

Pitch in Non-verbal Vocal Input

Adam J. Sporka

Department of Information and Communication, University of Trento
adam@sporka.eu

Introduction

Recently, numerous prototypes of user interfaces have been presented that are based on interpretation of non-verbal sounds produced by the users, such as humming, whistling, or hissing. These sounds can be characterized by numerous properties, such as pitch, volume, or timbre. The user may intentionally change these properties while producing the sound. The properties (or their profiles over time) can be mapped to different actions. In order to trigger an action or modify an input value, the user produces a corresponding acoustic gesture.

The non-verbal vocal input (NVVI) represents an inexpensive alternative to various state-of-art technologies and techniques, such as sip-and-puff controllers or eye trackers, as no hardware additional to a standard PC is necessary.

Applications of NVVI that have been developed so far include systems able to control mouse cursor pointer [2, 7], operate computer games [3, 6], create artwork [1], or emulate keyboard [8]. Some of the user interfaces are controlled by sound timbre or volume, such as [2], some others by pitch, such as [7] or [4]. Most of these systems have been developed and deployed in the context of accessibility.

In some non-speech user interfaces, the input from the users is solicited through absolute pitch of the tone they produce. For example, in [3] the absolute pitch of tone determines the absolute vertical position of a game character.

In the pitch-to-address mapping mode of the system for NVVI emulation of keyboard (see [8], p. 147) the users produce sequence of three tones (A1, A2, A3) in order to specify the coordinates of keys on the keyboard (see Fig. 1).

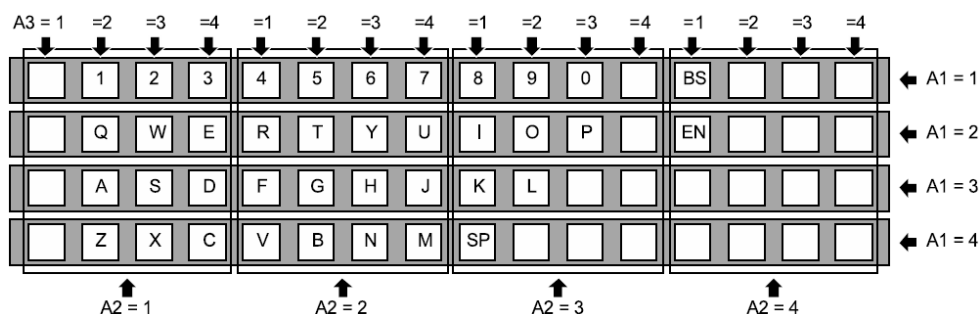


Figure 1. The address space of the keys for the NVVI keyboard. (From [8].)

Each coordinate is a number from a given interval of integers $\langle 1, N \rangle$ (The authors of the system selected $N = 4$). Each integer is assigned a pitch interval (see Fig. 2). In order to select the desired integer, the user produces a tone that falls into the corresponding interval. Significant for selection is the pitch before the very end of the tone. The users may

utilize the visual feedback that displays the current pitch of the tone in relation to the available intervals.

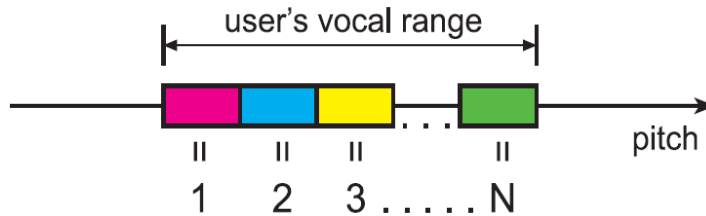


Figure 2. Selection of a value by absolute pitch of tone

There are two important physical human factors in this mode of interaction, determined by individual physiological properties of the user's vocal organs:

- The vocal range, i.e. the interval of pitch that the user is capable to produce.
- The precision of intonation, especially the ability of the user to finish a tone within given interval.

Obviously, the larger the user's vocal range is and the more precise the user's intonation is, the wider set of values can be presented to the user to choose from.

The aim of this paper is to investigate these two factors. We describe a new method for how the user's vocal range can be measured. Later in the text we present results of our study in which we quantify the ability of people to produce tones that fall into desired intervals as in the pitch-to-address method for different widths of the intervals.

