M N 1 2 3 4 5 6 7 8 9 10 11 12 .....A B C D E F G H I J K L M"*. Another similar but even more remarkable and unexpectedly effective attempt was by another participant (P9) who decided to tell the tree in an extraordinarily fast tempo about what she did during that day: *"It is very difficult to talk without stopping, without stopping, without stopping but what did I do today, I have been working on the computer. I have been taking the dog for a walk. I went out for lunch today. Also I went to the dentist……..".*

When later given the instruction “use your voice, but without using words, and aim to light the highest bulb on the tree”, the most frequent sounds that participants made included: “oooh”, “mmm”, “ahh”. Some participants also hummed, whistled or blew into the microphone. Other sounds included: “ la la la”, “laaa”, “eee”, “ouuu”, “beee”, “deebeedebedebede”, “ba ba ba”, “be be be”, “dum dum dum”, “ra ra ra”, “da da da”, “ho ho ho”, “ha ha ha”, “du du du”, “ru ru ru”, “bmm, bmm, bmm”, “ma ma ma”, “na na na”, and “oufff, oufff, ouff”.

Some participants also displayed unexpected patterns of interaction. They coughed, cleared their throats, clicked their tongues and snapped their fingers. Many participants also clapped (Figure 71).

Figure 71 : Many participants clapped to interact with the tree.

One participant (P39) persistently explored various forms of input until he discovered a trick to light up all the bulbs on the tree. He held the microphone very close to his mouth and started blowing by exhaling loudly and also by inhaling loudly. Thus, the microphone was continuously detecting the sound input. Unlike some participants who stopped because they *"ran out of breath"*, this participant gracefully utilized his running out of breath as an input. It is not surprising, thereafter, that he was one of the only two who preferred blowing. Another participant (P8) used the same technique but with 'ahh' rather than blowing. He inhaled loudly after every prolonged generation of the 'ahh' sound.

An unexpected observation was that during the vocal endurance test, the pitch and volume of vocalizations seemed to increase as participants lit higher bulbs on the tree. Although *Expressmas Tree* was designed to use the duration of vocalization to cause the bulbs to react, it seemed that the bulbs also had an effect on the characteristics of voice such as pitch and volume. This unforeseen two-way voice-visual feedback calls for further research into the effects of the visual output on the vocal input that produced it.

One thought-provoking observation was that there were some remarkable, and sometimes unusual, gestural and postural accompaniments to the prolonged paralinguistic vocalizations. A number of participants sat with their mouths very close to the bulbs while vocalizing. One participant (P8) performed a circular hand gesture before each vocalization, which made him seem to be about to practice some magic (Figure 72). Many participants had their hands in their pockets while vocalizing (Figure 73). This seems to be related to the common but under-researched hands-in-pockets posture that many people adopt while whistling. Many participants performed rhythmic gestures while generating a continuous stream. These were in the form of moving the hands from side to side in one case, moving the hands up and down in another case, and standing on the tip of the toes (Figure 74) and moving the body up (as the vocalization continues and as higher bulbs light up) and down (as the bulbs switch off upon stopping).

Mikael Wiberg [2006] undertook an empirical evaluation in which he video-recorded users attempting to understand and explore an interactive lamp which was designed to be switched on and off, brightened and dimmed, by non-contact hand movements. Intriguingly, he observed that users tended to form their actions into graceful hand gestures – though this was technically unnecessary.

There was a similar element of gracefulness in participants' interactions with *Expressmas Tree,* which seems to confirm Wiberg's view of *"graceful interaction"* as: "*any way in which a person interacts with an intelligent environment that is both effective and effortless to him/her while at the same time appearing to be rational and elegant to anyone else observing him or her while interacting with the system"* [Wiberg, 2006].

Figure 72 : This participant performed a circular hand gesture before each vocalization, which made him seem to be about to practise some magic.

Figure 73 : Participants moving their hands rhythmically while vocalizing.

Figure 74 : A participant standing on the tip of her toes and holding her stomach while vocalizing.

The most intriguing postural accompaniment of many participants' vocalizations involved tilting the body backwards and forwards (Figure 75) or in some cases from side to side, often in a rhythmical fashion. Many participants clearly tilted backwards when they approached the end of their prolonged vocalization. This accompaniment of *excessive* vocalization may be related to Darwin's observation that *"During excessive laughter the whole body is often thrown backward"* [Darwin, 1872:206]. Since the backward tilt of the body was mostly observed when participants approached the end of their vocalization, in other words when they nearly ran out of breath, the postulation that *"the backward tilt of the head facilitates the forced exhalations"* [Ruch and Ekman, 2001] may be an explanation for this observation.

Figure 75 : Participants tilting their bodies forward (right) and backward (left) while vocalizing.

The vocal endurance test results revealed that among the six tested vocal expressions, 'mmm' was, on average, the most prolonged expression that the participants generated, followed by 'ooh', 'ahh', 'sss', whistling, and then blowing, respectively (Figure 76). These results were based on finding the duration of each participant's most prolonged attempt for each type of vocal expression and then comparing it with the average duration of all participants' most prolonged attempts. The following equation was formulated to calculate the efficiency of the vocal expression:

Vocal expression efficiency = duration of the prolonged vocalization –duration of silence after the prolonged vocalization

This equation is based on postulating that the most efficient and less tiring vocal expression is the one that the participants were able to generate for the longest period and that required the shortest period of rest after its generation. Accordingly, 'mmm' was the most efficient and suitable input for an application that requires maintaining what I refer to as *vocal flow*: vocal control that involves the generation of a voice stream without disruption in vocal continuity. The expressions 'ahh' and 'ooh' were also considerably and almost equally efficient while 'sss', 'whistling' and 'blowing' were the least efficient.

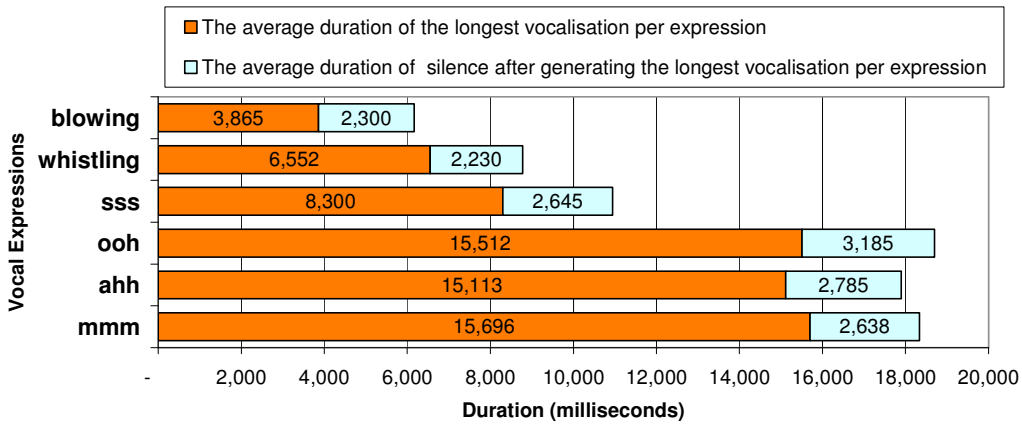


Figure 76 : The average duration of the longest vocal expression by each participant in experiment 2. Sounds are arranged on the ordinate from the most efficient (bottom) to the least efficient (top).

During the vocal endurance test and when instructed to generate 'sss', a couple of participants decided to generate 'zzz' instead of 'sss' as they seemed to have felt that, at a given volume, the voiced 'zzz' can be more easily prolonged than the voiceless 'sss'. This led to my postulation that voiced sounds such as 'mmm', 'ahh', and 'ooh' can probably be generated for longer durations than un-voiced sounds. Some participants also varied the pitch of their voices while vocalizing. Vocalizing melodically seemed, in some cases, to have the effect of prolonging the vocalization, perhaps by making it less tiring or tiresome. This possibility, however, calls for further research and if it is proven, it might explain why the participants in the experiment of Sporka and his colleagues found humming or singing less tiring than whistling [Sporka et al., 2005].

Other factors that can also probably affect the duration of vocalization include shyness, physiological limitations such as asthma, and possibly the commonly used phonemes in the vocalizer's mother tongue. The participant (P4) who generated the least prolonged 'ooh' was one of the participants who preferred whistling and who indicated that he stopped *"because he was embarrassed"*. The short duration of his vocalization may be due to his shyness. The participant (P12) who generated the least prolonged 'ahh' was German. His short vocalization may probably be related to the German production of vowels which differ from English vowel sounds in that there is "*no off-glide from one vowel sound to another. The short vowels are pronounced in a very clipped manner. German diphthongs (vowels which do glide from one vowel sound to another) are also pronounced in a clipped manner, never drawled*" [Smith, 1997]. The participant (P3) who generated the least prolonged 'mmm' was Taiwanese. The average duration of all the vocal attempts of a Chinese participant was also low relative to other participants. This too might be related to the Mandarin staccato speaking style. The exploitation of the possibilities of lengthening vowels differs from a language to another and may perhaps have an effect on the ability of a person to vocalize for a long duration.

On the other hand, the results of the preferences test revealed that 'ahh' was also the most preferred in this experiment, followed by 'mmm', whistling, 'ooh',

blowing, and 'sss' (Figure 77). Thirteen of the thirty-six participants preferred 'ahh', eleven preferred 'mmm', five preferred 'whistling', four preferred 'ooh', two preferred blowing, and only one preferred 'sss'. It is worth mentioning that among those who preferred 'ahh', males were over-represented. This coincides with the findings of the first experiment. It is remarkable to note the vocal preference of participant 5 who was noticeably very outgoing and who evidently had the lowest shyness score (shyness score = 26). His preference and pattern of interaction, as well as earlier observations of interactions with *sssSnake*, led to the inference that many outgoing people tend to prefer 'ahh', or even 'ooh', as input. Unlike whistling which is voiceless and involves slightly protruding the lips and unlike 'mmm' which does not involve opening the mouth, 'ahh' is voiced and involves opening the mouth expressively. Out of all participants, five made clear social references and uttered and/or wrote statements that indicated social discomfort. Three (P 4, 7, and 17) of these five preferred whistling. One of the participants (shyness score = 36) tried to utter 'ahh' but was too embarrassed to continue and he kept laughing during every attempt. He stated that he preferred whistling the most and that he stopped because he *"was really embarrassed"*. Another participant (P 7, shyness score = 44) also stated that she preferred whistling because while generating it she *"felt slightly less silly"*. When asked to use her voice to aim at the highest bulb, she laughed shyly, said: *"Does it work with volume? Or high pitch? Oh God! This is embarrassing"*, and whistled. The sample, however, is too small for these results to be statistically significant.

These participants' preference seems to verify the earlier hypothesis that shy people tend to prefer whistling to interact with a voice-controlled work. Another unexpected, though very reasonable, finding is that shy players were over-represented among those preferring 'mmm', most likely because it is *"less intrusive"* to generate and is *"more of an internal sound"* as one of the participants (P 34) who preferred it stated. This participant stated that she stopped because she was *"uncomfortable with using voice in such a quiet atmosphere"* and also said that she would probably feel more comfortable to make noises in public and in the presence of her friends than to make them in a

quiet room in front of an invigilator. The relationship between shyness and preferring whistling and 'mmm' and that between outgoingness and preferring 'ahh' is also reflected by the preferences of participants who had the maximum and minimum shyness scores. The participant (P34), who had the highest shyness score (shyness score = 45) among all, preferred 'mmm'. The participant (P7) who had the second highest shyness score (shyness score = 44) preferred whistling. Conversely, participants who preferred the vocal expression 'ahh' had the lowest average shyness scores in both experiments 1 and 2. In experiment 1, the least shy participants (P1 and P9, shyness scores = 29 and 21 respectively) preferred 'ahh' (Table 5). In experiment 2, the least shy participant (P5, shyness score = 26) also preferred 'ahh' (Table 6). However, since the experiment was in a closed room, there might have been some shy participants who preferred 'ahh' in that setting and who might not have the same preference if it was a public setting.

