Dealing with System Response Times in Interactive Speech Applications

Peter Fröhlich
ftw. Telecommunications Research Center Vienna
Donau-City-Str. 1, A-1220 Vienna, Austria
froehlich@ftw.at
+43/1/5052830-85

ABSTRACT
In this user study, we address several open issues in the design of waiting cues for system response time (SRT) in interactive telephony speech applications. User observations and structured preference tests indicate that silent waiting times should not be longer than 4 – 8 seconds. Already at short durations, music combined with speech was favored to silence. A preference test regarding several non-speech waiting cues proposed in literature suggests that music is preferred to more simple synthetic sounds and to natural sounds. The continuous indication of the remaining waiting time by speech was rated as much more pleasant and appropriate than a non-speech audio progress meter. Commercial announcements and navigational advice during waiting times were not accepted by the subjects. Empirically based guidelines for a maximum waiting duration in voice services is given. Implications for the design of auditory waiting cues for SRT are discussed.

Author Keywords
Speech I/O; Auditory I/O and Sound in the UI; Mobile Communication; System Response Time; User Preference.

ACM Classification Keywords
H5.2. Information Interfaces and Presentation: User Interfaces; H4.3. Information System Applications: Communication Applications.

INTRODUCTION

Speech dialogue systems have the potential to improve access to ever-smaller mobile devices [10]. Speech-based telephony booking services, banking applications or information services are emerging, and some of these have already achieved a considerable quality standard [9].

One of the challenges user interface designers of these applications are confronted with is system response time (SRT), often caused by complex and resource-demanding tasks like distributed speech recognition and synthesis, resource fetching and wireless data transmission. Apart from keeping SRT as short as possible, the basic user requirements in this regard are to diminish frustration caused by the task flow interruption and to reassure the user that the call is still connected.

For GUIs, [6] proposed that a waiting cue, i.e. an indicator telling the user that the system is processing, is needed for waiting times longer than 2 seconds. This (weakly empirically substantiated) rule of thumb cannot automatically be transferred to interactive speech applications, because time-related expectations in spoken interaction differ from GUI-based interaction and because telephony speech applications require SRT indication in the auditory modality. In this context the following questions arise:

What duration of silence do users tolerate while interacting with a speech application? From what duration on do users need a waiting cue?

[6] also proposes that, for longer waiting durations than 10 seconds, more specific information about system processing should be provided, e.g. by a progress bar [5]. In telephone-based interactive speech applications, information about remaining waiting time could either be provided by speech or by non-speech sounds like the audio progress meter proposed by [3]- a musical sequence with a tone changing its pitch towards a fixed reference frequency to convey the "distance" from current position to goal. Resulting questions are:

Does information about the remaining waiting time in an interactive speech application increase user satisfaction while waiting? Should this information be given by speech or by an audio progress meter?

Speech user interface experts (for instance [1] and [4]) have proposed several types of non-speech auditory waiting cues, namely natural sounds (e.g. water being poured into a cup or a clicking clock), synthetic ticking sounds, and musical sequences.

However, there have only been a few relevant empirical evaluations in the context of telephony speech applications. [7] and [8] have found evidence that clicks played at 1 – 2 second intervals decrease negative affect while waiting, but that jazz music and silence were preferred to clicks.

CHI 2005, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

We were interested in further empirical evidence about user preference of non-speech waiting cues, especially for short durations (5 seconds), in which a speech cue (e.g. "Please wait") might not yet be necessary.

Which type of waiting cue (natural sounds, synthetic ticks, or music) is most pleasant and most appropriate?

A practical consideration for service providers is whether waiting times could be used to make commercial announcements (e.g. about further mobile services) or to provide advice on how to use the voice service more efficiently. Several (not formally tested) expert-based guidelines [4,2] argue against this, because users might think that they are prevented from making progress with the application.

Can waiting time be used to make announcements of further service offers or to convey extra information on how to use the service?

An important design question is the maximum duration a user should be held waiting in case of an exceptionally long delay during the download of a required resource document. If it is known how long users would be prepared to wait for typical services offered in an interactive speech application, the downloading process could be aborted at the right time before risking a potential "hang-up".

How long are users prepared to wait for information in a voice application?

METHOD

General Setup

32 paid German native-speakers participated in individual test sessions. Gender, age (below and over 30 years) and professional status were balanced in order to represent the broad target population of mobile voice services. Our test prototype was a specifically adapted copy of Austria's leading voice portal, the "A1 Voice Service". Using speech commands, users could access and listen to several spoken information services (e.g. news, traffic, cinema or weather announcements) and organiser functions (calendar and Email reader) over a cellular phone.

