What does beep mean? – context free interpretation of short sinus wave stimuli

Matthias Wille, Sabine Theis, Peter Rasche, Christina Bröhl, Rebecca Kummer, & Alexander Mertens
Institute of Industrial Engineering and Ergonomics, RWTH Aachen University
Germany

Abstract

Although human-machine-interaction had become more technically mature through the last decades many devices still communicate with simple sinus beeps. Those beeps can be differentiated based on rhythm and tone pitch, but they are very abstract signals, which can be only interpreted by knowing the device very well or if the manual was studied profoundly. Therefore, it would be interesting if there is a kind of common structure in the beep appearance that is interpreted context free by all humans. For instance, fast sequences of beeps indicate an alarm or ascending pitch means a question while descending pitch means a statement (like typical human intonation). Hereby, possible age effects – based on technical generations and according experiences – should be taken into account. In our study 13 younger and 13 elderly participants were confronted with 8 sinus beeps that differ in rhythm and pitch and had to rate each of them on 5 continuous scales: alarm - note, question - confirmation, pleasant - unpleasant, important – unimportant, distinct – inconclusive. Results showed significant effects on 3 of the 5 scales meaning that a common interpretation in these dimensions exists. This can be important for designing auditory signals based on sinus waves.

Introduction

During the last decades technology and therefore also human-machine interaction has developed rapidly. Displays have become smaller, brighter and with higher resolution. Speech interaction has left its niche and is now possible on most smartphones. And with the internet of things machines have improved to communicate with themselves: A refrigerator can now order milk by itself without contacting its human master if it registers a lack of milk. But although communication with machines has developed so far, most of the machines still communicate with simple sinus tone sequences with their human masters for information input or input confirmation. These beeps can be sent by a washing machine, car radio, medical devices, smartphone, computer or anything else. The reason for these beeping machines may be the simple and unexpensive feasibility of a sinus tone generator which at least do the job to catch the human attention. The human operator might know the meaning of this beep-signal, because he is used to...
the machine or eventually (but unrealistic) he has read the manual. However, in most cases we only know that the machine wants something from us, but not the concrete content. This study investigates a common interpretation in beep sequences depending on pitch variation and speed.

The study took place within the “Tech4Age” Project (www.tech4age.de). Within this project we develop a pattern language of how human machine interaction should be designed for elderly users in healthcare context. In context of another study on multimodal perception (Wille et al., 2016) the question emerged, if sinus tone sequences alone can have some common meaning or interpretation that influence the perception and action of participants.

Theoretical background

Searching the literature about interpretation of sinus tone sequences no study or description can be found with that concrete topic. For sure, already Helmholtz (1863) had focused on sound perception and the research on auditory displays in common has a long history: Adam and Tucks (1976) for instance tested reaction time to different warning signals in different ambient noise enviroments. They found out that reaction time to these signals do not only refer to their intensity but also to their structure: Reaction to a “wail” sound (a slow variation of frequency between 400 and 925 Hz over several seconds) was much slower than to a “Yeow” sound (descending change in frequency from 800 to 100 Hz every second). But the faster reaction to more intense sounds is only true for simple reaction tasks and not for choice reaction tasks as Van der Molen and Kuess (1979) found out. That illustrates, that the structure of a sound has influence, but have to be seen in interdependency with the task. In common it is known that audiicons which simulate meaningfull sound are better for HCI than abstract earcons (Brewster et al., 1993; Gaver, 1986; Blattner et al., 1989 and many more). Garzonis et al. (2009) compared the effectiveness of earcons and audiicons in term of their intuitiveness, learnability, memorability and user preference and found out that audiicons significantly perform better than earcons across all four measures. If using abstract earcons Edworthy et al. (2011) have found that the learnability of a set of earcons is greatly enhanced by avoiding similar temporal patterns and increasing the range of type of sounds.

Many other aspects of sound has been investigated, even the pitch of sounds as a perceptual analogue to odor quality (Belkin et al., 1997). But if it comes to simple sinus tones and the interpretation of their sequence in a common, inherent and context free way literature is missing or at least was not found.