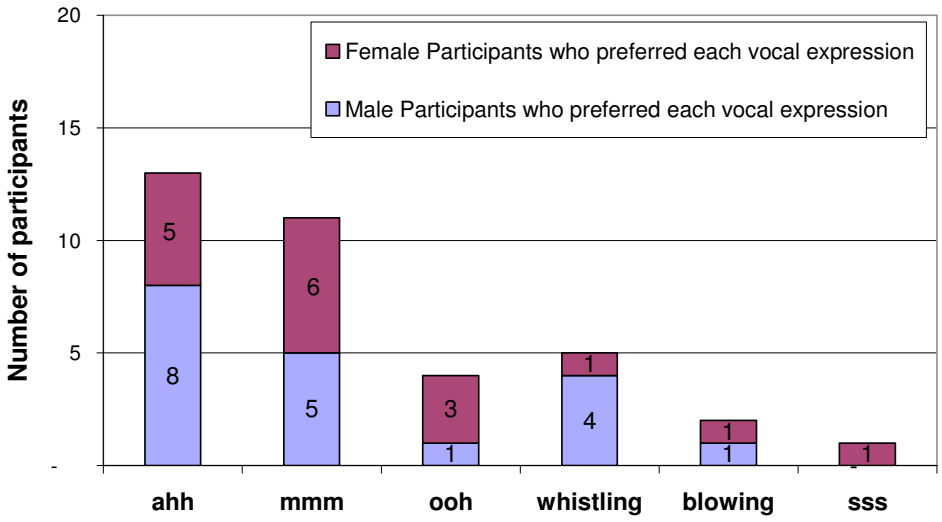


Figure 77 : Correlating the preferences and genders of participants in experiment 2.
The sound indicated is the one ranked as most preferred by each participant.
Sounds are arranged on the abscissa from the most preferred (left) to the least preferred (right).
according to the combined results of both experiments (Figure 78).

The reasons for preferring 'mmm' that participants stated included: *"because it's a pleasant sound and I could also breathe easier through my nose"*, *"it was more comfortable and I could sustain it longer. It irritated my throat less as I*

have a cough and cold at the moment", "less intrusive to make the sound…more of an internal sound". The reasons for preferring 'ahh' included: *"because it makes more noise", "it comes more naturally to say ahh out loud", "because it allows me to use my breath in a controlled way and to chose a pitch that comes naturally from my belly.", "can't whistle or do a very good sss sound"*. Unlike whistling which some people cannot generate and unlike 'sss' which people who have a lisp cannot produce accurately, the 'ahh' and 'mmm' sounds, which were the most preferred, were identified by some users as being *"natural"* as opposed, presumably, to "contrived". They are present from birth without any need to acquire the skills for producing them as in the case of 'sss' and whistling. It is therefore important to mention at this point that although five of the thirty-six participants preferred whistling, ten participants said that they could not whistle. Self-evidently, whistling is not an ideal primary input for any application, unless the objective is to teach users to whistle. It is also important to note that the preferences of the hearing-impaired may possibly differ from those with normal hearing. Before being restricted to the six tested sounds, one of the participants (P25) who was deaf preferred to utter "laaaa la la la". Her preference seems to have a connection with the "paaa" that the deaf participants, as will be discussed in the next chapter, preferred while controlling the screen-based version of *Expressmas Tree*. The commonality between the "paaa" and "laaaa la la" possibly lies in the visibility of their onsets to the deaf as they both start with consonants that are either bilabial (lips touch each other) as in "paaa" or dental (uttered with the tongue against the upper teeth) as in "la la la".

Combined results from both experiments revealed that nineteen participants preferred 'ahh', fourteen preferred 'mmm', six preferred 'ooh', five preferred whistling, two preferred blowing, and one preferred 'sss'. Most (twelve) of the participants who preferred 'ahh' were males, and most (eight) of those who preferred 'mmm' were females (Figure 78).

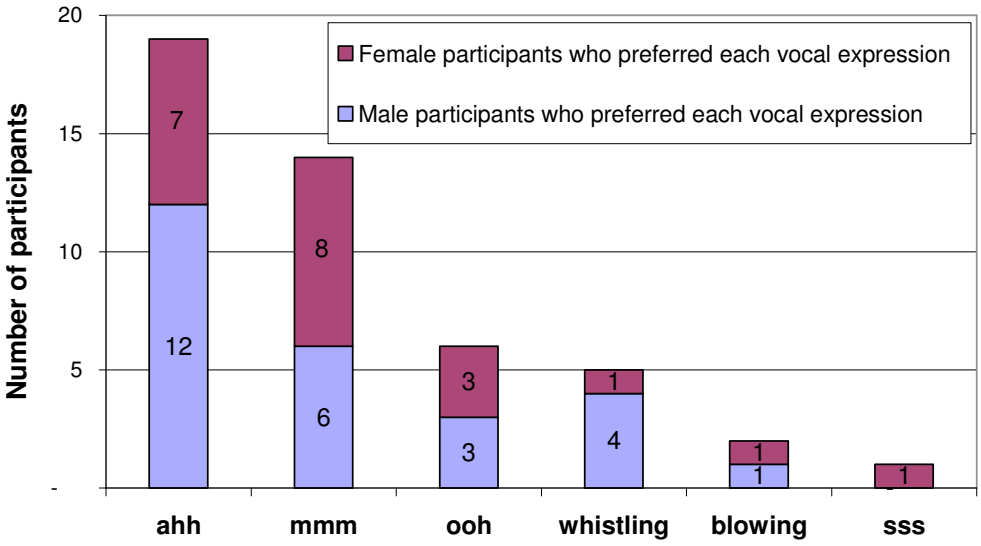


Figure 78 : Correlating the preferences and genders of participants in experiments 1 and 2. The sound indicated is the one ranked as most preferred by each participant. Sounds are arranged on the abscissa from the most preferred (left) to the least preferred (right).

4.2 Analysis

The results suggest that shy players are more likely to prefer whistling (possibly because it is a voiceless sound) and 'mmm' (possibly because it does not involve opening the mouth) while outgoing players are more likely to prefer 'ahh' (possibly because it is a voiced sound). It was also evident that many females preferred 'mmm' while many males preferred 'ahh'. The results also revealed that 'mmm', 'ahh', and 'ooh' are easier to generate for a prolonged period than 'sss', which is in turn easier to prolong than whistling and blowing. Accordingly, the vocal expressions 'mmm', 'ahh', and 'ooh', are more suitable than whistling or blowing for interactions that involve prolonged or continuous control. The reason could be that the nature of whistling and blowing mainly involves exhaling but hardly allows any inhaling, thus causing the player to quickly run out of breath. This, however, calls for further research on the relationship between the different structures of the vocal tract (lips, jaw, palate, tongue, teeth etc.) and the ability to generate prolonged vocalizations. It is hoped that these observations will help future developers anticipate vocal preferences and patterns in this new form of interaction.

5. Roles for Paralinguistic Voice Input

5. Roles for Paralinguistic Voice Input

While working on the projects described in the previous chapter, many other issues of potential importance arose, including the possible roles and applications of paralinguistic voice input. This chapter summarizes some of the ideas that arose and which seem worthy of further investigation, by myself or by others, and gives an account of one additional evaluation that I undertook.

5.1 The Role of Paralinguistic Vocalizations in Inducing Cathartic Experiences

The use of voice as an input might create what could be called *vocal disinhibition* or *vocally-induced catharsis*. For some players, who played *Sing Pong*, *sssSnake*, and *Expressmas Tree* 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. 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* and *sssSnake*, in this context, also provide an outlet for emotions and encourage social and vocal disinhibition through expression and discharge of embarrassment. The fact that they both incorporate a two-player experience encourages players to make voices in public more than it would if they were 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 considered of negative consequences is here encouraged and rewarded by winning the game. The vocal disinhibition and the vocally-induced catharsis that players seem to experience

are therefore a result of the negative becoming positive and the disapproved becoming approved in the context of the game. In such games, 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, *Sing Pong, sssSnake,* and *Expressmas Tree* involve a hidden form of social interaction and of social learning through which people learn by watching others who are rewarded. They also make certain actions and expressions more tolerable socially, and they increase the viewer's willingness to try new interactive experiences.

Moreover, the laughter and joy that playing such voice-controlled games brings to the players and audiences may also be cathartic to both. The screaming that these games may involve may also be considered cathartic. Some people scream when angry or stressed in order to release their negative emotions. Inspired by this thought, Kelly Dobson, a researcher at the MIT Media Lab, developed *Screambody*; a portable wearable body organ with an opening at the top for the user to scream into [Dobson, 2004]. *Screambody* silences the vocalizations of the user when the user is in a public area, and records them for later release in a private area (Figure 79).

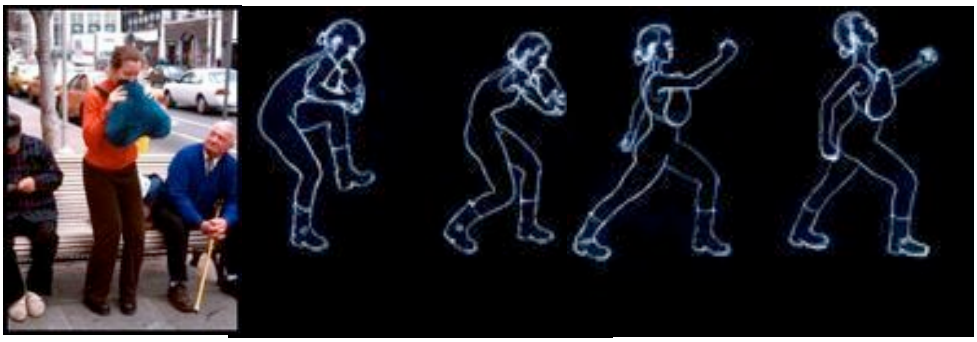


Figure 79: *Screambody*; *"A portable space for screaming"* [Dobson, 2004].

Screaming is not the only form of expression that may be considered cathartic. Some researchers believe that singing is also cathartic and that it is an artistic and biological means of emotional expression and release [Beeman, 1998]. Since the paralinguistic dimension of singing is what distinguishes it from normal everyday speech, the cathartic effect of singing may arguably be at least partly attributed to this dimension. The next section investigates the role of the paralinguistic dimension of singing in expressive communication.

5.2 The Role of Paralinguistic Vocalizations and Singing in Expressive Communication

It has also been reported that singing may *"unlock the brain"* and help in *"relearning communication skills"* [Elliott, 2005]. It may help in the treatment of Alzheimer and dementia as well as Parkinson's disease, strokes, and head injuries. Chreanne Montgomery-Smith who believes that memory can be improved by singing has recently founded "Singing for the Brain" sessions. She observed that these sessions enhanced participants' cognitive skills and that singing seemed to enable some speech-disabled participants to access words through the melody [Montgomery-Smith quoted in Elliott, 2005]. Similar evidence about the role of singing in enhancing memory and communication skills were found by Loewy [2004].

Singing was found to be of therapeutic benefit to neurologically impaired patients and of significance in enhancing the mental and physical health of the aged [Unwin et al., 2002]. However, such research on the effect of singing on the singer rather than the listener is still comparatively thin on the ground [Unwin et al., 2002].

Diane Austin, a music therapist who has undertaken a qualitative study of music psychotherapy, developed a vocal improvisation technique that she termed *"vocal holding"*. This technique involved exploring the voice through breathing, improvising, singing, and attuning the client's voice characteristics with those of

the therapist. According to Austin, this technique was useful to people who suffered from eating disorders and those who were sexually abused [Austin 2003 quoted in Loewy, 2004].

