A Comparative Study of Pitch-Based Gestures in Nonverbal Vocal Interaction

Ondřej Poláček, Adam J. Sporka, and Pavel Slavík

***Abstract*—Nonverbal vocal interaction (NVVI) is an input modality by means of which users control the computer by producing sounds other than speech. Previous research in this field has focused mainly on studying isolated instances of NVVI (such as mouse cursor control in computer games) and their performance. This paper presents a study with 36 elderly users in which basic NVVI vocal gestures (commands) were ranked by their perceived fatigue, satisfaction, and efficiency. The results of this study inspired a set of NVVI gesture design guidelines that are also presented in this paper.**

***Index Terms*—Elderly users, nonverbal vocal input, user study, vocal gestures.**

I. Introduction

Nonverbal *vocal interaction* (NVVI) can be described as a method of interaction in which sounds other than speech are produced by the user in order to control a computer application.[1] Several approaches are described in the literature, which include using the pitch of a tone, the length of a tone, volume, or vowels in order to control the user interface. NVVI is a technique that has already received significant focus within the research community.

Pitch-based input is the part of NVVI in which the computer is controlled by the fundamental frequency of a sound signal. The user is supposed to produce a sound from which the fundamental frequency can be extracted, e.g., humming, whistling, or singing. Pitch-based input has been used as an input modality for people with motor disabilities [1]–[3] and also as a voice training tool [4]. In these applications, short melodic and/or rhythmic patterns (further referred to as *vocal gestures*) are used.

Previous research in this field mainly studied isolated instances of NVVI (such as mouse cursor control or computer games) and their performance. In most setups (see the section "Related Work"), the choice of the vocal gestures was made in a more or less *ad hoc* fashion. Questions of whether users prefer certain gestures over others, and why, have not been addressed so far in the literature.

This paper presents a study with 36 participants. The goal of the study was to compare basic NVVI pitch-based gestures in terms of perceived fatigue, satisfaction, and efficiency. We used a paired comparison paradigm [5]. The results of the study inspired a set of NVVI gesture design guidelines that are presented at the end of this paper.

The most common pitch-based gestures in current NVVI systems were selected for the experiment (see Section III-B2): flat tones, rising or falling tones, and a combination of rising and falling tones.

Manuscript received January 25, 2011; revised November 6, 2011; accepted February 20, 2012. Date of current version October 12, 2012. This research has been partially supported by the Ministerstvo Školství, Mládeže a Tělovýchovy (MSMT) under the research program MSM 6840770014. This paper was recommended by Associate Editor D. Ellis.

The authors are with the Faculty of Electrical Engineering, Czech Technical University, Prague 16627, Czech Republic.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2012.2201937

[1] See, for example, http://www.youtube.com/watch?v=Yx-M1rcsM_s.

NVVI is a low-cost technique that is relatively easy to deploy and may play an important role in the development of user interfaces for users with temporary disabilities (e.g., broken arms). While these conditions restrict the user's ability to use a keyboard and a mouse, investment in a more expensive assistive device would not be cost effective due to the limited time for which the assistive device would be used. However, devices such as a mouse or a keyboard may be emulated by NVVI. This study was conducted within the framework of the Vital Mind project, which focuses on the use of technology by elderly users. Elderly users are considered to be one of the groups for which NVVI may be useful, as they are prone to temporary disabilities. For this reason, they were selected as participants in our study.

II. Related Work

The applications of nonverbal vocal input can be roughly divided into two categories: real time and non real time. **Real-time applications** (continuous input channel) provide immediate feedback to the user while the sound is still being produced. This is useful, for example, for computer games and interactive art installations. NVVI thus does not work like speech recognition, where the system waits for the utterance to be completed.

Igarashi and Hughes [6] proposed the use of nonspeech sounds to extend the interaction facilitated by automated speech recognition. They reported that nonspeech sounds were useful for specifying analog parameters. For example, the user could produce an utterance such as "volume up, *aaah*," to which the system would respond by increasing the volume of the television for as long as the sound was held.