However, sinus tone sequences have been investigated about their emotional implication (Scherer & Oshinsky, 1977). Descending melodies are associated with pleasantness and upscending melodies are also interpreted as being pleasant as long as the pitch variation is big enough. This stands in line with Smith and Cuddy (1986) who found out, that sinus tone sequences are interpreted in general as more pleasant as more melodic they are. So it can be seen as proved that sinus tone sequences have a common emotional interpretation. Therefore the question pops up if there is also a common rational interpretation of sinus tone sequences, like a sequence is interpreted as more or less urgent signal.
Method

To investigate if simple sinus tone sequences are interpreted in the same way from different people a laboratory-study was conducted where 8 sinus tone sequences with up to three tones in up to three pitches were randomly presented on a smartphone and participants rated their impression on 5 scales immediately after hearing the tone.

Participants

26 subjects participated in the study with an age of 16-76 years. The sample was divided by median split into two age-groups to investigate possible age effects. In the younger group were 13 subjects between 18-28 years (Mean = 22.38, SD = 2.434, 7 male /6 female). In the older group were 13 subjects aged 49-76 years (Mean = 64.23, SD = 9.391, 4 male / 9 female). All participants had normal or corrected to normal sight and normal hearing abilities.

Material – the sinus tone sequences

Each stimulus consisted of 5 time-units which were 100 msec long and were filled with a tone or a pause where no sound is played. So overall each stimuli was ½ second long. The tones were pure sinus waves with three different pitches that were matched harmoniously (C6 - 1046.50 Hz; G5 - 783.99 Hz; C5 - 523.25 Hz). Furthermore, all pitches were selected in an area where no influence of presbycusis is to be expected. Figure 1 shows the structure and the naming of the stimuli: Each stimulus was assigned by a 5-digit number which reflects the sequence of tones and pauses. A pause was indicated by 0, the lowest pitch (C5) by 1, the middle pitch (G5) by 2 and the highest pitch (C6) by 3. Longer tones in a sequence (like in 11033 or 22222) are played continuously without any break or new attack envelope in between.

<table>
<thead>
<tr>
<th>10203</th>
<th>11033</th>
<th>13013</th>
<th>20202</th>
</tr>
</thead>
<tbody>
<tr>
<td>22222</td>
<td>30201</td>
<td>33011</td>
<td>31031</td>
</tr>
</tbody>
</table>

Figure 1. Structure and naming of the sinus tone sequences.

Scales

Participants had to rate each stimulus on 5 continuous scales. These scales were based on expert interviews and a short evaluation regarding intelligibility with participants during pretesting (these participants are not part of the sample described here). The scales are listed below with the original German dimensions in brackets and followed by a brief explanation which was given to the participants also during introduction.
alarm - note („Alarm – Hinweis“)
An alarm calls for fast action while a note is not time critical.

question - confirmation („Frage – Bestätigung“)
A question requires a timely response, while during a confirmation no response is required (for example, confirmation-tone after a command entry)

pleasant – unpleasant („Angenehm – Unangenehm“)
A signal can be perceived as pleasant or unpleasant by you.

important – unimportant („Wichtig – Unwichtig“)
A signal can be perceived as important or unimportant by you.

distinct – inconclusive („Eindeutig – Uneindeutig“)
If the stimulus could be clearly assigned to the presented scales it is distinct if differentiation is more difficult, then the stimulus is inconclusive.

Material – the experimental application and used hardware
The experimental setup was built as a native Android app and presented on a Nexus 5 smartphone running Android 6.0. The volume was fully turned up to ensure all participants perceived the stimuli clear and with the same volume. Figure 2 shows a screenshot of the application. Sounds were played once by pressing the “next” button on the lower right, but can be repeated as often the participant want by pressing the “repeat” button on the top right. Scales were presented as continuous horizontal faders with both endpoints named in german language. By default the middle position was displayed for each new stimulus. Participants rated the scales by moving the fader into the desired position and each scale which was already rated became grey (like the first two in figure 2). Not yet rated scales were shown in black (as the last three in figure 2). Participants had to rate on all scales before the application allowed going to the next stimulus by pressing “next”. If not all scales were rated a message box popped up when pressing “next” and told to rate all scales of the actual stimulus first. This was to ensure that all participants rate all stimuli and not skip some scales by accident or by disinterest. If participants prefered the middle position given in the beginning they can rate it that way by just touching the scale for a short time and bringing the signifier back into the middle position. The internal resolution on each scale was 0-1 with 3 decimal places which was transferred to 0 – 1000 afterwards.
what does beep mean?