Singing and vocalizing are also argued to have a "*vibroacoustic stimulation"* effect on the brain. Jindrak and Jindrak [Jindrak, 1986] claim that vocalizations cause the vocal folds, the walls of the mouth, and the pharynx to vibrate. These vibrations are then transmitted by the interstitial fluid to the skull and lead to its vibration. The vibration of the skull causes the sphenoid bone to vibrate which in turn causes the parietal bone to vibrate. The parietal bone causes the vibration of the arachnoid membrane. In his explanation of the Jindrak postulate, Olav Skille compares the arachnoid membrane to a *"plastic bag"* that contains the brain and the cerebrospinal fluid (CSF). It, along with the brain-blood-barrier (BBB) prevents the blood from mixing with the brain. The arachnoid membrane, however, contains villi or granulations which are the only path through which the CSF can leak to the blood. The CSF functions mainly in the process of diffusion, transmitting nutrients into the brain and cleaning molecular wastes such as carbon dioxide (CO2). According to Jindraks' postulate, the arachnoid membrane with its granulations is similar to a sieve through which the passage of material (in this case the flow of the CSF) can be significantly facilitated by vibrations [Skille, undated]. Based on this theory, vocalization or *"vibroacoustic stimulation"* speeds up the circulation of the CSF and the elusion of waste from the brain. I do not have adequate knowledge to be sure of the validity of any of these claims.

5.3 The Role of Voice-Visualization in Voice and Speech Therapy

Medical involvement in the human voice dates back to the fifth century B.C since Hippocrates wrote about the significance of the laryngeal system in voice production [Harman 1991 quoted in Arizona Health Sciences Library, 2006]. According to Harman, the role of voice as a means of expressing emotions was first realized by Aristotle, but its scientific value was first investigated by

Claudius Galen who is considered the founder of voice science. The writings of Leonardo Da Vinci during the Renaissance also contributed to earlier speculations on voice production. Rhazes in Baghdad, on the other hand, made the earliest attempts to explore voice disorders and to advocate voice training [Harman 1991 quoted in Arizona Health Sciences Library, 2006].

Traditional forms of speech and voice therapy involve looking at the therapist's face and lips in order to establish correct voice production techniques and acquire proper articulation skills. Some forms even involve feeling the therapist's throat as a way of overcoming the invisibility of voice and perceiving its imperceptible production techniques. Other ways included the use of systems that consisted of a special indicative lamp or a deflective meter needle [Öster, 1996]. Oscilloscopes have also been used to explore further acoustic dimensions of speech but their usefulness was found to be limited because the visual feedback was difficult to understand, delayed, and unappealing to children [Öster, 1996].

A remarkable improvement in speech therapy was made in 1976 when Nickerson and Stevens developed the first *"computer-based speech training system"* [Öster, 1996]. Since then, a number of effective voice-visual training tools have been developed to assist dyslexic, hearing-impaired, speech-impaired, and asthmatic patients.

Computer-based speech and voice therapy is nowadays gaining increasing rehabilitative credibility and clinical popularity [Walker, 2006]. Computers, however, are not yet cognitively skilled enough to connotatively interpret and detect errors in continuous speech without human involvement [Walker, 2006].

One of the few interactive therapeutic speech training systems that exist today is *Video Voice*. This system consists of a number of entertaining therapy games that improve voice production and articulation by encouraging various speech activities. Some of the games may help improve pitch and volume control as well as breath control [Arizona Health Sciences Library, 2006]. Most of the games are designed to be used by the player in the presence of a therapist who

examines and discusses the speech errors that the system allows the therapist to capture and freeze. The speech patterns of which the system generates a variety of visual representations allow the patient and the therapist to judge voice production instantly.

Another therapeutic application is *SpeechViewer II* which comprises fifteen applications developed to improve the user's awareness of pitch, loudness, timing, and voicing. By matching a spectrogram of the sounds produced against a visual target model on a split screen, the hearing-impaired child knows if the sound has deviated from the therapist's correct production (Figure 80). Colors are used to differentiate between voiced sounds which are displayed in red and voiceless sounds which are displayed in green [Öster, 1996].

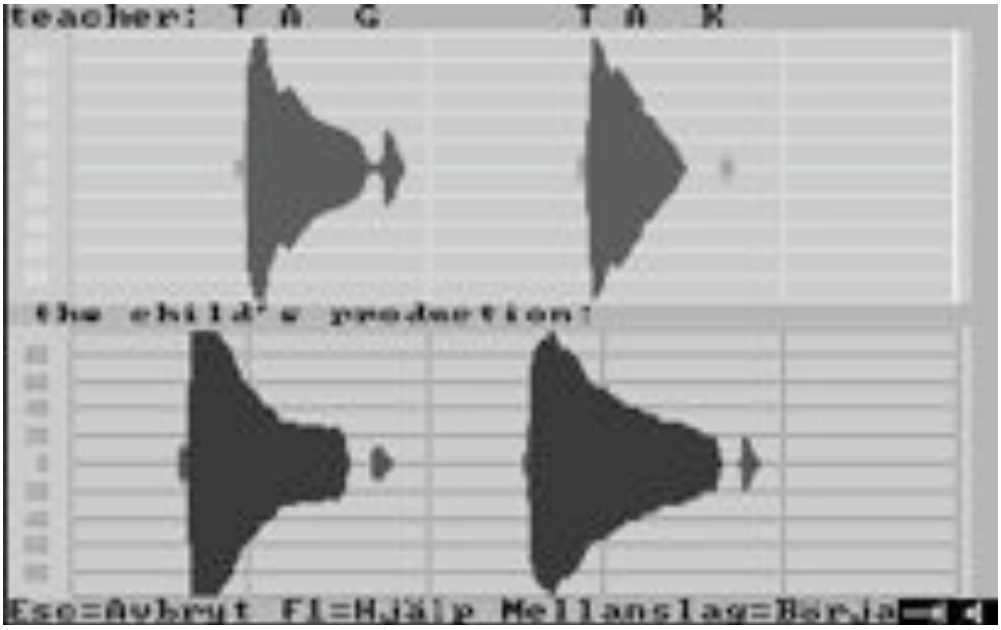


Figure 80: The upper part displays the therapist's model of a correctly produced vocalization. The lower part displays the child's attempt [Öster, 1996].

The real-time visual feedback allows for the visual exploration of auditory errors and voice disorders and the discovery of the *"kinesthetics"* [Walker, 2006] of voice production and articulation. It allows the eye to complement the ear in quickly assessing and correcting the otherwise invisible vocal impairments. Seeing the computer's feedback allows the impaired to be more involved in

judging their own speech errors and makes it less likely for them to argue with the therapist or even with the computer about the results [Walker, 2006].

Using voice-visualization in therapeutic sessions may also involve one major advantage over traditional therapeutic sessions that only involve a therapist. This advantage is that a voice-to-vision application may possibly be more engaging and appealing to the patient than a face-to-face session with a therapist.

The next section highlights the need for gathering a rich set of requirements and preferences for creating voice-controlled applications that are accessible and engaging to the deaf.

5.4 The Role of Paralinguistic Voice Input in Augmenting Awareness of Voice Characteristics in the Hearing-Impaired

Explaining voice characteristics to the deaf is not an easy undertaking for their instructors. Furthermore, many existing strategies for conveying these characteristics and teaching the deaf how to perceive them do not seem efficient, especially when dealing with the concept of pitch. As a result, some deaf people are not fully capable of differentiating between voice characteristics and some cochlear implant recipients *"often confuse the concept of pitch with loudness"* [Fearn, 2001]. Paradoxically, this does not only apply to the deaf, but some of the hearing cannot make a clear distinction between the concepts of pitch and loudness [Ma, 2001]; when asked to generate a higher-pitched sound, many generate a louder sound, which may of course be a linguistic rather than a cognitive confusion. These were some of the many reasons that led to my exploration of additional approaches to the visual representation of voice.

I employed *SpitSplat* and *Expressmas Tree* in analyzing the interaction patterns of seven deaf children. I slightly modified the screen version of *Expressmas Tree* to test it on deaf users. Instead of the audio feedback employed in the original version, I added a star that smiles when the user lights up a flashing

bulb, frowns when the user misses it, and displays a bigger smile when the user wins (Figure 81). The aim was to explore the potential role of paralinguistic vocal control of interactive media in enabling the deaf to have a greater understanding of voice and to offer their instructors more efficient and engaging strategies for explaining voice characteristics.

There are many reasons why non-speech voice-controlled games could be useful to the deaf as well as to the hearing. According to Goberis and Loraine *"Most young children, whether implant users or not, tend to use a voice that is too loud"* [Goberis and Loraine, 2006]. As a result, several loudness charts have been designed to aid the deaf in learning the variations between loud and soft voices. These usually contain a small figure representing a soft sound, a medium figure representing a medium sound, and a larger figure representing a loud sound. However, there are hardly any well-designed charts that establish an understanding of pitch or duration. Moreover, charts are unlikely to be as engaging, appealing and memorable as an interactive experience.

Figure 81: An illustration of a star was employed as a visual feedback signal in the virtual version *of Expressmas Tree.*

I carried out a small-scale experiment at the Prince Sultan Bin Abdul Aziz Al Saud Hearing and Speech Development Centre in Bahrain (Figure 82). The evaluation involved observing, writing field-notes, and video-recording seven children with varying degrees of hearing loss that ranged from mild (hearing threshold => 25 dB) to profound (hearing threshold => 95 dB). Their ages ranged from three and a half to seven years. They all had normal cognitive capabilities. Four children had cochlear implants and the rest wore hearing aids. Two were females and the rest were males.

Figure 82: Hearing-impaired children playing *SpitSplat*.

The initial plan was to evaluate each participant individually. I started with Zahra, a seven-year-old female (mild hearing loss) who had a cochlear implant. The instructor, Afrah Al Fardan, was present during the study to help in instructing the children. I started by demonstrating how to play *SpitSplat* by uttering 'ooh' into the microphone. When handed the microphone, Zahra started blowing into it rather than vocalizing. It seemed that, to Zahra as a deaf person, rounding or protruding the lips was probably associated with blowing rather than

with producing an 'ooh' sound. The instructor, hence, held Zahra's hand close to her mouth. She produced an air stream by blowing on her hand and instructed Zahra to avoid doing so. She then placed Zahra's hand on her own throat, generated an 'ooh' sound, and explained that that was the right way to play the game. Placing Zahra's hand on her instructor's throat allowed her to sense the vibration of the instructor's vocal folds. She immediately realized that she was expected to produce a voiced sound rather than a voiceless air-stream. Zahra, however, seemed too shy to generate a loud voice and was therefore not very willing to vocalize in the presence of an invigilator. I then allowed her to play *Expressmas Tree*. Despite receiving instructions to vocalize louder, Zahra was still either blowing into the microphone or producing a very soft 'ooh'. This led the instructor to demonstrate how to light the sequence of bulbs and stop at the flashing bulb by uttering a prolonged 'Paaaa'. This vocal expression seemed to be more convenient because unlike 'ahh' or 'ooh', it started with the bilabial (lips touch each other) consonant 'P'. Seeing the lips come together enabled Zahra to 'see' the onset of voice, which turned out to be an important factor as duration was the voice characteristic in question. Hence, she started interacting properly and loudly with the game.

Due to Zahra's shyness and the time it took to explain the game-play to her, it was decided that user-testing all the participants together in one room rather than singularly would be more efficient and encouraging. I invited the other six participants and the instructor explained and demonstrated *Expressmas Tree*. She then allowed them to take turns in playing.

Explaining that they should stop vocalizing upon reaching a flashing bulb was a very difficult task. Four of them seemed to have eventually realized it, while the rest confined themselves to watching how the bulbs lit sequentially as a visual reaction to their vocal input.

Figure 83: Hearing-impaired children playing *Expressmas Tree*.

One of the remarkable behaviors that I observed was that Hussain, a five-year-old male (mild hearing loss), who wore a hearing aid, tapped the floor with his foot while vocalizing (Figure 83). It seemed (to the instructor) that he performed this complementary behavior either to help him prolong his vocalization, or to help him estimate the required duration by synchronizing his tapping with the sequential lighting of the bulbs. Another interesting behavior was that Abdullah, a six-year-old male (mild hearing loss), who wore a hearing aid, insisted on vocalizing or pointing at the flashing bulb while others played in order to help them. He seemed to enjoy vocalizing simply for the sake of vocalizing. I also noticed Ali, who was a three and a half-year-old male (mild hearing loss) with a cochlear implant, blocking his ears. Ahmed, on the other hand, who was a four-year-old male (profound hearing loss), with a cochlear implant, seemed to have clearly understood the game. He even cupped his hands in front of his mouth to better direct and concentrate his voice. A notable observation was that while most of the children enjoyed watching, Zahra rewarded those who performed well by clapping.

