Preferences and Patterns of Paralinguistic Voice Input to Interactive Media

Sama'a Al Hashimi

Lansdown Centre for Electronic Arts
Middlesex University
Hertfordshire, England
Samaa.alhashimi@gmail.com

Abstract. This paper investigates the factors that affect users' preferences of non-speech sound input and determine their vocal and behavioral interaction patterns with a non-speech voice-controlled system. It throws 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, it presents and employs a voice-controlled Christmas tree in a preliminary 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.

Keywords: Paralanguage, vocal control, preferences, voice-physical.

1 Introduction

As no other studies appear to exist in the paralinguistic vocal control area addressed by this research, the paper comprises a number of preliminary experiments that explore the preferences and patterns of interaction with non-speech voice-controlled media. In the first section, it presents a general overview of the voice-controlled project that was employed for the experiments. In the second section it discusses the experimental designs, procedures, and results. In the third section it presents the findings and their implications in an attempt to lay the ground for future research on this topic. The eventual aim is for these findings to be used in order to aid the developers of non-speech controlled systems in their input selection process, and in anticipating or avoiding vocal input deviations that may either be considered undesirably awkward or serendipitously "graceful" [6]. In the last section, it discusses the conclusions and suggests directions for future research.

The project that propelled this investigation is *sssSnake*; a two-player voice-physical version of the classic 'Snake'. It consists of a table on top of which a virtual snake is projected and a real coin is placed [1]. The installation consists of four microphones, one on each side of the table. One player utters 'sss' to move the snake

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 3–12, 2007.

and chase the coin. The other player utters 'ahhh' to move the coin away from the snake. The coin moves away from the microphone if an 'ahhh' is detected and the snake moves towards the microphone if an 'ssss' is detected. Thus players run round the table to play the game.

This paper refers to applications that involve vocal input and visual output as voice-visual applications. It refers to systems, such as *sssSnake*, that involve a vocal input and a physical output as voice-physical applications. It uses the term vocal paralanguage to refer to a non-verbal form of communication or expression that does not involve words, but may accompany them. This includes voice characteristics (frequency, volume, duration, etc.), emotive vocalizations (laughing, crying, screaming), vocal segregates (ahh, mmm, and other hesitation phenomena), and interjections (oh, wow, yoo). The paper presents projects in which paralinguistic voice is used to physically control inanimate objects in the real world in what it calls *Vocal Telekinesis* [1]. This technique may be used for therapeutic purposes by asthmatic and vocally-disabled users, as a training tool by vocalists and singers, as an aid for motor-impaired users, or to help shy people overcome their shyness.

While user-testing *sssSnake*, shy players 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 inquiry about the factors that influence users' preferences and patterns of interaction with a non-speech voice-controlled system, and that developers should, therefore, 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, cultural background, social context, and physiological limitations. There are other aspects to bear in mind. The author of this paper, for instance, prefers uttering 'mmm' while testing her projects because she noticed that 'mmm' is less tiring to generate for a prolonged period than a whistle. This seems to correspond with the following finding by Adam Sporka and Sri Kurniawan during a user study of their Whistling User Interface [5];

"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." [5]

The next section presents the voice-controlled Christmas tree that was employed in investigating and hopefully propelling a wave of inquiry into the factors that determine these preferences and interaction patterns. The installation was initially undertaken as an artistic creative project but is expected to be of interest to the human-computer interaction community.

2 Expressmas Tree

2.1 The Concept

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 (Fig. 1 shows an example). Longer vocalizations switch more bulbs on, thus allowing for new forms of expression resulting in vocal decoration of a Christmas tree.

Expressmas Tree employs 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.

Fig. 1. A participant uttering 'aah' to control *Expressmas Tree*

2.2 The Implementation

The main hardware components included 52 MES light bulbs (12 volts, 150 milliamps), 5 microcontrollers (Basic Stamp 2), 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 Christmas tree.

The application was programmed in Pbasic and Macromedia Director/Lingo. Two Xtras (external software modules) for Macromedia Director were used: asFFT and Serial Xtra. asFFT [4], which employs the Fast Fourier Transform (FFT) algorithm, was used to analyze vocal input signals. On the other hand, the Serial Xtra is used for serial communication between Macromedia Director and the microcontrollers.

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.