Figure 2. Screenshot of the experiment app.

Procedure

This study was conducted as an appendix to another study about multimodal perception (Wille et al., 2016; more to come), where participants had to categorize multimodal stimuli consisting of visual, audio and tactile material whether they were rhythmic or not. The sinus wave stimuli in this previous experiment matched the middle high of this study (G5 – 783.99 Hz) but no variation in pitch was given and the focus layed on all three dimensions, not only on tones. Between the last stimuli of the previous study and the first trials of this study a break of about 15 minutes was given. As the previous study focused on fast reaction to any stimuli and contained no pitch variation or rating of stimuli an influence on this follow up study might be controllable.

When starting the here described study, participants first got an instruction of what they will have to do followed by an introduction to the scales, where a written description of each scale was given to them alike the description given here a subchapter above. Finally, stimuli were played once in random order to familiarize the participants with the set of stimuli. The experiment started after participants had the chance to pose remaining questions. During the study the investigator monitored the hand position of the subjects and corrected them if they cover accidently the speakers of the presentation phone. Conducting the experiment was accounted for 15 minutes.

Experimental design and variables

This study followed a repeated measurement plan, where each participant rated each stimulus once on all scales. Independent variable was the stimulus – 8 different sinus tone sequences. The age of subjects was a between subject factor (median split into 2 groups). Main dependent variables were the ratings of each stimulus on the 5 scales. As participants had the choice to repeat the stimulus as often as they want during their rating, the number of repetitions is interpreted as a factor of uncertainty:
If one stimulus is repeated more often than others participants had more difficulties to rate it. The rating time was measured from the beginning of stimulus presentation until the rating was finished and is interpreted as a factor of difficulty. All data was analysed in SPSS 22 using Anovas with repeated measures.

**Results**

The result section initially contemplates the number of stimulus repetitions and the time for rating the participants needed. Subsequently, on each scale differences between stimuli and age-group were reported. This is followed by a correlation matrix to see if the scales intercorrelate. Finally an overview is given for the ratings of each stimulus on the significant scales.

**Repetition count and rating time**

A repeated measurement Anova with stimulus as independent variable, age-group as between subject factor and repetition count as dependent variable showed no effect for stimulus \[F(4.69, 112.62) = 0.773, p = .564, \eta_p^2 = .031\], age-group \[F(1, 24) = 0.848, p = .402, \eta_p^2 = .029\] or the interdependency of those factors \[F(4.69, 112.62) = 0.641, p = .659, \eta_p^2 = .026\]. Mauchly’s test indicated that the assumption of sphericity had been violated for the stimuli \[X^2(27) 56.03, p < .001\], therefore Greenhouse-Geisser corrected tests are reported (\(E = .67\)). Overall each stimulus was repeated less than one time (0.71) with a maximum of 6 repetitions for one stimulus.

As a factor of difficulty the rating time was measured from the beginning of stimulus presentation until the rating was finished. A repeated measurement Anova with stimulus as independent variable, age-group as between subject factor and rating time as dependent variable showed no effect of stimulus \[F(3.99, 38.38) = 0.811, p = .522, \eta_p^2 = .037\] but a significant effect of age-group \[F(1, 21) = 8.204, p = .009, \eta_p^2 = .281\]: older participants took about 36 seconds to rate while younger participants required about 22 seconds. An interdependency between age and stimulus was not found \[F(3.99, 38.38) = 0.972, p = .972, \eta_p^2 = .006\]. Again Mauchly's test indicated that the assumption of sphericity had been violated \[X^2(27) 71.65, p < .001\] therefore Greenhouse-Geisser corrected tests are reported (\(E = .67\)).