Although both games did involve visual rewards, the children seemed to find these feedback signals less enticing and eye-catching than I anticipated. This was probably because they were overshadowed by the central game-play elements. Displaying highly animated feedback signals may solve this problem and counterbalance the attention given to the central elements. This calls for further research, in this context, on the use of visual signals as an alternative to audio signals for the deaf.

Another aspect that needs to be further researched is the appropriate strategies for representing voice characteristics. The children's instructor mentioned that she uses a tall versus a short tree to represent long versus short vocal duration. She also represents duration by comparing it with the child's mother's hair length. As for pitch, she usually represents low pitch by a man and high pitch by a woman. She finds that loudness is the easiest characteristic to represent, and that representing a soft voice by a small light-colored circle and a loud voice by a large intense/dark-colored circle is a very efficient strategy.

I hope that my study will further progress towards understanding the most efficient visual mappings of voice characteristics and the most convenient vocalizations for the deaf. The efficiency in choosing the appropriate visual representations of voice characteristics, however, is not the only factor that contributes in the development of a comprehensible voice-controlled application. Another important requirement for a successful voice-visual aid is immediacy: *"The visual feedback of the child's voice and articulation should be shown immediately and without delay"* [Öster, 1996]. The next section, therefore, addresses and calls attention to the importance of investigating the effect of latency on the perception of causality in voice-visual and *voice-physical* media.

5.5 The Role of Vocal Telekinesis in the Perception of Causality in Interactive Media

The movement of the coin and snake in reaction to voice induced forms of causal perception in many players' minds during the exhibition of the screen-based version of *sssSnake*. Most players asked: *"what causes the sss to move the snake, and the ahh to move the coin"*. This recurring question prompted me to wonder about the cognitive aspects involved in using voice as an input and to study users' attribution of causality in voice-controlled applications. No exhaustive studies of how users of an audio-visual or voice-visual installation attribute causality relationships appear to exist. When voice is used as a causal input, it is crucial to understand the perception of causality.

Michotte defines causal perception as the establishment of a causal link between two events whereby one event *"produces"* another [Michotte, 1963]. Most of our everyday uses of our non-verbal aspects of voice involve its generation as an effect; *"just as tears or groans are an effect of sorrow, so laughter is an effect of joy"* [Aquinas, 1947]. The act of generating non-speech voice as a cause is unfamiliar, in comparison, and calls for the investigation of the perception of voice as a cause in voice-controlled media.

In *sssSnake*, I noticed that the causal relationship between the vocal input and visual output, which is maintained by a responsive interface and a near-immediate output, improves the user's engagement levels with the application and augments immersiveness. As discussed in the first chapter, the use of paralinguistic vocal input supports real-time output in comparison with linguistic input which involves checking the recognition result against a previously stored model.

However, I had further observations about causality while testing vocal control of the plotter during the first phase of developing the *voice-physical* version. Technical limitations obliged me to use a parallel interface rather than a serial one. This in turn, combined with issues of buffering, meant that it was

impossible to abort data transfer once the plotter head started moving. The resulting latency caused the plotter head to keep moving for a few seconds after the test subject stopped vocalizing. This also caused the plotter head to keep moving towards a certain direction for a few seconds even when the test subject has already vocalized into another microphone in order to change the direction. During that early stage of developing the game, test subjects could not directly perceive the causal relationship between the voice input and the direction of the plotter head. They did, however, perceive the causality in the voice moving the plotter head. This observation reflects the relationship between latency and the perception of causality.

In the 1950s and 1960s, Albert Michotte carried out many studies about the perception of causality involved in the interaction between visual events. He found that when the delay between events exceeded 150 ms, they were no longer perceived to have a causal link [Cavazza et al., 2005]. In my work, however, the causal relationship does not involve an interaction between visual events entailing two physical objects hitting or contacting each other but rather an interaction between an invisible vocal event and a visual event. A *voice-physical* event is not an everyday event to which the observer's mind is used. This could perhaps mean that latency may be of even more critical consequence in the perception of causality in *voice-physical* events than in a physical-physical event.

Most causality experiments to date have concerned a physical object causing another physical object to move, or a physical event causing sound. As no causality studies appear to exist in the voice-physical area addressed by this research, my future work will involve designing a usability test of the perception of causality in voice-controlled applications, with particular emphasis on the issue of latency. I hope that this research will lead to systematic studies about the perception of causality in voice-controlled applications.

5.6 The Role of Paralinguistic Vocalizations in Transforming Users into Performers

In *Sing Pong, sssSnake*, and *Expressmas Tree* all devices are hidden and the interface extends beyond the screen. It embraces the players as well as the physical space allocated for movement and tracking. Unlike conventional interfaces, the button that initializes *Sing Pong*, for instance, is a round mat positioned at a specific location on the floor. When players stand on the mat, their shadows appear in the middle of the projected screen and the game is initialized. From a spectator's point of view, the performers are an integral part of the game. Their shadows are images, their voices are sound effects, their movements are animations, and their interactions are events. They do not need to execute any keyboard commands or even to conventionally 'use' the computer. These players are performers who are interacting with the computer *"not as a tool, but as a medium"* [Laurel, 1991].

Sing Pong involves a quasi-theatrical setting. Two mats are placed on the floor to define each performer's virtual stage (Figure 29). The lighting and the design of the space are important aspects of the game. *sssSnake* and *Expressmas Tree* as well as *Sing Pong* require a relatively anechoic space. The freedom of movement and expression and the naturalness of the vocal interaction are comparable to the actions that are usually demonstrated in a performance. Such paralinguistic forms of interaction as well as the display of kinaesthetic and coordination skills are essential theatrical elements, and can be thought of as theatrical signs such as *"gesture, color, scenic elements, or paralinguistic elements (patterns of inflection and other vocal qualities)"* [Laurel, 1991]. Moreover, the employment of expressive voice-visualization and body movement as means of controlling games can involve players as well as spectators in highly immersive experiences. According to Laurel *"Tight linkage between visual, kinaesthetic, and auditory modalities is the key to the sense of immersion that is created by many computer games, simulations, and virtual-reality systems"* [Laurel, 1991]. I have aimed to offer a similarly tight linkage

by harnessing the relative immediacy of non-verbal as compared to verbal control.

Voice-controlled games provide a promising option for the evolution of computer games into theatre games and they allow players to use human-human interaction styles with which they are already familiar and in which they are probably skilled. They could be a further step in the already existing process of transforming players into performers and computer games into computer theatre games.

The PlayStation game, *Sing Star*, for instance, is a recent innovation that allows players to perform, sing, and demonstrate their vocal skills. Players are prompted to sing along with any various songs recorded in a game disc. The singers' vocal skills are evaluated by – in a sense – measuring their pitch, tone, and rhythm.

A noteworthy collection of performance games developed by Sony is *EyeToy*: *Play* [Sony Computer Entertainment Europe, 2003] which features a motion sensitive camera that detects players' position and actions. In 'Kung foo', for instance, players are prompted to move and use their hands and legs to control the game. Watching a player move crazily in a public context can be a very entertaining show for passers-by. Thus the player is not only a player or a passive viewer, but an active performer who is part of the game.

One intriguing example of the employment of non-verbal communication in an interactive game performance is *Ghost in the Cave* (Figure 84). This collaborative game investigates the use of expressive gestural and vocal input as a basis for audience collaboration and participation [Rinman et al., 2003] (Figure 85). The game takes place in a mixed reality environment and involves two teams. Each team controls a fish avatar displayed on a large screen, and tries to navigate in a virtual environment in order to find three caves. The avatar can be controlled either by gestural input into a video camera or by vocal input into microphones. Players' voices determine the direction and speed of the avatar;

"the sound level in the left or right microphone controls the direction and the number of note onsets influences the speed." [Rinman et al., 2003:562]. Alternatively, the direction of the avatar is determined by players' arm movements, while its speed is determined by the extent of arm movement. The background music is controlled by the rest of the players' movements in front of a camera; *"One team is controlling the drums and the other team is controlling the accompaniment"* [Rinman et al., 2003:562]. The speed of the avatar is also determined by the amount of the players' movement.

A notable technique in this game is mapping the left-right direction of the avatar to the left-right channels of the microphone; detecting input signal through the right microphone steers right, detecting input signal through the left microphone steers left, and detecting input signals through both microphones moves the avatar forward.

Figure 84: *Ghost in the Cave*; an interactive performance that investigates the use of expressive gestural and vocal input [Rinman et al., 2003].

Figure 85: The audience members as well as the performer participating expressively in *Ghost in the Cave* [Rinman et al., 2003]

A very recent performance game is *Kick Ass Kung Fu*; an interactive martial arts game that uses advanced *"real-time image processing"* and *"computer vision"* [Hämäläinen et al., 2005:781] techniques which allow players to watch themselves moving, jumping, punching, and kicking virtual enemies on two projected screens (Figure 86). The two screens act as mirrors allowing the player to turn around and defeat enemies who try to attack from behind. Players' movements and jumps are exaggerated in order to make them feel superhuman. Shouting supercharges a player's virtual body:

> *"Kick Ass Kung-Fu can be seen as part of an emerging phenomenon of computer games as performances. […] The common factor is that playing a game can actually be great entertainment not only for the player, but also for people around"* [Hämäläinen et al., 2005:789].

Such performance games should therefore be viewed from the audience's perspective as well as the player's perspective.

Figure 86: *Kick Ass Kung Fu* performed on a theatrical stage [Hämäläinen et al., 2005].

Jennifer Sheridan and her colleagues refer to improvised performances which occur spontaneously and are carried out by novice performers as *"DIY performances"* [Sheridan et al., 2006]. They developed *iPoi*: an interactive system specifically designed to encourage spectators to make the transition to participants by spinning a ball around their bodies to interact with visuals.

In this context, *Sing Pong* and *sssSnake* allow players to express themselves vocally, which in turns allows for the richness of expression and demonstration that is usually experienced in live performances. These games may embrace both the novice as well as the vocally skilled player because they exploit natural and intuitive human capabilities as means of controlling them. Although the players are spontaneous untrained performers, I noticed that their first trial of *Sing Pong* during its exhibition in London resembled a form of rehearsal. We, as the developers of *Sing Pong*, became directors who instructed players on how and where to move, and how to use their voices to interact with the game. Once

players got used to the game, they rapidly cast themselves in the role of experts. Many players started to instruct their friends and show them how to play the game, and how to initialize it. One of the players pushed her friend out of her zone because she noticed that he was interfering with the movement of her paddle whenever his shadow was tracked in her area. Another player pulled her partner towards her while she was standing in the middle of the stage in order to initialize the game. Many visitors brought new friends to watch them play when they visited the show again.

Viewing the voice-controlled installations discussed above from an audience perspective leads to the inference that spectators may take on an active role and their actions around the stage can also affect the performance. As a result, the notion of audience as passive observers disappears. This exemplifies Sgouros' remark that stimulating audience participation is an important criterion for judging the effectiveness of an interactive performance [Sgouros, 2000:197]. Furthermore, the fact that any member of the audience can step into the performance area and take the place of a performer confirms Sparacino's observation that *"Many efforts of today's multimedia technologies are aimed at prompting a more active and direct participation of the public in the artwork"* [Sparacino et al., 1999:4]. Active involvement in the performance space may well lead to the formation of a digital version of *"theatre in the round"* [Farlex Inc., 2004] where the audience members can walk through the performers from different directions and vice versa during an interactive performance. Markus Montola and Annika Waern use the term *"socially expanded games"* to refer to games of which the social boundary expands to include the active involvement or participation of bystanders [Montola and Waern, 2006].