Note: Unless explicitly stated, all pitch values are written in MIDI note numbers. According to the MIDI Tuning Standard [5], note number p is obtained from a frequency f as $p = 69 + 12 \log_2(f / 440)$. The tone middle C (261.6 Hz) corresponds to the MIDI note number 60. A tone n semitones above or below corresponds to the MIDI note number $60 \pm n$.

Comfortable Range of Pitch

Constrained by anatomical and physiological properties of the voice and the speech organs, each person has their own general range of pitch $\langle p1, p2 \rangle$ that he or she is able to produce (see fig. 3). The trained singers can reach as much as two octaves while the range of pitches of the untrained people is usually narrower. The difficulty of producing a tone varies with its pitch. There is only a subrange $\langle pc1, pc2 \rangle$ in which lay the tones that are comfortable enough to be produced over an extended period of time (the comfortable range of pitch, CPR).

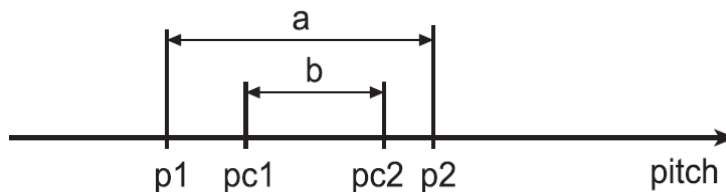


Figure 3. User Pitch Ranges. a—general range, b—comfortable range. See text or explanation of the variable names.

In order to minimize the fatigue, it is necessary to match the non-speech user interface to the comfortable range during the system calibration, i.e. to specify the $\langle pc1, pc2 \rangle$ interval. We have developed a two-step method for estimation of the values $p1$, $p2$, $pc1$, and $pc2$ in a user. The method does not require any previous musical training on the users' part. It is based on the visualization of the pitch being produced on a tonal scale. It is a two-step process, as described below:

- **Step 1: Familiarizing the voice.** The user is made aware of their voice, especially how easy it is to produce tones at different pitch. The user goes through a sequence of attempts. In each attempt, the user is requested to produce a tone that is within a given target interval, as shown in the stimulus (see Fig. 4). The sequence of attempts should be designed in a way that the user experiences all possible pitches and his or her ability or inability to produce these.
- **Step 2: Self-assessment of difficulty.** The users go through the same sequence of attempts. However, in this step, the user is asked to evaluate how comfortable for them it was to produce a tone within the given target interval by assigning grades 1 (easy) through 5 (impossible) after producing each tones. A typical user's response is shown in the chart on Fig. 5. The pair of values $pc1$ and $pc2$ corresponds to the horizontal position of the bottom of the 'canyon' in the chart. Similarly, the pair of values $p1$ and $p2$ correspond to the edge of the canyon.

A User Study

We have used this method to measure the pitch range in 13 adult participants (9 males, 4 females, average age 37 years, SD 17), each one in an individual session taking place in a quiet room. Three of the participants had taken some music training as children. One of the participants was an advanced piano player. All participants wore headsets during the experiment. The sound signal was recorded and analyzed with their consent.

The intervals in the stimuli used in our experiment were 4 semitones wide. The intervals were selected so that they cover the complete range of human voice (55 to 1760 Hz) and that no more than two target intervals would overlap at any given pitch. A stimulus presented to the user during one attempt is shown in Fig. 4.