3 Experiments and Results

3.1 First Experimental Design and Setting

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

Experimental Procedures. Four female students and seven male students volunteered to participate in this experiment. Their ages ranged from 19 to 28 years. The experiment was conducted in the canteen with one participant at a time while passers-by were watching. Each participant was given a wireless microphone and told the following 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. Their voice signals were also recorded in Praat [2], at a sampling rate of 44,100 Hz and saved as a 16 Bit, Mono PCM wave file. Their voice input patterns and characteristics were also analyzed in Praat.

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) (scoring over 49= very shy, between 34 and 49 = somewhat shy, below 34 = not particularly shy) [3]. The aim was to find correlations between shyness levels, gender, and preferences and interaction patterns.

Results. Due to the conventional use of a Christmas tree, passers-by had to be informed that it was an interactive tree. Those who were with friends were more likely to come and explore the installation. The presence of friends encouraged shy people to start playing and outgoing people to continue playing. Some outgoing players seemed to enjoy making noises to cause their friends and passers-by to laugh more than to cause the bulbs to light. Other than the interaction between the player and the tree, the game-play 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

Table 1. Profile of participants in experiment 1

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	Zimbabw	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

(participant 6) tried persistently to convince his friends to play the game. When he stopped playing and handed the microphone back to the invigilator, he said that he would have continued playing if his friends joined. Another male player (participant 3) stated "my friends weren't playing so I didn't want to do it again" in the questionnaire. This could indicate embarrassment; especially that participant 3 was rated as "somewhat shy" on the shyness scale (Table 1), 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" (participants 1, 2, 4, and 10). One participant wrote that he stopped because he was "embarrassed" (participant 5). Most of the rest stopped for no particular reason while a few stopped for various other reasons including that they lost. 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 to play a voice-controlled game such as *Expressmas Tree*.

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'. Six of them 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 (Fig. 2 shows a graph).

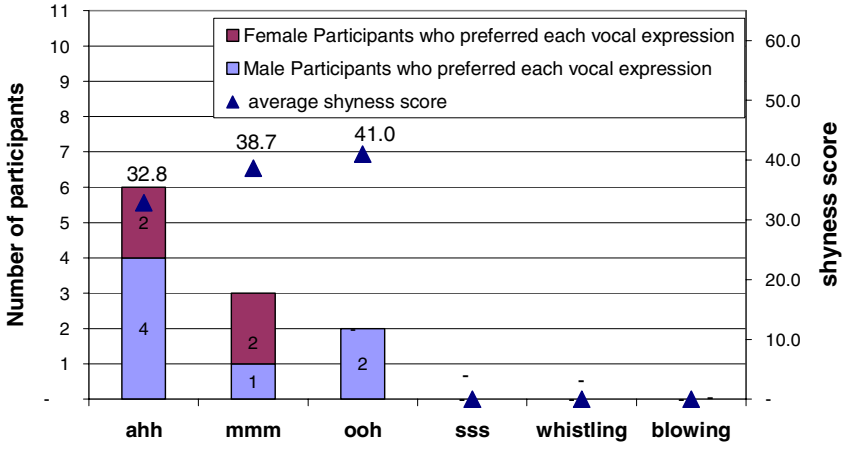

Fig. 2. Correlating the preferences, genders, and shyness levels of participants in experiment 1. Sounds are arranged on the abscissa from the most preferred (left) to the least preferred (right).

3.2 Second Experimental Design and Setting

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

Experimental Procedures. Two female students and five male students volunteered to participate in this experiment. Their ages ranged from 19 to 62 years. The simplified version of the game that the participants were presented with 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 vocalize and light up the sequence of bulbs consecutively from the bottom of the tree to the top. The experiment was conducted with one participant at a time. 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". During the first minute of game play, the number of linguistic and paralinguistic interaction attempts were noted. If the player continued to use a linguistic command beyond the first minute, the invigilator gave him/her another note with the instruction: "make non-speech sounds and whenever you want to stop, say 'I am done' ".

The participants' interaction patterns and their mostly used non-speech sounds were carefully observed and video-recorded. Their voice signals were also recorded in Praat [2], at a sampling rate of 44,100 Hz and saved as a 16 Bit, Mono PCM wave file. The duration of each continuous voice stream and silence periods were detected by the asFFT Xtra. Voice input patterns and characteristics were analyzed in Praat.

Each participant underwent a vocal endurance test, in which s/he was asked to try to light up the highest bulb possible by continuousely generating each of the following six vocal expressions: whistling, blowing, 'ahhh', 'mmm', 'ssss', and 'oooh'. These were the six types that were mostly observed by the author during evaluations of her previous work. A future planned stage of the experiement will involve more participants who will perform the sounds in a different order, so as to ensure that each sound gets 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 preferrence (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 Cheek and Buss Shyness Scale [3].