**Scales**

The ratings on scale 1 alarm – note showed a significant effect for stimulus \[F(7, 168) = 2.313, p = .028, \eta_p^2 = .088\] but no effect of age-group \[F(1, 24) = 1.086, p = .308, \eta_p^2 = .043\] or interdependency between stimulus and age-group \[F(7, 168) = 0.974, p = .452, \eta_p^2 = .039\]. Here Mauchly’s test indicated no violation of sphericity. The scale alarm – note is shown in figure 3.

Scale 2 question – confirmation showed no significant effect: neither for stimulus \[F(4.32, 103.62) = 1.602, p = .175, \eta_p^2 = .063\], nor for age-group \[F(1, 24) = 1.751, p = .198, \eta_p^2 = .068\] or interdependency of both factors \[F(4.32, 103.62) = 1.515, p = .199, \eta_p^2 = .059\]. Here Mauchly's test indicated that the assumption of sphericity
had been violated \( \chi^2(27) = 49.80, p = .005 \) therefore Greenhouse-Geisser corrected tests are reported \( (E = .62) \).

Scale 3 pleasant – unpleasant showed a significant effect for stimulus \( F(7, 168) = 5.710, p < .001, \eta^2 = .192 \), but not for age-group \( F(1, 24) = 0.634, p = .434, \eta^2 = .026 \) or interdependency \( F(7, 168) = 0.720, p = .655, \eta^2 = .029 \) (see figure 4).

Scale 4 important – unimportant showed again a significant effect for stimulus \( F(4.65, 111.61) = 3.529, p = .006, \eta^2 = .128 \), but no effect for age-group \( F(1, 24) = 0.094, p = .761, \eta^2 = .004 \) or interdependency of stimulus and age \( F(4.65, 111.61) = 1.684, p = .149, \eta^2 = .066 \) (see figure 5). On this scale Mauchly’s test indicated that the assumption of sphericity had been violated, \( \chi^2(27) = 50.70, p = .004 \), therefore Greenhouse-Geisser corrected tests are reported \( (E = .66) \).

![Figure 3. Rating of 8 stimuli at scale 1 alarm – note. Error bars reflect the 95% confidence interval.](image)
Figure 4. Rating of 8 stimuli at scale 3 pleasant – unpleasant. Error bars reflect the 95% confidence interval.

Figure 5. Rating of 8 stimuli at scale 4 important - unimportant. Error bars reflect the 95% confidence interval.
The scale 5 distinct – inconclusive showed no significant effect at all: Not for stimulus \[ F(7, 168) = 0.820, \quad p = .572, \quad \eta^2 = .033 \], nor for age-group \[ F(1, 24) = 0.008, \quad p = .930, \quad \eta^2 = .000 \] or interdependency of both factors \[ F(7, 168) = 1.095, \quad p = .369, \quad \eta^2 = .044 \].

To sum it up: Two scales failed to show significant differences between the stimuli: scale 2 (question - confirmation) and 5 (distinct - inconclusive). In that scales no pattern could be found that indicate a common interpretation. But three scales (1 alarm – note; 3 pleasant – unpleasant; 4 important – unimportant) showed significant effect of stimulus, which means that participants have a common way to rate the stimuli in that dimensions.

**Correlations between scales**

Table 1 shows the correlations between the scales. These correlations are based on the ratings of all participants across all stimuli (N = 26 participants * 8 stimuli = 208). As table 1 shows the scales do highly intercorrelate, meaning the ratings are not independent from each other. Focusing on the scales that showed a significant effect (1,3,4) it can be said that stimuli that are rated rather as note than alarm (high score on scale 1) are associated with being pleasant (low score on scale 3, with negative correlation) and less important (high score on scale 4). On the other hand alarms are more unpleasant and important. Although these associations make sense in real life it means statistically that the alpha risk of the Anovas for each scale is enlarged (because asking several times for the same phenomenon) and has to be corrected from 5% to 1% (divided by the number of scales or times asking for the same phenomenon). In that case scale 1 would be no more significant, while scale 3 and 4 still hold their significance.