An example of emulating a mouse device is described by Sporka *et al.* [2], in which different nonverbal gestures control the movement of the mouse cursor and also the mouse buttons. This system was evaluated in a longitudinal study by Mahmud *et al.* [7]. A similar approach has also been used by Bilmes *et al.* [1].

NVVI has been employed successfully as a means for controlling computer games: Hämäläinen *et al.* [4] presented platform arcade games for children. Sporka *et al.* [8] demonstrated how the Tetris game can be controlled by humming, and Harada *et al.* [9] employed NVVI in several other games. NVVI has also been used as a means of artistic expression. For example, Al-Hashimi [10] described an NVVI-controlled plotter.

Non-real-time applications (event input channel) of NVVI are applications where the user is expected to finish producing nonspeech sounds before the system responds. Interaction with these systems follows the query–response paradigm, as in the case of speech-based systems. Applications of this kind are important for people who are not capable of achieving the level of speech articulation required by current automatic speech recognizers.

Ghias *et al.* [11] described query by humming, a method allowing the user to retrieve information on music tracks stored in a database, indexed by the melodies contained in them, simply by humming the melody for which the user was searching.

Watts and Robinson [12] proposed a system where the sound of whistling triggers commands in the environment of a UNIX operating system. Sporka *et al.* [13] demonstrated that non-real-time NVVI can be used for emulating a computer keyboard.

NVVI shares some similarities with speech input (typically realized through automatic speech recognition). It utilizes the vocal tract of the user and a microphone that picks up the audio signal. However, the two interaction modalities are better fitted to different scenarios, so NVVI should be considered as a complement to speech input rather as

a replacement for it. When comparing NVVI and speech input, several differences may be identified:

1) NVVI is better fitted to continuous control than speech input [8];
2) NVVI is cross cultural and language independent [14];
3) NVVI generally employs simple signal processing methods [6];
4) NVVI has limited expressive capabilities, and so speech input is better at triggering commands, macros, or shortcuts [13].

The performance of NVVI is usually lower than that of traditional input methods, e.g., a mouse or a keyboard, but it is still sufficient for cases when no alternative is available. Moving the mouse using NVVI is about three times slower [7], and the NVVI-emulated keyboard can yield type rates up to about 25 characters/min [3].

III. Experiment

The aim of this experiment was to rank the selected NVVI gestures by perceived fatigue, satisfaction, and efficiency, based on the participants' personal experience of producing these gestures.

The participants underwent a pretest interview and a training period. They were asked to use the gestures in a test application in order to accomplish a series of tasks in a simple interactive scenario. Later, they were asked to perform the following: 1) pairwise comparisons of the gesture sets and 2) a comparison of individual gestures within each set, using a forced-choice questionnaire. They were asked which gesture seemed to them more tiring, more appealing, and yielding a quicker reaction from the system. Finally, insights were solicited from the participants in a posttest interview. All comparisons were within subject.

We used the two-alternative forced-choice experiment paradigm for ranking the gesture sets. This paradigm is commonly used in human–computer interaction research to obtain reliable subjective rankings of multiple objects or categories. For example, Čadík used a similar setup to rank color-to-grayscale image conversion methods in a subjective study [15]. Ledda *et al.* employed Law of Comparative Judgement (LCJ) in a study of high dynamic range imaging [16].

A. Organization

A total of $n = 36$ participants were recruited among students of the University of the Third Age. Each participant was asked to attend three sessions in the course of a single week. The data were collected after the last session. One session typically lasted half an hour. The participants received at least one day of rest between the sessions.