During the exhibition of *Sing Pong* in London, many visitors enjoyed watching players' amusing interactions with the game. A visitor, who was watching the performance for a while, when asked if he wanted to try the game, said that he was actually enjoying watching it. However, he could not resist trying it later on in order to beat the skilled player whom he was watching earlier. Another spectator was a child who enjoyed projecting his shadow to interfere while his

dad and sister played the game. Such an interactive performance not only builds a relationship between the performer and the game but between the performer and the audience as well. Some spectators decided to be fans who supported a certain player and opposed the other. This performance-audience relationship, where the actions of the players induced audience reactions, and the audience reactions provoked performers' actions, was an important factor in maintaining audience involvement.

The fact that all the devices required for the setup of *Sing Pong, sssSnake, and Expressmas Tree* are hidden is one of the factors that create a hybrid connection between these games and theatrical performances. According to Laurel, audience members are usually not aware of the technical aspects of a theatrical performance and *"the action on the stage is all there is"* [Laurel, 1991].

Hiding the computer and the technical aspects involved in an interactive work extends the interface beyond the confines of the traditional desktop paradigm. It shifts the focus from the computer as a source of output to the user as a rich source of input. Traditionally, computer games are played on a screen, thus discouraging physical activity and obliging the users to enter the computer's virtual world [Pinhanez, 2002]. Moving away from the conventional screen may encourage whole-body movement and help users acquire a different interface, the real world.

6. Conclusions

6. Conclusions

This chapter provides a concise summary of the key achievements and findings in my research and suggests some directions for future work in the area.

As I have stressed in this thesis, speech recognition has dominated both research and deployment in the use of voice as input to interactive systems. However, there is increasing interest in the paralinguistic aspect, as evidenced by recent publications and projects described in this thesis.

In the current immature state of the field, I found it essential to review a wide range of aspects of voice including the physiological, the artistic-performative, the paralinguistic, the communicative, the cultural, and the technological. I believe that my review of the literature and of others' work across these various aspects will be of use to others. The table (Table 1 in the second chapter) that summarizes the paralinguistic information conveyed by various voice characteristics, for instance, may provide insights into the mappings that developers decide to establish between these characteristics and visual parameters or other interactive data. The section that outlines the classifications of vocalizations in the same chapter may help developers adopt a standard term to refer to this new research area and to define vocal input in relation to the linguistic/paralinguistic continuum.

At this stage it may be useful to restate my research questions and to comment on how far I have answered them.

What are the factors that may determine users' preferences and interaction patterns during non-speech voice control?

Creating my own projects allowed me to engage in an open-ended exploration leading to a number of original observations, many of which I had not anticipated. In summary, I found that paralinguistic vocal control of interactive media may be affected by a number of factors including gender, shyness,

tiredness, social and environmental context, cultural background, impairment, the type of vocalization that the developer expects or the user adopts as an input, and possibly the type of visuals that are used as an output. I hope that my observations will help future developers anticipate vocal preferences and patterns besides possible limitations in this new form of interaction.

I also found that new issues demand to be addressed in relation to the user's perception of causality when voice is used as input. I therefore hope that my observations will lead to studies of causal perception in voice-controlled applications. In the light of this, another fundamental area that would repay examination concerns the perceptual aspects that inform – or should inform– the mappings between the vocal and the visual that developers adopt. I found that the way users perceive the mappings is a further factor that may be of significance in determining their interaction patterns.

Another intriguing observation that calls for further investigation is the gestural accompaniments of non-speech vocalizations. The hands-in-pockets posture that many people adopt while whistling, the rhythmic gestures that many people perform while vocalizing or singing, and the backward tilt of the body that I observed when many participants approached the end of their vocalizations were unforeseen interaction patterns. My findings provide a starting point for further studies of this relationship between vocalization, posture and gestural movement.

What are the aspects of voice that have not yet been adequately explored and exploited in the field of interactive media? What benefits and what limitations are there in using paralinguistic rather than or along with linguistic input? And what is the future potential of paralinguistic voice input?

I embarked on this thesis arguing that there are aspects of voice which have not yet been tapped in the field of interactive media and which if appropriately exploited may enhance the user's experience of multimedia and offer certain benefits especially when used as a complementary input to speech recognition and other input mechanisms. I now have evidence to support this argument, but

also a much greater insight into the complexities and potential problems of paralinguistic vocal control.

One major benefit of the use of paralinguistic voice input is the possibility of maintaining a near-real-time and continuous response. However, one major drawback of this type of input is that it is tiring. This makes it crucial to further explore its use with other input modalities, especially speech, in order to create combinations that would be both compensatory and synergistic. I developed *Blowtter* in an attempt to investigate – in detail and on a practical level – the pros and cons of combining features of speech input with features of non-speech input. In speech-controlled applications, the technology is still limited, especially in recognizing continuous speech. In non-speech-controlled applications, the technology excels in responding to continuous sounds but users' ability to generate sound continuously is limited. I hope that the results of my evaluation of *Blowtter*, which I set out in the third chapter, may advance understanding of how non-speech input can complement attempts to use speech recognition, and contribute to the research being done on how voice can be used both on its own and in combination with other emerging input mechanisms to control interactive media. The combined use of the linguistic and the paralinguistic is in itself a huge topic which deserves substantial further work.

A further benefit of non-speech sounds is that they are not restricted to particular cultures. It is, however, possible that the ability to generate certain sounds may differ from one culture to another. The perception of voice-visual mappings may similarly be culturally dependent. Hence, although non-speech sounds are not linguistic, some of their aspects may nonetheless be language-specific or at least culture-specific. Certain sounds may have connotations inherent in a particular culture that impede users' perception of the correlation that the developer intends to conjure. It is therefore vital to augment research on the cultural factors that may affect the perception of voice-visual mappings.

Vocal input may prove a useful therapeutic intervention in asthma, voice disorders, social anxiety, shyness, and similar disorders. It may also be used as

a tool for people suffering from motor impairment. Disorders of this kind are frequently accompanied by speech impairment, making speech recognition an unrealistic option and hence suggesting a role for non-speech vocal control. The fact that any associated *cognitive* impairment may itself impede the perception of voice-visual or voice-physical mapping would of course be a significant challenge in this area.

Despite these limitations, developments such as those described in this thesis may contribute to the new generation of interactive applications that extend beyond the use of sound for sound effects, and beyond enabling users to interact with computers as mere users. They may contribute to a fundamental shift in multimedia by helping to emancipate it from the confines of the conventional desktop into the real environment. It may be a step forward in the development – already begun – of theatrical games that integrate virtual and real game play and blur the boundaries between virtual and real environments. This may help to reacquaint the user with the richest interface of all, the real world. One of my goals is to harness a range of emerging non-conventional input mechanisms going far beyond those described in this paper. These would exploit the potential of the body to recast itself as a rich source of input and – in synergy with technology – a still richer source of output to others.

References

References

Al Hashimi, S. (2005) Beyond Using Voice as Voice. *In Proceedings of the 16th International Conference for Advanced Studies in Systems Research, Informatics and Cybernetics*. Baden-Baden, Germany.

Allchin, D. (2002) 'Newton's Colors', http://www1.umn.edu/ships/updates/newton1.htm (January, 2004).

Allott, R. (1989*) The Motor Theory of Language Origin*. Lewes: Book Guild. (An adaptation is available on the Web at http://www.percepp.demon.co.uk/motor-i.htm).

American Public Works Association (2001) 'Diversity Resource Guide', http://www.apwa.net/Documents/MembersOnly/APWA_Diversity_Guide.pdf (March, 2007).

Aquinas, St. T. (1947) *The Summa Theologica of St. Thomas Aquinas*. trans. Fathers of the English Dominican Province. Benzinger Brothers, New York. (Available on the Web at http://www.newadvent.org/summa/).

Ardó, Z. (2001) 'Emotions, Taboos, and Profane', http://www.accurapid.com/Journal/16review.htm (August, 2006).

Arizona Health Sciences Library (2006) 'Health Problems of Musicians', http://www.ahsl.arizona.edu/about/ahslexhibits/musicianmedicalmaladies/instruments.cfm (February, 2006).

Armand, A. and Dauw, B. (2004) 'Commotion', http://www.francobelgedesign.com/ (August, 2005).

BALTIC Centre for Contemporary Art (2007) 'Marcus Coates', http://www.balticmill.com/whatsOn/future/ExhibitionDetail.php?exhibID=61 (January, 2007).

Ballora, M. (2006) 'Pitch Vs. Frequency', http://emusician.com/tutorials/emusic_pitch_vs_frequency/ (March, 2007).

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A. and Nöth, E. (2001) Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground. *In Proceedings of the European Conference on Speech Communication and Technology*. Aalborg, Denmark, Vol. 4, pp.2285-2288. (Available on the Web at http://www.ims.uni-stuttgart.de/~moebius/papers/prosmod_paper.pdf).

Batliner, A. and Nöth, E. (2003) Prosody and Automatic Speech Recognition - Why not yet a Success Story and where to go from here. *In Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*. Tokyo, pp. 357-364. (Available on the Web at http://www5.cs.fau.de/Forschung/Publikationen/2003/Batliner03-PAA.pdf?language=en).

Batliner, A. and Möbius, B. (2005) Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground? In William J. Barry and Wim A. van Dommelen (Eds.). *The Integration of Phonetic Knowledge in Speech Technology*. Springer, Dordrecht, pp.21-44. (Available on the Web at http://www.ims.uni stuttgart.de/~moebius/papers/batliner_moebius_phonknow2005.pdf).

Bauer, I. (2001) *Diaper Free: The Gentle Wisdom of Natural Infant Hygiene*. Natural Wisdom Press.

BBC News (2004) 'Shouting Men of Finland Perform Ice Breaker', http://news.bbc.co.uk/2/hi/europe/3454567.stm (August, 2005).

Beeman, W.O. (1998) 'The Mystery of Singing', http://www.brown.edu/Departments/Anthropology/publications/MysteryofSinging.htm (February, 2005).

Belin, P., Fectea, S. and Bédard, C. (2004) Thinking the Voice: Neural Correlates of Voice Perception. *TRENDS in Cognitive Sciences,* Vol. 8, pp.129-135. (Available on the Web at http://www.indiana.edu/~iung/fmriPapers/Belin(2004)_voice.pdf).

Belin, P., Fillion-Bilodeau, S. and Gosselin, F. (2005) 'The "Montreal affective voices": A validated set of non-linguistic emotional vocal expressions for research on auditory affective processing', http://vnl.psy.gla.ac.uk/articles/submitted_or_in_press/Montreal%20Affective%20Voices.pdf (December, 2006).

Berlin, B. (2005) Just another fish story? Size-symbolic properties of fish names. In *Animal Names*. A. Minelli, G. Ortalli, G. Singa (eds.). Venezia: Instituto Veneto di Scienze, Lettre ed Arti, pp. 9 - 21.

Bilmes, J., Li, X., Malkin, J., Kilanski, K., Wright, R., Kirchhoff, K., Subramanya, A., Harada, S., Landay, J.A., Dowden, P. and Chizeck, H. (2005) The Vocal Joystick Demo at UIST05: A Voice-Based Human-Computer Interface. *In Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*. Seattle.

Bohlen, M. and Rinker, JT. (2004) Unexpected, Unremarkable, and Ambivalent or How the Universal Whistling Machine Activates Language Remainders. *Computational Semiotics for Games and New Media, COSIGN2004*. University of Split, Croatia. (Available on the Web at http://www.realtechsupport.org/pdf/Cosign2004.pdf).

Borland, R., Findley J. and Jacobs, M. (2002) 'Milliefiore Effect', http://www.siggraph.org/artdesign/gallery/S02/onfloor/borland/1technicalstatement.html (July, 2004).

Bradbury-Carlin, R. (2007) *All my Shoes and Glasses*. Gothic Street Press, Massachusetts. (Available on the Web at http://www.allmyshoesandglasses.com/archives/ShoeAndGlasses.pdf).