Results. When given the instruction "See what you can do with this tree", some participants didn't 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 the baubles in an effort to rearrange them. 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 aesthetic purposes seemed to have overshadowed its interactive application, despite the presence of the microphone and the computer.

Only two participants realized it was interactive; they thought that it involved video tracking and moved backward and forward to interact with it.

When given the instruction "use your voice and aim to light the highest bulb on the tree", four of the participants initially uttered verbal sounds; three uttered "hello" and one 'thought aloud' and varied his vocal characteristics while saying: "perhaps if I speak more loudly or more softly the bulbs will go higher". The three other participants, however, didn't start by interacting verbally; one was too shy to use his voice, and the last two started generating non-speech sounds. One of these two, generated 'mmm' and the other cleared his throat, coughed, and clicked his tongue.

When later given the instruction "use your voice, but without using words, and aim to light the highest bulb on the tree", two of the participants displayed unexpected patterns of interaction. They coughed, cleared their throats, and one of them clicked his tongue and snapped his fingers. They both scored highly on the shyness scale (shyness scores = 40 and 35), and their choice of input might be related to their shyness. One of these two participants 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 most of the other 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 the only participant who preferred blowing as an input. A remarkable 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 voice to cause the bulbs to react, it seems 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. Recent focus on investigating the feedback loop that may exist between the vocal input and the audio output seems to have caused the developers to overlook the possible feedback that may occur between the vocal input and the visual output.

The vocal endurance test results revealed that among the six tested vocal expressions, 'ahh', 'ooh', and 'mmm' were, on average, the most prolonged expressions that the participants generated, followed by 'sss', whistling, and blowing, respectively (Fig. 3 shows a graph). These results were based on selecting and finding the duration of the most prolonged attempt per each type of vocal expression. The following equation was formulated to calculate the efficiency of the vocal expression:

$$\text{Vocal expression efficiency} = \text{duration of the prolonged vocalization} - \text{duration of silence after the prolonged vocalization} \quad (1)$$

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, 'ahh', 'ooh', and 'mmm' were more efficient and suitable for an application that requires maintaining what this paper refers to as *vocal flow*: vocal control that involves the generation of a voice stream without disruption in vocal continuity.

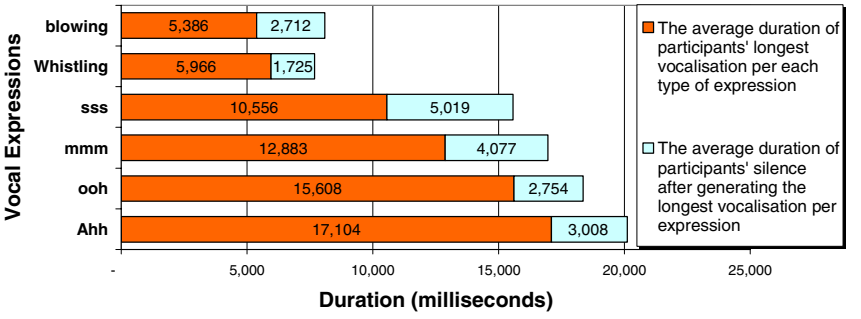

Fig. 3. The average duration of the longest vocal expression by each participant in experiment 2

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, and blowing. None of the participants preferred 'sss' or 'ooh'. The two females who participated in this experiment preferred 'mmm'. This seems to coincide with the results of the first experiment where the majority of participants who preferred 'mmm' where females. It is remarkable to note the vocal preference of one of the participants who was noticeably very outgoing and who evidently had the lowest shyness score. 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' as input. Unlike whistling which is voiceless and involves slightly protruding the lips, 'ahh' is voiced and involves opening the mouth expressively. One of the participants (shyness score = 36) tried to utter 'ahh' but was too embarrassed to continue and he kept laughing before and after every attempt. He stated that he preferred whistling the most and that he stopped because he "was really embarrassed". This participant's