1) *First session*. The purpose of the experiment was explained to the participant. A pretest interview was conducted to learn more about the participant. The experimenter explained and demonstrated the function of the test application and the task that was prepared for the participant (see the next section for details). The participant was trained to produce gestures in sets A and B and then carried out the task (using each set twice). The participant qualified for the experiment after reaching 75% accuracy, which was typically after 15 min of training.
2) *Second session*. The participant's ability to produce the gestures from sets A and B was verified. Then, the participant was trained to produce the gestures in sets C and D. Then, the participant performed the task twice, using each of sets A to D.
3) *Third session*. The participant's ability to produce the gestures from sets A to D was verified. The participant was trained to produce the last gesture set, i.e., set E. Then, the participant performed the task twice, using each of sets A to E. The order of the gesture sets in this session was counterbalanced to control for learning effects. After all the tasks had been completed, the participant was asked to fill out the quantitative questionnaire and was interviewed and debriefed by the experimenter.

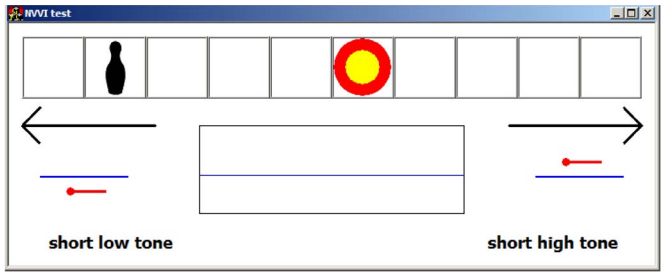


Fig. 1. User interface of the application used in the experiment for movement along horizontal axis. The cursor is in the form of a ninepin.

B. Apparatus

A test application and a quantitative posttest questionnaire were developed for the experiment.

1) NVVI Test Application: A simple test application implementing an environment for synthetic GUI tasks was developed. The user interface of the application is shown in Fig. 1.

The task for the participants was to move the cursor (represented as a black ninepin) to the target (red and yellow circle) by producing the corresponding vocal gesture from the set that was being tested. This was repeated four times in each task. The direction of movement during a task was twice to the left and twice to the right. The positions of the target and the direction toward it were randomized in each run. The distance to travel was kept constant at five cells.[2]

The rectangle below shows the immediate feedback on the voice: The red line symbolizes the pitch of the tone, and the blue line indicates the threshold pitch, separating the low and high tones. The threshold pitch can be adjusted to match the vocal range of each user. The vocal gestures to be used were depicted on the sides of the application window.

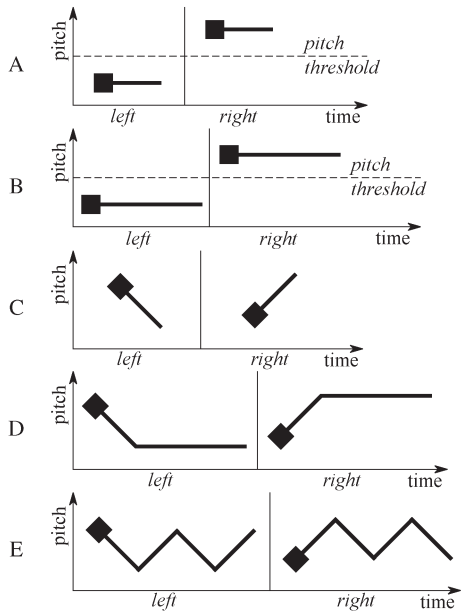
2) Selected Gestures: In this experiment, we used five different vocal gesture sets, as shown in Fig. 2. These gestures were commonly present in the current NVVI applications and research prototypes: flat tones (differing by pitch, as in [3]), rising or falling tones (tones with increasing or decreasing pitch, as in [2]), and a combination of rising and falling tones (vibrato, as in [13]).

There were only two gestures in each gesture set. They were mapped to leftward and rightward movements. NVVI applications typically employ more than two gestures. The purpose of this setup however was not to test the simultaneous use of multiple gestures but rather to expose the users to multiple gestures in a sequence, so that they could experience different kinds of gestures *in the same context of use* and thereby be able to compare them.