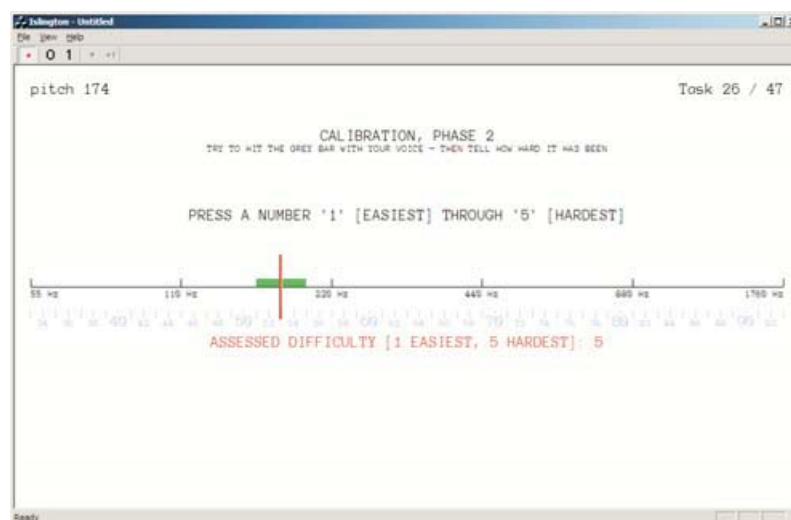


Figure 4. A stimulus. Green bar—the target interval. Vertical red line—the current pitch of tone.

The values gathered during the test are shown in Table 1. The p-values are the centers of the test intervals. The lines of the table are ordered by the width of the comfortable range interval $\Delta c12 = pc2 - pc1$.

Age / Gender	p1	pc1	pc2	p2	$\Delta c12$
29/M	43	47	51	65	4
27/M	39	45	49	69	4
22/M	39	43	51	59	8
49/F	55	57	67	73	10
59/M	35	43	53	61	10
22/M	35	43	55	81	12
24/M	39	45	59	65	14
64/F	39	47	61	71	14
51/F	43	47	61	69	14
25/M	39	45	61	71	16
61/M	35	39	57	61	18
28/F	47	53	73	77	20
24/M	39	41	63	67	22

Table 1: Participants' Ranges

The average comfortable range was 12.7 semitones, SD 5.57. Only less than one third of the participants was not able to produce tones in ranges at least 10 semitones wide.

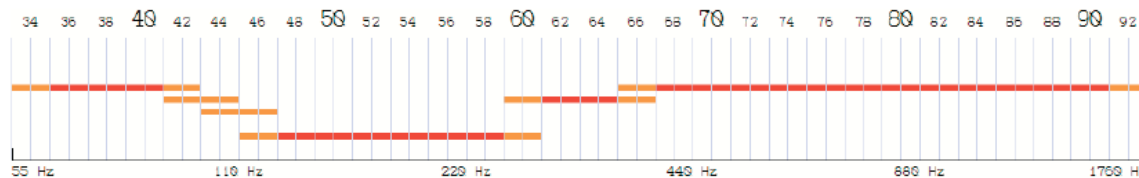


Figure 5. An example of pitch difficulty chart.
Horizontal axis: pitch [MIDI note numbers, Hz]. Vertical axis: user-reported difficulty.

Width → [semitones]	1	2	4	8
Too low	20%	14%	8%	11%
Correct	48%	54%	77%	81%
Too high	32%	32%	15%	8%
# of datapoints	112	112	111	75

Table 2. Classification of attempts by target interval width and outcome

Precision of Intonation

The purpose of this study was to compare the difficulty of intonation of tones for different required precision of this intonation. The precision was determined by the width of target interval. E.g. when a 4-semitone target interval C4–E4 was presented, the user was supposed to produce a tone that would terminate at pitch that would be between these two notes.

Method

19 users (13 males, 6 females, average age 32 years, SD 15.9) participated in the study. Two of the participants had taken some music training as children. One of them was an active musician. None of the participants had any prior experience with using this input method. All participants wore headsets during the experiment. The participants were informed on the purpose of the study and gave the consent to record and process the voice signal. The session with each participant consisted of the following two steps:

- Step 1. The participant's comfortable pitch range $pc1$, $pc2$ was measured by the method described in the previous section.
- Step 2. The participant went through a sequence of attempts. The goal of each attempt was to hum a tone whose pitch at the very end of the tone falls into a target interval. The stimulus for each attempt was very similar to the one as shown in Fig. 4. The attempts varied in target interval width (1, 2, 4, or 8 semitones) and in the placement of the target interval on the tonal scale. Only such target intervals were selected that would fit into the participant's comfortable pitch range. This way, there were usually 10 to 20 attempts in the sequence for each participant. The following data were recorded for each attempt: The width and the placement of the target pitch interval, pitch profile of the tone, duration of the tone, and the overall outcome (either "success", "tone released too low", or "tone released too high").