Breazeal, C. and Aryananda, L. (2002) Recognition of Affective Communicative Intent in Robot-Directed Speech. *Autonomous Robots,* Vol. 12, pp.83-104.

Bregman, A. (1990) *Auditory Scene Analysis; The Perceptual Organisation of Sound*. MIT Press, Cambridge, MA.

Breidegard, B. and Balkenius, C. (2003) Speech development by imitation. In C.G. Prince, L. Berthouze, H. Kozima, D. Bullock, G. Stojanov and C. Balkenius (Eds.). *In Proceedings of the Second InternationalWorkshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems.* Lund University Cognitive Studies, pp. 57-64.

Carey, J. (1980) Paralanguage in Computer Mediated Communication. In N.K. Dondheimer (Ed.). *In Proceedings of the Eighteenth Annual Meeting of the Association for Computational Linguistics and Parasession on Topics in Interactive Discourse*. Philadelphia: University of Pennsylvania Press, pp. 67-69. (Available on the Web at http://www.cs.mu.oz.au/acl/P/P80/P80-1018.pdf).

Cavazza, M., Lugrin, J., Crooks, S., Nandi, A., Palmer, M. and Le Renard, M. (2005) Causality and Virtual Reality Art. *In Proceedings of the 5th Conference on Creativity &Amp; Cognition (C&C '05).* ACM Press, New York, pp.4-12.

Cheek, J.M. (1983) 'The Revised Cheek and Buss Shyness Scale', http://www.wellesley.edu/Psychology/Cheek/research.html#13item (December, 2006).

Clement, C.J., Den Os, E.A. and Koopmans-van Beinum, F.J. (1994) The Development of Vocalizations of Hearing Impaired Infants. *In Proceedings of the Institute of Phonetic Sciences Amsterdam*, Vol.18, pp.65-76.

Clement, C. J.,Koopmans-van Beinum, F. J. and Pols, L. C. W. (1996) Acoustical Characteristics of Sound Production of Deaf and Normally Hearing Infants. *In Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp.1549-1552. (Available on the Web at http://www.asel.udel.edu/icslp/cdrom/vol3/724/a724.pdf).

Collopy, F. (2000) Color, form, and motion: Dimensions of a musical art of light. *Leonardo*, Vol. 33, No. 5, pp. 355-360.

Communication, Cultural, and Media Studies (CCMS) (2003) 'NVC: Communication and Speech', http://www.cultsock.ndirect.co.uk/MUHome/cshtml/nvc/nvc6.html (January, 2005).

Corballis, M.C. (2003) From mouth to hand: gesture, speech and the evolution of right-handedness. *Behavioral and Brain Sciences,* Vol. 26, pp.198-208.

Correa, B. (2004) 'Blendie (2000), Kelly Dobson' http://www.mediamatic.net/popup.php?id=200.8288&label=FIG00 (February, 2005).

Cox, AL. and Walton, A. (2004) Evaluating the Viability of Speech Recognition for Mobile Text Entry. *In Proceedings of HCI 2004: Design for Life.*

Darwin, C. (1871) *The Descent of Man, and Selection in Relation to Sex* (2 vols). John Murray, London, UK.

Darwin, C. (1872) *The Expression of the Emotions in Man and Animals.* John Murray, London, UK [3rd edition edited by Paul Ekman. Oxford University Press, New York, 1998].

Darwin, C. and Ekman, P. (Ed.) (1998) *The Expression of the Emotions in Man and Animals*. Oxford University Press, Oxford, UK.

Davies, L. (2003) 'The Shy Child', http://www.kellybear.com/TeacherArticles/TeacherTip31.html (March, 2007).

de Götzen, A. (2004) The sounding gesture: an overview. *In Proceedings of the International Conference on Digital Audio Effects (DAFx).* Naples, Italy. (Available on the Web at http://dafx04.na.infn.it/WebProc/Proc/P_005.pdf).

Department of Linguistics and Modern English Language (LAMEL) at Lancaster University (2002) 'The EAGLES Spoken Language Working Group Work, Package 4', http://bowland-files.lancs.ac.uk/eagles/glossar1.htm (January, 2006).

Dissanayake, E. (2005) A review of The Singing Neanderthals: The Origins of Music, Language, Mind and Body by Steven Mithen. *Evolutionary Psychology*, Vol. 3, pp.375-380. (Available on the Web at http://human-nature.com/ep/reviews/ep03375380.html).

Dobson, K. (2003) 'Blendie 2000', http://web.media.mit.edu/%7Emonster/blendie/ (February, 2004).

Dobson, K. (2004) 'Screambody', http://web.media.mit.edu/~monster/screambody/ (February, 2004).

Dobson, K, Whitman, B. and Ellis, D. P. (2005) Learning auditory models of machine voices. *In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Dozza, M., Chiari, L. and Horak, F.B. (2005) Audio-Biofeedback Improves Balance in Patients With Bilateral Vestibular Loss. *Archives of Physical Medicine Rehabilitation*, Vol. 86, No.7, pp.1401-1403.

Dunbar, R.I.M. (1993) Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, Vol.16, No.4, pp.681-735.

Dyer, F. L. (1929) 'Edison, His Life and Inventions', http://www.worldwideschool.org/library/books/hst/biography/Edison/chap10.html (February, 2005).

Elliott, J. (2005) 'How Singing Unlocks the Brain', BBC News, http://news.bbc.co.uk/1/hi/health/4448634.stm (November, 2005).

Elmorex Ltd (2000) 'Rring!: Generation of Ringing Tones', http://playsingmusic.com/products/rring_white_paper.pdf (August, 2005).

Encyclopedia of Nursing and Allied Health (2007) 'Voice Disorders', http://health.enotes.com/nursing-encyclopedia/voice-disorders (March, 2007).

Farlex, Inc. (2004) 'Theatre In The Round', http://encyclopedia.thefreedictionary.com/Theatre%20in%20the%20round (August, 2004).

Fearn, R.A. (2001) *Music and Pitch Perception of Cochlear Implant Recipients*. Ph.D. Thesis, University of New South Wales. (Available on the Web at http://www.phys.unsw.edu.au/jw/reprints/FearnThesis.pdf).

Fernald, A. (1992) 'Mothers' melodies teach babies lyrics of language', http://news-service.stanford.edu/pr/92/920212Arc2430.html.

Finch, P. (2003) 'Sound Poetry', http://www.57productions.com/article_reader.php?id=11 (January, 2006).

Ford Motor Company (2006) 'Fiesta Connexxion', http://www.fordgame.be/ (October, 2006).

Fox, K. (2003) 'SIRC Guide to Flirting; What Social Science can Tell you about Flirting and how to do it', Social Issues Research Centre, http://www.sirc.org/publik/flirt.pdf (April, 2005).

Friend, M. and Bryant, J. B. (2000) A Developmental Lexical Bias in the Interpretation of Discrepant Messages. *Merrill-Palmer Quarterly*, Vol. 46, pp.140-167. (Available on the Web at http://www.findarticles.com/p/articles/mi_qa3749/is_200004/ai_n8899221).

Fruhlinger, J. (2004) 'Universal Whistling Machine whistles', http://www.engadget.com/entry/1234000730022889/ (December, 2004).

Gaver, W. (1989a) The Sonic Finder: An Interface that Uses Auditory Icons. *Human-Computer Interaction*, Vol. 4, pp. 67-94.

Gaver, W. (1989b) Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, Vol. 2, pp.167-177.

Gaye, L. and Holmquist, L.E. (2004) In Duet with Everyday Urban Settings: a User Study of Sonic City. *In Proceedings of the New Interfaces for Musical Expression*.

Givens, D.B. (1998a) 'Tone of Voice', Center for Nonverbal Studies, http://members.aol.com/nonverbal2/tone.htm (August, 2004).

Givens, D.B. (1998b) 'Human Brain', http://members.aol.com/nonverbal3/human.htm (January, 2005).

Goberis, D. and Loraine, S. (2006) 'A Child with a Cochlear Implant is Joining my Classroom – What in the World do I do?', http://www.medel.com/Shared/pdf/en/HEARSAY_vol2.pdf. (January, 2007).

Grauwinkel, K. and Fagel, S. (2006) Crossmodal Integration and McGurk-Effect in Synthetic Audiovisual Speech. *In Proceedings of the International Conference on Speech and Computer*. St. Petersburg.

Gray, H., Warwick, R. and Williams, PL. (Eds.) (1973) *Gray's Anatomy*. 35th ed. Longman, London, UK.

H2 Training and Consultancy (2006) 'How to Make a Good Impression on the Telephone', http://www.h2training.com/telephone_tips.pdf (March, 2007).

Hale, M. and Kissock, M. (2002) 'Introduction to Linguistic Science', http://modlang-hale.concordia.ca/IntroBook.pdf (March, 2007).

Hall, D. C. (2002) 'Interactive Sagittal Section', http://www.chass.utoronto.ca/~danhall/phonetics/sammy.html (March, 2007).

Hall, E. (1966) *The Hidden Dimension*. Doubleday & Company, New York.

Hämäläinen, P., Maki-Patola, T., Pulkki, V. and Airas, M. (2004) Musical Computer Games Played by Singing. *In Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04)*, Naples, Italy, pp. 367-371.

Hämäläinen, P., Ilmonen, T., Höysniemi, J., Lindholm, M. and Nykänen, A. (2005) 'Martial arts in artificial reality'. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM Press, New York, NY, pp.781-790.

Harada, S., Landay, J., Malkin, J., Li, X. and Bilmes, J. (2006) The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques. *In Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility.* Portland, Oregon.

Hartman, M. J. (2004) 'People With and Without Disabilities: Interacting & Communicating', http://eeo.gsfc.nasa.gov/disability/publications.html (March, 2004).

Hartmann, W.M. (1998) *Signals, sound, and sensation*. Springer_Verlag, New York.

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

Herr (2002) 'Tip of the Tongue Experience', http://www.everything2.com/index.pl?node_id=1078328 (April, 2005).

Hirst, D.J. and Di Cristo, A. (Eds.) (1998) A survey of intonation systems. In Hirst and DI. Cristo (Eds.). *Intonation Systems: a Survey of Twenty Languages*. Cambridge: Cambridge University Press, pp.1-44.

Ho Ching-Hsiang (2001) *Speaker Modelling for Voice Conversion*. Ph.D. thesis, Department of Electronic and Computer Engineering, Brunel University.

Hooshmand, H. (2000) 'Anatomy Sketches', Neurological Associates Pain Management Center, http://www.rsdrx.com/Anatomy%20Sketches.htm (March, 2007).

Huerta, J., Lubensky, D., Nahamoo, D., Pieraccini, R., Raman, T.V. and Wiecha, C. (2004) 'Reusable Dialog Components; Mainstreaming speech-enabled web applications', http://www-128.ibm.com/developerworks/web/library/wa-dialogcomp/ (February, 2005).

Huron, D., Kinney, D. and Precoda, K. (2000) 'Relation of Pitch Height to Perception of Dominance/Submissiveness in Musical Passages', http://csml.som.ohio- state.edu/Music829D/smile.html (January, 2005).

Iga, S. and Higuchi, F. (2002) Kirifuki: Inhaling and Exhaling Interaction for Entertainment Systems. *Transactions of the Virtual Reality Society of Japan TVRSJ*, Vol. 7, No.4, pp.445-452.

Igarashi, T. and Hughes, F. (2001) Voice as Sound: Using Non-verbal Voice Input for Interactive Control. *In Proceedings of the 14th Annual Symposium on User Interface Software and Technology*, ACM UIST'01, Orlando, Florida, pp.155-156.

International Kabaddi Federation (2006) 'Origin, History and Development of Kabaddi', http://www.kabaddiikf.com/history.htm (December, 2006).

Izdebski, K., Shipp, T. and Dedo, H.H (1978) Voice onset and offset RTs in spastic dysphonia. *The Journal of the Acoustical Society of America*, Vol. 64, No. S1, pp. S51-S52 .