Both absolute-pitch and relative-pitch gestures and also those employing a continuous input channel and an event input channel were used. An autocorrelation method [17] was used to detect the pitch of the sound. The method computes the fundamental frequency in a sound, so the participant could use any sound that contained this frequency. This includes humming "*hmmm*" as well as vowels "*a*," "*ae*," "*uw*," "*ow*," etc.

The gesture set A [Fig. 2(a)] contained short flat tones. The cursor was moved by one position after recognizing the gesture (discrete event-based control). Gesture set B [Fig. 2(b)] contained long flat tones. The cursor moved continuously until silence was detected (continuous control). The gestures in sets A and B differed in the pitch of the tone (the threshold pitch was calibrated at each session during training).

[2]A demonstration of the task performed using each of the gesture sets is shown in http://www.youtube.com/watch?v=LPoSIg7uNHY.



Set	Description	Channel	Pitch
A	short flat tones	event	absolute
B	long flat tones	continuous	absolute
C	short inflected tones	event	relative
D	long inflected tones	continuous	relative
E	vibrato	event	relative

Fig. 2. Gestures used in the experiment.

Gesture sets C and D [Fig. 2(c) and 2(d)] were similar to A and B, but a relative-pitch approach was used. The movement of the cursor was determined by the initial tonal inflection of a gesture. A rising tone triggered movement to the right, while a falling tone triggered movement to the left [8].

Gestures in set E (2-E) were tones with oscillating pitch (*vibrato*). The first tonal inflection determined the movement of the cursor. With each following inflection, the cursor was moved by one cell (event-based input). The *vibrato* gestures were designed for rapid movement as long as the users could modulate their voice quickly.

3) Quantitative Questionnaire: The questionnaire was in two parts: 1) a pairwise comparison of gesture sets and 2) a pairwise comparison of the two gestures within each set. A total of five gesture sets (A to E) were compared. The comparisons were based on the following forced-choice questions. The same questions were used for both 1) and 2) with a slight difference of wording. The version of the questions for 2) is marked by brackets [].

1) (Q1) Which of these two sets of gestures [which of these two gestures] was more tiring for your vocal cords?
2) (Q2) Which of these two sets of gestures [which of these two gestures] did you like more?
3) (Q3) To which of these two sets of gestures [which of these two gestures] did the system react better?

Q1 was used as a definition of the physical difficulty of producing the gestures. Q2 and Q3 were aimed at satisfaction and efficiency, the usability attributes mentioned in International Organization for Standardization (ISO) 9241-11 (via [18]).

For the five sets of gestures, there would be ten pairwise comparisons for each question. In order to reduce the time burden, each participant performed only five randomly selected pairwise comparisons for each question.

TABLE I
PREFERENCE MATRICES FOR Q1, Q2, AND Q3

Q1	Set A	Set B	Set C	Set D	Set E
Set A	0.500	0.529	0.818	0.688	0.880
Set B	0.471	0.500	0.684	0.750	0.846
Set C	0.182	0.316	0.500	0.500	0.429
Set D	0.312	0.250	0.500	0.500	0.750
Set E	0.120	0.154	0.571	0.250	0.500

Q2	Set A	Set B	Set C	Set D	Set E
Set A	0.500	0.619	0.001	0.150	0.167
Set B	0.381	0.500	0.071	0.111	0.125
Set C	0.999	0.929	0.500	0.706	0.250
Set D	0.850	0.889	0.294	0.500	0.278
Set E	0.833	0.875	0.750	0.722	0.500

Q3	Set A	Set B	Set C	Set D	Set E
Set A	0.500	0.812	0.091	0.250	0.053
Set B	0.188	0.500	0.154	0.048	0.001
Set C	0.909	0.846	0.500	0.467	0.278
Set D	0.750	0.952	0.533	0.500	0.250
Set E	0.947	0.999	0.722	0.750	0.500

TABLE II
z-SCORES OF GESTURE SETS FOR QUESTIONS Q1 TO Q3

	Set A	Set B	Set C	Set D	Set E
Q1	0.00 (5)	0.11 (4)	0.84 (2)	0.63 (3)	1.07 (1)
Q2	1.84 (1)	1.71 (2)	0.00 (5)	0.66 (3)	0.21 (4)
Q3	1.74 (2)	2.53 (1)	0.86 (3)	0.84 (4)	0.00 (5)

Note: The order is shown in brackets.