Results

Success rate. The Table 2 shows the counts of the stimuli, grouped by the interval width and the outcome of each attempt. The data show a trend that with increasing width of the target interval the success rate rises.

Time Requirements. The average time needed to produce a tone within a given interval width is shown in Fig. 6. The time needed to produce a tone within 1-semitone interval was significantly greater than for wider intervals (a *t*-test; $p \leq .05$). Only successful attempts of all users were included into this analysis. Fig. 7 shows the histogram of tone durations with 0.5 second bins. The majority of all tones were around 1 and 1.5 seconds of duration.

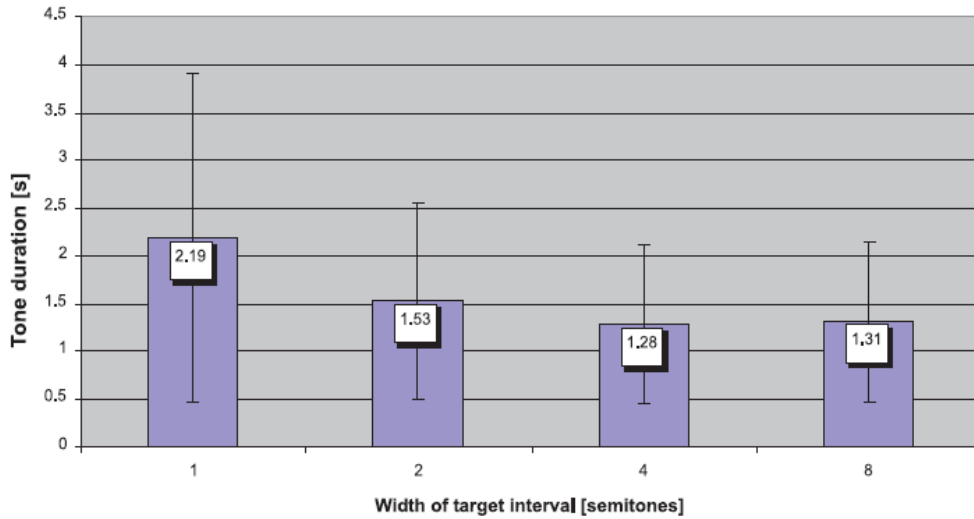


Figure 6. Histogram of tone durations in successful attempts

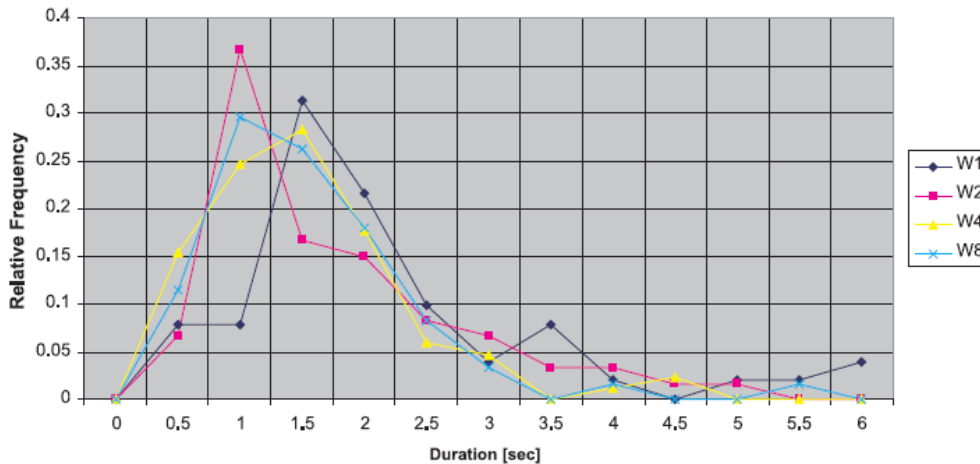


Figure 7. Histogram of tone durations in successful attempts

Discussion

On average, the comfortable pitch range was slightly wider than one octave (12.7 semitones). However, there were two people who were unable to produce a wider range than 4 semitones.

The participants were generally more successful when hitting wider target intervals. The success rate for 4-semitone interval was 77%. A further increase of width brought only a minimal improvement. We conclude that the width of 4 semitones was the most suitable target interval when considering the average width of the comfortable range. This way, the users would be able to select from 3 or 4 different values at a time.