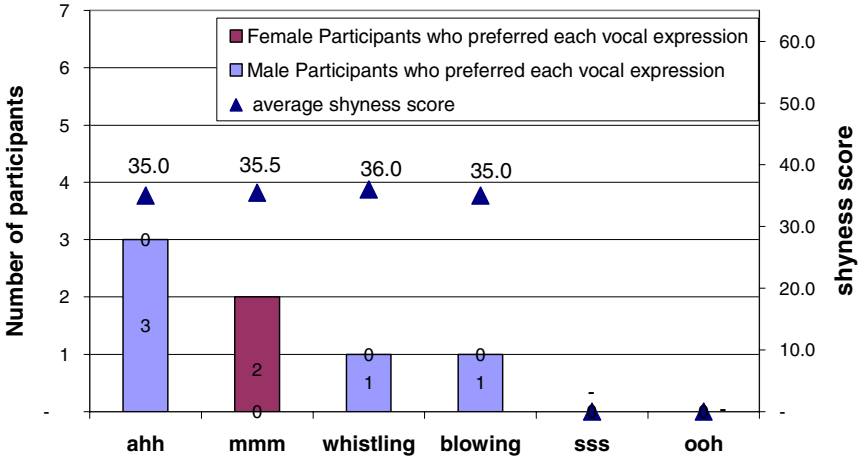

Fig. 4. Correlating the preferences, genders, and shyness levels of participants in experiment 2.Sounds are arranged on the abscissa from the most preferred (left) to the least preferred (right)

preference seems to verify the earlier hypothesis that many shy people tend to prefer whistling to interact with a voice-controlled work. This is also evident in the graphical analysis of the results (Fig. 4 shows an example) in which the participants who preferred whistling had the highest average shyness scores among others. Conversely, participants who preferred the vocal expression 'ahh' had the lowest average shyness scores in both experiments 1 and 2.

Combined results from both experiments revealed that nine of the eighteen participants preferred 'ahh', five preferred 'mmm', two preferred 'ooh', one preferred whistling, one preferred blowing, and no one preferred 'sss'. Most (seven) of the participants who preferred 'ahh' were males, and most (four) of those who preferred 'mmm' were females. One unexpected but reasonable observation from the combined results was that the shyness score of the participants who preferred 'mmm' was higher than the shyness score of those who preferred whistling. A rational explanation for this is that 'mmm' is "less intrusive to make", and that it is "more of an internal sound" as a female participant who preferred 'mmm' wrote in the questionnaire.

4 Conclusions

The paper presented a non-speech voice-controlled Christmas tree and employed it in investigating players' vocal preferences and interaction patterns. The aim was to determine the most preferred vocal expressions and the factors that affect players' preferences. The results revealed that shy players are more likely to prefer whistling or 'mmm'. This is most probably because the former is a voiceless sound and the latter doesn't involve opening the mouth. Outgoing players, on the other hand, are more likely to prefer 'ahh' (and probably similar voiced sounds). It was also evident that many females preferred 'mmm' while many males preferred 'ahh'. The results also revealed that 'ahh', 'ooh', and 'mmm' 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 'ahh', 'ooh', and 'mmm' 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. In a future planned stage of the experiments, the degree of variation in each participant's vocalizations will also be analyzed as well as the creative vocalizations that a number of participants may generate and that extend beyond the scope of the six vocalizations that this paper explored. It is hoped that the ultimate findings will provide the solid underpinning of tomorrow's non-speech voice-controlled applications and help future developers anticipate the vocal preferences and patterns in this new wave of interaction.

Acknowledgments. I am infinitely grateful to Gordon Davies, for his unstinting mentoring and collaboration throughout every stage of my PhD. I am exceedingly grateful to Stephen Boyd Davis and Magnus Moar for their lavish assistance and supervision. I am indebted to Nic Sandiland for teaching me the necessary technical skills to bring *Expressmas Tree* to fruition.

References

1. Al Hashimi, S., Davies, G.: Vocal Telekinesis; Physical Control of Inanimate Objects with Minimal Paralinguistic Voice Input. In: Proceedings of the 14th ACM International Conference on Multimedia (ACM MM 2006). Santa Barbara, California, USA (2006)
2. Boersma, P., Weenink, D.: Praat; doing phonetics by computer. (Version 4.5.02) [Computer program]. (2006) Retrieved December 1, 2006 from http://www.praat.org/
3. Cheek, J.M.: The Revised Cheek and Buss Shyness Scale (1983) http://www.wellesley.edu/Psychology/Cheek/research.html#13item
4. Schmitt, A.: asFFT Xtra (2003) http://www.as-ci.net/asFFTXtra
5. Sporka, A.J., Kurniawan, S.H., Slavik, P.: Acoustic Control of Mouse Pointer. To appear in Universal Access in Information Society, a Springer-Verlag journal (2005)
6. Wiberg, M.: Graceful Interaction In Intelligent Environments. In: Proceedings of the International Symposium on Intelligent Environments, Cambridge (April 5-7, 2006)