C. Participants

The participants ($mean\ age = 66$, $SD = 5.9$) were recruited by an advertisement in a local newspaper and from the University of the Third Age. There were 22 females and 14 males. One-fourth of the participants had an academic degree, and the others had a completed secondary education. The following information was collected in the pretest questionnaire:

1) **Health state**. Five participants reported problems with their vocal cords, including hoarse voice and a mild form of dysphonia. One participant had difficulty in producing long tones, due to asthma. One participant had previously had a thyroid gland operation, which affected her performance. Three participants had a partial hearing loss; one wore a hearing aid.
2) **Music experience**. Thirteen participants reported that they used to sing or play a musical instrument. Ten participants did not sing and had no music experience. Previous music experience was not observed to impact on performance in producing voice gestures.
3) **Computer experience**. Eleven participants had a computer at home or at work, while three participants did not use computers at all. Some of them played logic games on their computers, such as cards, crosswords, sudoku, or chess.

IV. RESULTS

A. Quantitative Results

1) Comparison of the Gesture Sets: The first part of the questionnaire yielded a total of 180 pairwise comparisons for each of the questions Q1, Q2, and Q3 from the total of 36 participants.

A frequency matrix of preferences was constructed for each question (see Table I). We used Thurstone's law of comparative judgments (Case V) [5] to obtain the interval z-score scales for the gestures. The z-scores are presented in Table II.

The quantitative results are the following:

- Q1. Set A (short flat tones) was the least tiring, closely followed by set B (long flat tones). The least favorable was set E (vibrato).

TABLE III
PAIRWISE COMPARISONS BETWEEN GESTURES OF THE SAME SET

	Set	Votes for L	Votes for R	p-value	p-value Bonferroni
Q1	A: Short Flat	20	16	0.618	7.412
Q1	B: Long Flat	20	16	0.618	7.412
Q1	C: Short Inflected	10	26	0.0113	0.136
Q1	D: **Long Inflected**	9	**27**	0.00393	**0.0472**
Q2	A: Short Flat	16	20	0.618	7.412
Q2	B: Long Flat	17	19	0.868	10.415
Q2	C: Short Inflected	24	12	0.0652	0.782
Q2	D: Long Inflected	24	12	0.0652	0.782
Q3	A: Short Flat	19	17	0.868	10.415
Q3	B: Long Flat	19	17	0.868	10.415
Q3	C: **Short Inflected**	**30**	6	0.00006	**0.00072**
Q3	D: Long Inflected	26	10	0.0113	0.136

Legend: #L – number of votes in favor of gesture *Left* and #R – number of votes in favor of gesture *Right*. Significant differences are set in bold.

- Q2. Set A followed by set B was the most liked among the gestures. Set C (short inflected tones) and set E were the least favored in this aspect.
- Q3. The best response from the system was reported by the participants when using set B. The worst response was reported when using set E.

2) Comparison of the Gestures Within the Sets: The same group of participants also completed the second part of the questionnaire, in which gestures belonging to the same set were compared. Gesture set E was excluded from further data analysis as it was, in general, poorly accepted by the participants. Thirty-six participants performed one pairwise comparison per set of gestures (A to D) per question (Q1 to Q3). In each comparison, they could vote either for gesture *Left* or for gesture *Right*.

These comparisons could answer the following question: "For one gesture set, is there a significant preference among the participants for one gesture over the other?" This is a Bernoulli experiment [19] for which a binomic test can be used. The null hypothesis holds that the true probability of either choice is 0.5.