Jäncke, L., Wüstenberg, T., Scheich, H. and Heinze, H. (2002) Phonetic Perception and the Temporal Cortex. *NeuroImage*, Vol.15, pp.733–746.

Jindrak, K.F. (1986) *Sing, Clean Your Brain and Stay Sound and Sane: Postulate on Mechanical Effect of Vocalization on the Brain*. Forest Hills Station, New York: Karel F. Jindrak and Heda Jindrak.

Johnstone, I. T. (1996) Emotional Speech Elicited Using Computer games. *In Proceedings of the 4th International Conference on Spoken Language Processing*, Vol. 3, pp.1985-1988. (Available on the Web at http://www.unige.ch/fapse/emotion/publications/pdf/icslp96.pdf).

Johnstone, T., van Reekum, C. M., Hird, K., Kirsner, K. and Scherer, K. R. (2005) Affective Speech Elicited with a Computer Game. *Emotion*, Vol. 5, pp.513-518. (Available on the Web at http://brainimaging.waisman.wisc.edu/~tjohnstone/2005_12_Johnstone_Emotion .pdf).

Kappas, A. (2003) 'Emotions in the Voice', http://www.iu-bremen.de/hss/akappas/31033/index.shtml (February, 2004).

Karpf, A. (2006) *The Human Voice*. Bloomsbury, London, Great Britain.

Kataria, M. (2002) *Laugh for no Reason*. Madhuri International, Mumbai, India.

Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., Itou, K., Ito, A., Yamamoto, M., Yamada, A., Utsuro, T. and Shikano, K. (2000) Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition. *In Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Vol. 4, pp. 476-479.

Nelson, D.G. K., Hirsh-Pasek, K., Jusczyk, P.W. and Cassidy, K.W. (1989) How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, Vol.16, pp. 55-68.

Kirchhoff, K. and Schimmel, S. (2005) Statistical properties of infant-directed vs. adult-directed speech: insights from speech recognition. *Journal of the Acoustical Society of America,* Vol.117, No.4, pp. 2224-2237.

Kirschner, R., Morawe, V., and Reiff, T. (2001) 'His Master's Voice', http://www.fursr.com/details.php?id=3&pid=3 (October, 2005).

Kirschner, R., Morawe, V., and Reiff, T. (2003) 'SoundSlam', http://www.fursr.com/ (October, 2005).

Knuuti, S. (2003) The Serious Business of Shouting. *Finnish Music Quarterly 4*. Finnish Music Information Centre. (Available on the Web at http://www.fimic.fi/fimic/fimic.nsf/mainframe?readform&E348BCE6BE7F4F21C2256F51002B2986).

Kob, M. (2002) *Physical modeling of the singing voice*. Ph.D. thesis, Aachen University (RWTH). (Available on the Web at http://sylvester.bth.rwth-aachen.de/dissertationen/2002/132/02_132.pdf).

Lane, H., and Tranel, B. (1971) The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, Vol.14, pp.677-709.

Lattner, S., Meyer, M. and Friederici, A. D. (2005) Voice Perception: Sex, Pitch, and the Right Hemisphere. *Human Brain Mapping*, Vol. 24, pp.11-20.

Laukka, P. (2004) Vocal Expression of Emotion: Discrete-Emotions and Dimensional Accounts (Doctoral dissertation, Uppsala University, 2004). *In Comprehensive summaries of Uppsala dissertations from the faculty of social sciences*. Uppsala, Sweden: Acta Universitatis Upsaliensis, Vol. 141, pp. 1-80. (Available on the Web at http://publications.uu.se/theses/abstract.xsql?dbid=4666).

Laurel, B. (1990) *The Art of Human-Computer Interface Design*. Addison-Wesley, Reading, Massachusetts.

Laurel, B.(1991) *Computers as Theatre*. Addison-Wesley, Reading, Massachusetts.

Levin, G. (1999) 'Interface Metaphors and Signal Representation for Audiovisual Performance Systems', http://acg.media.mit.edu/people/golan/thesis/proposal/ (October, 2005).

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

Levin, G. and Lieberman, Z. (2004) In-Situ Speech Visualization in Real-Time Interactive Installation and Performance. *In Proceedings of The 3rd International Symposium on Non-Photorealistic Animation and Rendering*, Annecy, France.

Levin, G. (2005) 'Synopses of Major Projects', http://www.flong.com/resume/projects.html (January, 2006).

Levin, T.C. and Edgerton, M.E. (1999) 'The Throat Singers of Tuva', http://www.sciam.com/print_version.cfm?articleID=00080AA2-BA32-1C73-9B81809EC588EF21 (January, 2006).

Lewis, J. (2002) *Forest Hunter-Gatherers and their World: A Study of the Mbendjele Yaka Pygmies of Congo-Brazzaville and their Secular and Religious Activities and Representations*. Ph.D. Thesis, London School of Economics and Political Science.

Liljedahl, M., Lindberg, S. and Berg, J. (2005) Digiwall – an Interactive Climbing Wall. *In Proceedings of ACE2005*. Valencia, Spain, pp. 15-17.

Limber, J. (1982) What can chimps tell us about the origins of language. In S. Kuczaj (Ed.). *Language Development*. Hillsdale, NJ: L. E. Erlbaum,Vol. 2, pp. 429-446. (Available on the Web at http://pubpages.unh.edu/~jel/JLimber/What_can_chimps.pdf).

Lipscomb, S.D. and Kim, E.M. (2004) Perceived match between visual parameters and auditory correlates: An experimental multimedia investigation. In S.D. Lipscomb, R. Ashley, R.O. Gjerdingen and P. Webster (Eds.). *In Proceedings of the 8th International Conference on Music Perception & Cognition*. Sydney, Australia: Causal Productions, pp.72-75.

Loewy, J. (2004) 'Integrating Music, Language and the Voice in Music Therapy', http://www.voices.no/mainissues/mi40004000140.html (December, 2005).

Lowrey, T.M. and Shrum, L. J. (2007) Phonetic Symbolism and Brand Name Preference. *Journal of Consumer Research.* (Available on the Web at http://faculty.business.utsa.edu/ljshrum/JCR2006.Final.doc).

Ma, J. (2001) 'Sound and Hearing – Formative Evaluation; Sound Spectrogram', http://www.exploratorium.edu/partner/pdf/soundSpect_rp_06.pdf (April, 2007).

Madan, A. (undated) 'Jerk-O-Meter: Speech-Feature Analysis Provides Feedback on Your Phone Interactions', http://www.media.mit.edu/press/jerk-o-meter/ (August, 2005).

Madan, A., Caneel, R. and Pentland, A. (2005) Voices of Attraction. *In Proceedings of Augmented Cognition, HCI 2005*, Las Vegas.

Malkin, J. and Brandi, H. (2007) 'A Simulated Robotic Arm', http://ssli.ee.washington.edu/vj/video_demos.htm (May, 2007).

Matsumura, E. (2005) 'Blowing "Windows"', http://www.interaction.rca.ac.uk/alumni/04-06/Eriko/html/01_puppet.htm (August, 2005).

McCaffery, S. (1978) 'Sound Poetry: A Survey', http://www.ubu.com/papers/mccaffery.html (January, 2006).

Mcginty, S. (2005) 'Can't hear you, dear ... blame my brain', http://news.scotsman.com/scitech.cfm?id=1736172005 (June, 2006).

McGurk, H. and MacDonald, J. (1976) Hearing lips and seeing voices. *Nature*, Vol. 264, pp.746–748.

Meade, L. (2002) 'Nonverbal Communication', http://lynn_meade.tripod.com/id56.htm (January, 2005).

Mehrabian, A. (1981) *Silent messages: Implicit communication of emotions and attitudes*. Wadsworth, Belmont, CA.

Michotte, A. (1963) *The perception of causality*. Basic Books, New York.

Milestone International (2006) 'The Pretender Voice Changer', http://www.milestonesafety.com/pretender-voice-changer.html (April, 2007).

Mithen, S. (2005*) The Singing Neanderthals: The Origins of Music, Language, Mind and Body*. Weidenfeld & Nicolson, London, UK.

Montola, M.and Waern, A. (2006) Participant Roles in Socially Expanded Games. *In Proceedings of the Third International Workshop on Pervasive Gaming Applications, Pervasive Conference*. Dublin, Ireland.

Moritz, W. (1997) 'The Dream of Color Music, And Machines That Made it Possible', http://www.awn.com/mag/issue2.1/articles/moritz2.1.html (January, 2004).

National Center for Voice and Speech (2005) 'A Few Acoustics and Physics Basics', http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/refresh.html (March, 2005).

Ng, K., Popat, S., Stefani, E., Ong, B. and Cooper, D. (2000) Music via Motion: Interactions between Choreography and Music. *In Proceedings for the Inter-Society for the Electronic Arts (ISEA) Symposium*, France. (Available on the Web at http://www.isea2000.com/actes_doc/48_kia_ng.rtf).

Ng, K. (2004) 'Music via Motion', http://www.kcng.org/mvm/ (October, 2005).

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

Norman, D. (1999) *The Invisible Computer: Why good products can fail, the personal computer is so complex and information appliances are the solution*. MIT Press, Cambridge.

Nowak, R. (2003) 'New software allows you to log on by laughing', http://www.newscientist.com/article.ns?id=dn3921 (December, 2006).

Nwokah, E. E., Hsu, H., Davies, P. and Fogel, A. (1999) The Integration of Laughter and Speech in Vocal Communication: A Dynamic Systems Perspective. *Journal of Speech, Lang & Hearing Res*, Vol. 42, pp.880-894.

Okimoto, S. (2005) 'Innovative Gameplay in Nintendo DS Launch Titles', http://www.experimental-gameplay.org/2005/nintendo_ds.ppt#256,1,Innovative Gameplay in Nintendo DS Launch Titles (September, 2006).

Ostendorf, M., Shafran, I. and Bates, R. (2003) Prosody Models for Conversational Speech Recognition. *In Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pp. 147-154.

Öster, A-M. (1996) Clinical applications of computer-based speech training for children with hearing impairment. *In Proceedings of the International Conference on Spoken Language Processing (ICSLP).* Philadelphia, USA. pp.157-160.

O'Sullivan, D. and Igoe, T. (2004) *Physical Computing: Sensing and Controlling the Physical World with Computers*. Thomson Course Technology PTR, Boston.

Patten, J., Recht, B. and Ishii, H. (2002) Audiopad: A Tag-based Interface for Musical Performance. *In Proceedings of the New Interfaces for Musical Expression (NIME '02).*

Picard, R.W (2000) Toward Computers That Recognize and Respond to User Emotion. *IBM Systems Journal*, Vol. 39, Nos. 3 and 4, pp.705-719.

Pinhanez, C. (2002) Creating ubiquitous interactive games using everywhere displays projectors. *In Proceedings of the International Workshop on Entertainment Computing*.

Provine, R.R. (1996) Laughter. *American Scientist,* Vol. 84, No.1 (Jan-Feb, 1996): pp.38-47. (Available on the Web at http://www.americanscientist.org/template/AssetDetail/assetid/24591?fulltext=true&print=yes).

Quast, H. (2002) Automatic Recognition of Non-verbal Speech; An Approach to Model the Perception of Para- and Extralinguistic Vocal Communication with Neural Networks. *Machine Perception Lab Tech Reports.*

Quast, H. (2003) 'Robust Machine Perception of Nonverbal Speech', http://ergo.ucsd.edu/~holcus/Speech.html (January, 2005).

Rinman, M. Friberg, A., Bendiksen, B., Kjellmo, I., Cirotteau, D., McCarthy, H., Mazzarino, B. and Dahl, S. (2003) Ghost in the Cave: An Interactive Collaborative Game Using Non-Verbal Communication. In R. Bresin (Ed.). *In Proceedings of SMAC03*, Vol. 2, pp. 561- 563.

Rizzolatti G. and Arbib, M.A. (1998) Language within our grasp. *Trends in Neuroscience*, Vol. 21, pp.188–194.

Roads, C. (1996) *The Computer Music Tutorial*. MIT Press, Cambridge, MA.