Producing tones in order to fit 1-semitone intervals took 2.2 seconds on average, which was significantly longer than aiming at wider ones. The difference between duration of tones for 1-semitone and 4-semitone interval was almost 1 second. The shortest tones were produced for 4- and 8-semitone intervals.

Conclusion

In this paper we discussed some of the human factors that affect the performance of the pitch-to-address selection method described in [8]. Depending on the pitch of tone produced by user, a value is selected. The performance of this input method is affected by the users' vocal range and their ability of precise intonation.

We have described a new method of measurement of the comfortable range of pitch that does not require any previous musical training of the user and is easy to deploy in the end-user applications. Using this method we have measured the comfortable range of pitch in a small sample of general population. Our user study of the precision of intonation has indicated that most people can select from three or four different discrete values at once by means of pitch-sensitive non-speech sound input devices. Thus, we have verified the design of the original prototype of the pitch-to-address method of non-speech emulation of keyboard.

In some applications of the non-speech input (such as emulation of keyboard) it is necessary that users do not rely on visual feedback to the sound they produce in order to eliminate the second focus of attention. Our future work should therefore include a study of precision of intonation in which the users would not be presented the visual feedback. It may be expected that people with musical skills would outperform people without these skills. A larger user study that would be able to quantify the difference in performance of these two groups should be performed as well.

Acknowledgments

We would like to express our thanks to the students who volunteered to run the experiments. Namely, to Vojtech Jirkovsky, Ivo Jirele, Martin Strnad, Jiřka Trojankova, and Lukas Wroblewski. This work has been supported by internal grant 712913 (2007) of the Czech Technical University in Prague. The work has been also supported by Student Research Project contest held by the CTU in Prague and IBM Czech Republic.

References

- [1] S. Al-Hashimi. Blowtter: A voice-controlled plotter. In Proceedings of HCI 2006 Engage, The 20th BCS HCI Group conference in co-operation with ACM, vol. 2, London, England, September 2006.
- [2] 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. In Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing, October 2005.
- [3] P. Hämäläinen, T. Mäki-Patola, V. Pulkki, and M. Airas. Musical computer games played by singing. In G. Evangelista and I. Testa, editors, Proceedings of 7th International Conference on Digital Audio Effects, Naples, Italy, pages 367–371, 2004.
- [4] T. Igarashi and J. F. Hughes. Voice as sound: using non-verbal voice input for interactive control. In UIST'01: Proc 14th Annual ACM Symp on User Interface Software and Technology, pages 155–156, New York, NY, USA, 2001. ACM Press.
- [5] MIDI Manufacturers Association. Complete MIDI 1.0 Detailed Specification. March 1996.
- [6] A. Sporka, S. H. Kurniawan, M. Mahmud, and P. Slavik. Non-speech input and speech recognition for real-time control of computer games. In The Proceedings of The Eighth

International ACM SIGACCESS Conference on Computers & Accessibility, ASSETS 2006, Portland, Oregon. ACM, 2006.

- [7] A. J. Sporka, S. H. Kurniawan, and P. Slavík. Accoustic control of mouse pointer. *Universal Access in the Information Society*, 4(3):237–245, 2006.
- [8] A. J. Sporka, S. H. Kurniawan, and P. Slavík. Non-speech operated emulation of keyboard. In J. Clarkson, P. Langdon, and P. Robinson, editors, *Cambridge Workshop on Universal Access and Assistive Technology, CWUAAT 2006. Designing Accessible Technology*, pages 145–154. Springer-Verlag London Ltd, 2006.

About the authors:



Adam J Sporka is a senior research fellow (Marie Currie programme) at the University of Trento, Italy. He received his PhD at the Czech Technical University in Prague. In his research, he focuses on speech user interfaces and non-verbal vocal input for emulation of input devices of personal computing equipment. He wrote or contributed to about 25 papers and articles published in scientific journals and proceedings of various international conferences. He was one of the organizers of a first workshop on non-verbal vocal interaction at the ACM CHI 2007 conference. He is also a freelance consultant in HCI and software development. His clients include Czech Academy of Sciences and Prague Philharmonic Choir.