Since a total of 12 comparisons were performed (three questions $\times$ four sets of gestures), in order to reduce the risk of type I error, a Bonferroni adjustment [20] of the p-value level was performed. In order for a result to be considered significant, the p-value must be less than 0.05/12. An overview of the results is shown in Table III.

Gesture *Right* of set D (long inflected tones) was significantly more tiring (Q1) for the participants than gesture *Left*. A similar trend could be observed for set C, but the difference was not significant. For set C, the system was perceived to react significantly better to gesture *Left* than to gesture *Right*.

B. Qualitative Results

1) Short Flat Tones: Participants did not have serious problems when producing short flat tones (set A). Two participants produced *"la la la"* instead of humming. This was not considered as an error, as the input was based on the pitch of the tone only. Several participants were confused about the direction of movement at the beginning of the task, and two participants had difficulties producing a correct tone, although they were able to complete the tasks successfully.

2) Long Flat Tones: The long flat tone (set B) task was also completed by all participants. They mostly appreciated the immediate feedback of movement and the simplicity of the gestures. They identified those gestures as easier and less fatiguing than other gestures, mainly because they did not need to repeat gesture by gesture and could do everything by one long tone.

3) Short Tones with Inflection: Most participants struggled with short tones with inflection (set C). Six participants were not able to learn these gestures at all, and therefore, they could not complete the task. Approximately half of the rest had significant problems producing these gestures. Only one participant stated that these gestures were simpler than the others, because the absolute pitch of the tone was not important, and another participant enjoyed this task. Other comments were mainly negative. We observed that participants were more successful when producing the rising tone than when producing the falling tone.

4) Long Tones with Inflection: Participants faced similar problems with long tones with inflection (set D) to the problems with short ones. Again, falling tones were more difficult for some participants to produce than rising tones. Nine participants were not able to complete this task successfully.

5) Vibrato: The most difficult task was the vibrato (set E). Twelve participants skipped this task. They were usually confused by the direction of the gestures. Several participants identified these gestures as the worst. Only one participant liked the vibrato gestures more than short inflexion tones.

Participants differed in their comparisons of the long and short tones. Several participants claimed that long tones were more demanding than short ones, because they needed to hold their breath for a long time. On the other hand, several participants said that short tones were more demanding for them, because they needed to start the tone over and over.

The participants were asked to identify their favorite and least favorite gesture set. Seven participants liked flat tones, e.g., "They were easier for me" and "I did not feel embarrassed." Nine participants disliked one of inflected tones (including vibrato), e.g., "I do not have my voice trained enough." The qualitative results suggest that flat tones (sets A and B) are more accepted than inflected tones (sets C, D, and E).

6) Perception of Humming: Twelve participants did not feel comfortable. They mainly made comments such as "I felt like a fool," "It was funny," and "I felt like a small child." Several participants also reported that the voice gestures reminded them of animal sounds. However, five participants reported that they did not feel any embarrassment when producing humming.

7) Voice Fatigue: Ten participants reported that they did not feel any fatigue during the experiment. Four participants complained about mild fatigue.

V. DISCUSSION

The results presented earlier indicate that gestures using absolute pitch mapping (gesture sets A and B—flat tones) were well accepted by the users. Preference for a higher tone or for a lower tone was highly individual. These gestures can be used in both event and continuous input channels. The disadvantage of these gestures is the need for manual threshold pitch adjustment.

Gestures that use relative-pitch mapping (gesture sets C and D—inflected tones) were found to be more difficult to produce and were therefore not very well accepted by the users. An interesting point is that rising tones were significantly better accepted than falling tones.

Very few users accepted vibrato (gesture set E).

A. Guidelines for the Design of Pitch-Based Gestures

We have summarized the results into four guidelines for the use of designers of future pitch-based applications.