**Table 1. Correlations among scales (**p < .001)**

<table>
<thead>
<tr>
<th></th>
<th>Scale 1</th>
<th>Scale 2</th>
<th>Scale 3</th>
<th>Scale 4</th>
<th>Scale 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale 1</td>
<td>-.184**</td>
<td>-.220**</td>
<td>.476**</td>
<td>.026</td>
<td></td>
</tr>
<tr>
<td>Scale 2</td>
<td>-.220**</td>
<td>-.073</td>
<td>.189**</td>
<td>-.075</td>
<td>.107</td>
</tr>
<tr>
<td>Scale 3</td>
<td>-.200**</td>
<td>-.073</td>
<td>-.099</td>
<td>-.193**</td>
<td></td>
</tr>
<tr>
<td>Scale 4</td>
<td>.476**</td>
<td>.189**</td>
<td>-.099</td>
<td>-.193**</td>
<td></td>
</tr>
<tr>
<td>Scale 5</td>
<td>.026</td>
<td>-.075</td>
<td>.107</td>
<td>.193**</td>
<td></td>
</tr>
</tbody>
</table>

**Interpretation of stimuli**

Figure 6 shows the mean ratings on the two remaining significant scales for each stimulus. Age-group as factor was dropped as no age effects showed up during the analysis of each scale. The stimuli “10203”, “30201” and “20202” are interpreted as being most pleasant, while “13013” and “31031” are being perceived as most unpleasant. That means that pauses between the tones – independent from pitch variation – are more pleasant to the subjects, while fast pitch jumping without any pauses in between is perceived as unpleasant. The most important stimuli were “20202” – the one which was also identified mostly as alarm – and “31031” – the one which was the most unpleasant. The most unimportant stimuli were “10203”,...
“11033”, “30201” and “33011” – all of the stimuli with a slow pitch variation separated by a pause.

Figure 6. Rating of 8 stimuli at the two significant scales: Scale 3 (pleasant – unpleasant) and 4 (important – unimportant). Error bars reflect the 95% confidence interval.

Discussion

This study followed an explorative setting to investigate if simple and abstract sinus wave sequences do carry an inherent and context free meaning if applied in human computer interaction. The results have to be interpreted as relative within the set of given stimuli. This set was very minimalistic regarding number of stimuli and also regarding variation within those stimuli to identify certain factors that influence the perception of simple sinus wave sequences.

Only two of five scales showed significant effects: Scale 3 (pleasant – unpleasant) and 4 (important – unimportant) and the effect size is rather small. However, even with that given restricted set of only 8 stimuli and 26 participants it is now proved that the sinus wave stimuli are interpreted in the same way from participants in these dimensions. This effect is independent from age or cohort as young and elderly participants showed the same reaction. This does not only stand for a “design for all” approach but also once again it can be seen that the signals are interpreted consistently the same way by both groups. So it can be nomore denied that even abstract sinus wave signals contain context free characteristics that will influence the interpretation and reaction to these signal. But it is too early to build common rules
for that interpretation based on pitch and sequence of the stimulus. Therefore more studies have to be done with a wider range of stimuli.

The present finding that pitch variation is interpreted to be more pleasant than stimuli without pitch variation stands in line with the findings of Scherer and Oshinsky (1977) and Smith and Cuddy (1986). Additional we can say that pitch variation separated by pauses is interpreted as more unimportant signal compared to signals without pitch variation or without pauses. However, scales are highly intercorrelated, which means they are not independent and seem to charge on a concept we do not know yet and which is somehow a mix of unpleasant importance on one side of the scale and pleasant unimportance on the other side of the scale. This has to be further evaluated in future research, where scales have to be validated and ideally be renamed in one global concept or other independent scales that do not correlate have to be found. But although the behind concept is not fully clear yet, this work has proven that there is some global, context free interpretation of abstract sinus wave signals, that will influence the perception of and reaction to these sounds. And as long as machines use those sounds (which are not ideal at all, see theoretical background) understanding this concept is crucial for developing auditory displays.

Acknowledgement

This publication is part of the research project “TECH4AGE”, which is funded by the German Federal Ministry of Education and Research (BMBF, Grant No. 16SV7111) supervised by the VDI/VDE Innovation + Technik GmbH.

References