Roosegaarde, D. (2005) '4D- Pixel', http://www.daanroosegaarde.nl/ (October, 2005).

Ruch, W. and Ekman, P. (2001) The Expressive Pattern of Laughter In A.W. Kaszniak (Eds.) Emotion Qualia, and Consciousness. *Word Scientific Publisher*. Tokyo, pp. 426-443.

Saffran, J. R. and Griepentrog, G. J. (2001) Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology,* Vol. 37, pp.74–85.

Scherer, K. R. (1986) Vocal Affect Expression: A review and a Model for Future Research. *Psychological Bulletin*, Vol. 99, pp.143-165.

Scherer, K. R. (1995) Expression of emotion in voice and music. *Journal of Voice*, Vol. 9, No.3, pp.235-248. (Available on the Web at http://affectco.unige.ch/system/files/1995_Scherer_JVoice.pdf).

Scherer, K. R. (1996) Adding the Affective Dimension: A New Look in Speech Analysis and Synthesis. *In Proceedings of the International Conference on Spoken Language Processing (ICSLP).* Philadelphia.

Scherer, R. C. (2000) 'A Basic Overview of Voice Production', http://www.voicefoundation.org/VFScherervoiceprod.html (March, 2004).

Schmitt, A. (2003) 'asFFT Xtra', http://www.as-ci.net/asFFTXtra (May, 2004).

Schötz, S. (2002) *Linguistic & Paralinguistic Phonetic Variation in Speaker Recognition & Text-to-Speech Synthesis*. Term paper for course in Speech Technology, GSLT. (Available on the Web at http://www.speech.kth.se/~rolf/gslt_papers/SusanneSchotz.pdf).

Sgouros, N. M. (2000) Detection, Analysis and Rendering of Audience Reactions in Distributed Multimedia Performances. *In Proceedings of the eighth ACM international conference on Multimedia*, ACM, Los Angeles, pp.195-200.

Sheridan, J., Bayliss, A. and Bryan-Kinns, N. (2006). iPoi: acceleration as a medium for digital live art. Demonstration. *In Proceedings of The 8th International Conference on Ubiquitous Computing.* Los Angeles, USA.

Skille, O. (undated) 'Mechanical Cleaning of Brain Cells and Muscle Cells by Sound Vibration', http://members.tripod.com/~quadrillo/VAT/e_jindrak-2.html (March, 2005).

Slaney, M. and McRoberts, G. (1998) Baby Ears: A Recognition System for Affective Vocalization. *In Proceedings of ICASSP 1998.*

Smith, G. (1997) 'German Pronunciation', http://www.wm.edu/modlang/gasmit/pronunciation/vowels.html (March, 2007).

Snibbe, S. (2005) 'Blow up',
http://www.snibbe.com/scott/breath/blowup/index.html (October, 2005).

Sokhi, D.S., Hunter, M.D., Wilkinson, I.D. and Woodruff, P.W.R.(2005) Male and Female Voices Activate Distinct Regions in the Male Brain. *NeuroImage*, Vol. 27, No. 3, pp. 572-578.

Someth;ng (2005) 'Headspin',
http://www.somethingonline.org/index.php?main=in_progress&sub=headspin (April, 2005).

Sony Computer Entertainment Europe (2003) 'What is EyeToy? '
http://www.eyetoy.com/english/index.html (September, 2004).

Sørensen, K. and Petersen, K.Y. (2000) 'Smiles in Motion',
http://www.boxiganga.dk/english/enindex.html (February, 2006).

Sparacino, F., Wren, C., Davenport, G. and Pentland, A. (1999) Augmented performance in dance and theatre. *In Proceedings of the International Conference on Dance and Technology (IDAT99).* Arizona State University, Tempe, AZ, pp.25-28.

Sperberg-McQueen, C.M. and Burnard, L. (Eds.) (2005) 'Transcriptions of Speech: Text Encoding Initiative', http://www.tei-c.org/P5/Guidelines/TS.html (January, 2006).

Sporka, A.J., Kurniawan, S.H. and Slavik, P. (2005) Acoustic Control of Mouse Pointer. *Universal Access in Information Society, a Springer-Verlag journal*.

Srinivasan, R. V. and Lim, J. (2000) The Impacts of Involuntary Cues on Media Effects. *In Proceedings of the 33rd Hawaii International Conference on Systems Sciences, Maui (HICSS 2000).* (Available on the Web at http://csdl.computer.org/comp/proceedings/hicss/2000/0493/01/04931021.pdf).

Standley, J.M. (2000) The effect of contingent music to increase non-nutritive sucking of premature infants. *Pediatric Nursing*, Vol. 26, No.5, pp.493-495, 498-499.

Stoltzman, W. (2005) 'How Convincing Am I?',
http://groupmedia.media.mit.edu/elev.php (August, 2006).

Strubbe, B. (2003) 'Getting Serious about Laughter',
http://www.worldandi.com/newhome/public/2003/march/nspub.asp (August, 2005).

Technovelgy.com (2004) 'Universal Whistling Machine - The Future of Non-Verbal Communications', http://www.technovelgy.com/ct/Science-Fiction-News.asp?NewsNum=284 (December, 2004).

Thompson, W.F., Schellenberg, E.G. and Husain, G. (2004) Decoding speech prosody: Do music lessons help?. *Emotion*, Vol. 4, pp. 46-64. (Available on the Web at http://www.erin.utoronto.ca/~w3psygs/Thompson.pdf).

Tidwell, C. H. (2003) 'Non-Verbal Communication Modes', http://www.andrews.edu/~tidwell/lead689/NonVerbal.html (January, 2005).

Tourette Syndrome Association (2005) 'Tourette Syndrome Fact Sheet', http://www.ninds.nih.gov/disorders/tourette/detail_tourette.htm#What%20is%20Tourette%20syndrome? (April, 2005).

Trehub, S. E. and Nakata, T. (2002) Emotion and music in infancy. *Musicae Scientiae*, pp.37-61.

Unwin, M., Kenny, D.T. and Davis, P. J. (2002) The Effect of Singing on Mood. *Psychology of Music*, Vol. 30, No.2, pp.175-185.

Useher, R. (2004) 'A Whistle a Day Keeps Globalization Away', http://www.time.com/time/europe/magazine/article/0,13005,901040726-664985,00.html (December, 2004).

Van Leeuwen, T. (1999) *Speech, Music, Sound.* Macmillan Press LTD, London.

Von Kriegstein, K., Kleinschmidt, A. and Giraud, AL. (2006) Voice Recognition and Cross-Modal Responses to Familiar Speakers' Voices in Prosopagnosia. *Cerebal Cortex 2006*, Vol.16, pp.1314-1322. (Available on the Web at http://cercor.oxfordjournals.org/cgi/reprint/bhj073v1).

Wakefield, J. (2004) 'Body Movement to Create Music', http://news.bbc.co.uk/2/hi/technology/3873481.stm (October, 2005).

Walker, C.F. (2006) 'The Computer's Role in Speech Therapy', http://www.videovoice.com/vv_crole.htm (February, 2006).

Wendt, L. (2000) 'Narrative as Genealogy: Sound Sense in an Era of Hypertext', http://cotati.sjsu.edu/spoetry/nghome.html#sp (January, 2006).

Wharton,T. (2003) Interjections, Language and the 'Showing-Saying' Continuum. *Pragmatics and Cognition,* Vol. 11, No.1, pp.39-91.

Wiberg, M. (2006) Graceful Interaction in Intelligent Environments. *In Proceedings of the International Symposium on Intelligent Environments,* Cambridge.

Xiaoqing, Y. (2003) 'Nonverbal Sound Patterns, Paralanguage', http://www.cs.uta.fi/~grse/ACAI_2003/NonVerbalCommunication/SoundPatterns.ppt (January 2005).

Yacoub, S., Simske, S., Lin, X. and Burns, J. (2003) 'Recognition of Emotions in Interactive Voice Response Systems', http://www.hpl.hp.com/techreports/2003/HPL-2003-136.pdf (March, 2004).

Zivanovic, A. (2000) http://www.senster.com/ihnatowicz/ (August, 2005).

Practical Projects

Practical projects submitted to Middlesex University in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Practical Projects

The enclosed DVD contains video documentations of *sssSnake*, *Blowtter*, and *Expressmas Tree*, which are submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy. Gordon Davies contributed to the development of the source code.

Appendices

Appendix A: The Revised Cheek and Buss Shyness Scale

How Shy Are You?

Read each item carefully, and decide to what extent it is characteristic of your feelings and behavior. Answer each question by choosing a number from the scale below.

1 = Very uncharacteristic or untrue, strongly disagree
2 = Uncharacteristic
3 = Neutral
4 = Characteristic
5 = Very characteristic or true, strongly agree

Questions

____ 1. I feel tense when I'm with people I don't know well.

____ 2. I am socially somewhat awkward.

____ 3. I find it difficult to ask other people for information.

____ 4. I am often uncomfortable at parties and other social functions.

____ 5. When in a group of people, I have trouble thinking of the right things to say.

____ 6. It takes me long to overcome my shyness in new situations.

____ 7. It is hard for me to act natural when I am meeting new people.

____ 8. I feel nervous when speaking to someone in authority.

____ 9. I have doubts about my social competence.

____ 10. I have trouble looking someone right in the eye.

____ 11. I feel inhibited in social situations.

____ 12. I find it hard to talk to strangers.

____ 13. I am more shy with members of the opposite sex.

Now, add up your score. If you scored over 49, you're probably very shy. If your score is between 34 and 49, you're somewhat shy. If you scored below 34, you're probably not a particularly shy person, although you may feel shy in one or two situations. Most shy people score over 39 and a few reach the possible high score of 65.

Appendix B: Previous Practical Projects

The enclosed DVD contains video documentations of *SpitSplat* and *Sing Pong*. I developed *SpitSplat* in collaboration with Gordon Davies, Sumetha Nagalingam, and Athanasios Anthopoulos, and I developed *Sing Pong* in collaboration with Gordon Davies and Sumetha Nagalingam.

Appendix C: List of Relevant Publications by the Author

Parts of this thesis have been published in the following conference proceedings and journals:

Chapter 3:

Al Hashimi, S. (2005) Beyond Using Voice as Voice. *In Proceedings of the 16th International Conference for Advanced Studies in Systems Research, Informatics and Cybernetics*. Baden-Baden, Germany.

Al Hashimi, S. (2006) Blowtter: a Voice-Controlled Plotter. *In Proceedings of the 20th BCS HCI Group Conference in Cooperation with ACM*. Vol 2. London, UK.

Al Hashimi, S., Davies, G. (2006) Vocal Telekinesis; Physical Control of Inanimate Objects with Minimal Paralinguistic Voice Input. *In Proceedings of ACM Multimedia 2006*. Santa Barbara, California, USA.

Al Hashimi, S. (2006) Users as Performers in Vocal Interactive Media: The Role of Expressive Voice-Visualizations in Transforming Users into Performers. *International Journal of Performance Arts and Digital Media*. 2.3, London, UK.

Chapter 4 (parts of it)**:**

Al Hashimi, S. (2007) Preferences and Patterns of Paralinguistic Voice Input to Interactive Media. *In Proceedings of HCI International 2007, 12th international conference on Human-Computer Interaction*. Beijing, China.

Chapter 5 (Section 5.4)**:**

Al Hashimi, S. (2007) The Role Of Paralinguistic Voice Input of Interactive Media in Augmenting Awareness of Voice Characteristics in the Hearing-Impaired. *Extended Abstracts In Proceedings of the twenty-fifth annual CHI conference (CHI 2007)*, California, San Jose, April 2007.

Chapter 5 (Section 5.6)**:**

Al Hashimi, S. (2006) Users as Performers in Vocal Interactive Media: The Role of Expressive Voice-Visualizations in Transforming Users into Performers. *International Journal of Performance Arts and Digital Media*. 2.3, London, UK.

Appendix D: Samples of Relevant Publications by the Author

The enclosed CD contains some samples of my relevant publications.