1) **Use flat tones if possible**. This experiment demonstrated that flat tones were easiest for the users to produce. This is consistent with the finding reported by Sporka *et al.* [13]. Any design of

NVVI gesture assignment should therefore commence with flat tones. Other types of gestures should be used only in addition to flat tones.

2) **Use absolute pitch mapping if possible**. Absolute pitch mapping is better accepted by the users. Relative pitch should therefore be used only when more vocal gestures need to be assigned. In addition, splitting the vocal range into more than two vocal gestures is tricky, as more precise intonation is needed [21].
3) **Use positive rather than negative inflection gestures**. This experiment demonstrated that tones with decreasing pitch were more difficult to produce. If there is a need for relative-pitch mapping, rising tones should be preferred over falling tones.
4) **Do not use more than one inflection per gesture**. Gestures with pitch oscillation were not well accepted by the users in this experiment. The existing literature reports successful use of gestures with a single inflection [8] but difficulties with complex gestures [13].

NVVI applications support more complex tasks than 1-D movement. An example of a complex task of this kind is playing the Tetris game [8] or controlling a mouse cursor [2]. Designers may combine various types of gestures when a higher number of input signals are needed. Frequent operations should be assigned to simple gestures. For example, in a hand-free mouse [22], the short tone was used for the most frequent operation—left click—while scrolling was mapped to inflected tones.

VI. Conclusion

The study described in this paper has shown how users perceived various aspects of NVVI gestures: fatigue, satisfaction, and efficiency. Among numerous gestures that we could have chosen from, we focused on the basic pitch gestures that are present in the current NVVI literature: flat tones (i.e., tones with constant pitch), rising or falling tones, and gestures with oscillating pitch.

The study was performed with a group of 36 elderly users. Simple horizontal cursor motion tasks were used as stimuli for the participants. Each task could be carried out using only two gestures, for leftward motion or for rightward motion. The participants were exposed to five sets of gestures (ten gestures in total). They experienced different sets of gestures in the same context of use and could therefore make a comparison between them. We used the paired comparison paradigm, which is commonly employed in the field of human–computer interaction for subjective ranking of stimuli.

The most acceptable sets were those with tones of constant pitch, followed by gestures with rising or falling pitch. Gestures with multiple changes of pitch (vibrato) were found unacceptable. Individual gestures were compared within the gesture sets. The users reported that a short rising tone was significantly less fatiguing than a falling tone.

A small number of design recommendations for pitch-based gestures were formulated. Any design of NVVI gestures should start with flat tones, and other types of gestures should be included only when the required number of the gestures increases.

Future Work. This study has focused on vocal gestures produced in a laboratory environment. A further study is needed to investigate the acceptability of NVVI in environments with a reduced amount of privacy: streets, offices, etc. In this paper, NVVI was used by elderly Czech users. Levels of acceptance may vary in different social and cultural contexts. It will be interesting to study this aspect of NVVI in a cross-cultural experiment.

References

[1] J. A. Bilmes, X. Li, J. Malkin, K. Kilanski, R. Wright, K. Kirchhoff, A. Subramanya, S. Harada, J. A. Landay, P. Dowden, and H. Chizeck, "The vocal joystick: A voice-based human–computer interface for individuals with motor impairments," Univ. Washington, Seattle, WA, Tech. Rep. UWEETR-2005-0007, 2005.

[2] A. J. Sporka, S. H. Kurniawan, and P. Slavík, "Whistling user interface (U3I)," in *Proc. 8th ERCIM Int. Workshop 'User Interfaces for all'*, vol. 3196, *LCNS*, Vienna, Austria, Jun. 2004, pp. 472–478, Springer-Verlag Berlin Heidelberg.

[3] A. J. Sporka, T. Felzer, S. H. Kurniawan, O. Poláček, P. Haiduk, and I. S. MacKenzie, "CHANTI: Predictive text entry using non-verbal vocal input," in *Proc. Annu. CHI*, 2011, pp. 2463–2472.

[4] P. Hämäläinen, T. Mäki-Patola, V. Pulkki, and M. Airas, "Musical computer games played by singing," in *Proc. 7th Int. Conf. Digital Audio Effects*, G. Evangelista and I. Testa, Eds., Naples, Italy, 2004, pp. 367–371.

[5] L. L. Thurstone, "A law of comparative judgements," *Psychol. Rev.*, vol. 34, no. 273–286, 1927.

[6] T. Igarashi and J. F. Hughes, "Voice as sound: Using non-verbal voice input for interactive control," in *Proc 14th Annu. ACM Symp UIST*, 2001, pp. 155–156.

[7] M. Mahmud, A. J. Sporka, S. H. Kurniawan, and P. Slavík, "A comparative longitudinal study of non-verbal mouse pointer," in *Proc. INTERACT, Part II*, vol. 4663, *Lecture Notes in Computer Science (LNCS)*, Rio de Janeiro, Brazil, 2007, pp. 489–502, Springer-Verlag: Berlin, Gemany.

[8] A. J. Sporka, S. H. Kurniawan, M. Mahmud, and P. Slavík, "Non-speech input and speech recognition for real-time control of computer games," in *Proc. 8th Int. ACM SIGACCESS Conf. Comput. Accessibility. ASSETS*, 2006, pp. 213–220.

[9] S. Harada, J. O. Wobbrock, and J. A. Landay, "Voice games: Investigation into the use of non-speech voice input for making computer games more accessible," in *Proc. 13th IFIP TC Int. Conf. Human-Comput. Interact.—Volume Part I, INTERACT'11*, 2011, pp. 11–29.

[10] S. Al-Hashimi, "Blowtter: A voice-controlled plotter," in *Proc. HCI Engage, 20th BCS HCI Group Conf. Co-Oper. ACM*, London, U.K., Sep. 2006, vol. 2, pp. 41–44.

[11] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: Musical information retrieval in an audio database," in *Proc. 3rd ACM MULTIMEDIA*, 1995, pp. 231–236.

[12] R. Watts and P. Robinson, "Controlling computers by whistling," in *Proc. Eurograph.*, Cambridge, U.K., 1999.

[13] A. J. Sporka, S. H. Kurniawan, and P. Slavik, "Non-speech operated emulation of keyboard," in *Proc. CWUAAT Des. Access. Technol.*, J. Clarkson, P. Langdon, and P. Robinson, Eds., 2006, pp. 145–154.

[14] A. J. Sporka, "Non-speech sounds for user interface control," Ph.D. dissertation, Czech Technical Univ. Prague, Prague, Czech Republic, 2008.

[15] M. Čadík, "Perceptual evaluation of color-to-grayscale image conversions," *Comput. Graph. Forum*, vol. 27, no. 7, pp. 1745–1754, Oct. 2008.

[16] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 640–648, Jul. 2005.

[17] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 1, pp. 24–33, Feb. 1977.

[18] A. Holzinger, G. Searle, T. Kleinberger, A. Seffah, and H. Javahery, "Investigating usability metrics for the design and development of applications for the elderly," in *Proc. 11th ICCHP*, 2008, pp. 98–105.

[19] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984, ch. Bernoulli Trials, pp. 57–63.

[20] J. P. Shaffer, "Multiple hypothesis testing," *Annual Review of Psych.*, vol. 46, pp. 561–584, Feb. 1995.

[21] A. Sporka, S. Kurniawan, and P. Slavík, "Physical human factor in non-speech input," in *Proc. CHI 2007 Workshop Striking C[h]ord: Vocal Interaction in Assistive Technologies, Games, More*, 2007, pp. 13–16.

[22] O. Poláček and Z. Míkovec, "Hands free mouse: Comparative study on mouse clicks controlled by humming," in *Proc. 28th Int. CHI EA*, 2010, pp. 3769–3774.