The general procedure for each test session consisted of an introduction phase, an observation phase, and a preference testing phase. In the introduction phase, the subjects got acquainted with the system's service features and its usage (e.g. navigation by means of speech commands and barge-in). In the observation phase, the test persons performed specific tasks, navigating through the voice portal, e.g.: "Please find out the weather for the region of Vienna!" The system prototype was adapted to observe the spontaneous reactions of the test persons to various forms of waiting phases and SRT feedback. In the preference testing phase, the subjects were explicitly asked to rate specific alternatives with regard to subjective satisfaction issues (e.g. pleasantness, and appropriateness).

Several sub-tests were designed to provide answers to the research questions mentioned above. In all of these sub-tests, the presentation order of durations and alternatives was varied between the subjects to prevent learning and preference effects. Not all subjects performed all sub-tests, since the overall duration would have exceeded the 60 min individual test session duration assumed to be reasonable. The number of subjects included in each sub-test is indicated in the respective sub-test description. In the following, the methodology of each sub-test is described.

Durations for Silent SRT and Waiting Cues

In the observation phase, subjects performed several specified tasks and were confronted with intentionally implemented silent waiting times after uttering a speech command (4, 8, 12, and 16 seconds). User actions, errors (e.g. re-issuing a speech command) and comments were protocolled by the operator.

In the preference testing phase, 27 subjects navigated to specified parts of the system and were explicitly asked to provide subjective ratings after each experienced silent waiting time (2, 4, 6, 8, 12, and 16 seconds), according to the following 7-point rating scales: "pleasantness of waiting time" and "necessity of a waiting cue". We separated the sample into 2 groups who did (N=13) and did not (N=14) receive a spoken input confirmation (e.g. "OK") before being exposed to the silent waiting time.

The group receiving an input confirmation also performed another set of tasks, in which they encountered waiting times filled with music and a speech announcement ("The requested information is being processed. Please hold the line"; durations 2, 4, 6, 8, 12, and 16 seconds). For each duration, the subjects rated the "pleasantness of waiting time" and the "necessity of the speech announcement".

Indication of Remaining Time

In the preference testing phase, N=11 subjects navigated to specified system parts and encountered waiting durations (16 and 32 seconds), filled with the "audio progress meter" described above. Another subject group (N=13) was exposed to 16 and 32 second waiting durations consisting of a voice announcing the remaining waiting time and background music.

Comparison of Waiting Sounds

In the preference testing phase, N=11 subjects were asked to navigate to certain parts within the system, each being associated with a 5 sec waiting cue before the actual content was played. The following sounds were rated with regard to pleasantness of waiting time and appropriateness of the waiting cue:

- 2 natural sounds: a ticking clock and water being poured into a cup
- A sequence of synthetic ticks (90 BPM), which has frequently been proposed by speech interface experts

- 3 music sequences: an instrumental groove (bass, drums and keyboards, 90 BPM), a drum-only loop, and a sequence with minimal synthetic percussive sounds

Commercial Announcements and Navigational Advice

In the observation phase, while waiting for the information relevant during 6 specified usage tasks, N=13 subjects were exposed to a) 2 different commercial announcements, b) 2 advices about further available speech commands they could use to improve usage efficiency, and c) 2 normal speech waiting cues ("Please wait"). Usage behaviour (especially retrying the command during the waiting phase and gestures indicating impatience) and comments were logged, and the subjects were interviewed at the end of the test.

Patience Threshold

In the preference testing phase, N=15 subjects browsed through information services (sports and politics news). Furthermore, they were introduced into a usage scenario to find out specific information in their fictive email reader inbox about meeting dates, appointments, and contact data relevant for their weekly schedule. While accessing the news and email items, the subjects were confronted with 1 min waiting sequences. To get an overall estimation of tolerable waiting times for typical services offered in interactive speech applications, they were asked to access two mail and two news items and to give a sign at the moment until which they would be willing to wait.

RESULTS

Durations for Silent SRT and Waiting Cues

During the observation phase, the subjects only started to get impatient and to make errors (for example retrying a speech command) at the relatively long silent durations of 12 seconds (see Table 1). Interestingly, users made less errors in the 16 seconds silent duration. Although the order of presentation was varied between subjects, this result might be due to generally strong learning effects towards silent waiting times.

Durations (sec)	4	8	12	16
Errors (Re-utterance of the command)	0	1	3	0
Impatient behaviors	2	2	5	6

Table 1. Number of coded errors and impatient behaviors for different silent system response time durations during the observation phase (N=12)

As expected, silent waiting times after a confirmation of user input (e.g. “OK”) were perceived as significantly more pleasant than silent waiting times without input confirmation (Mann-Whitney-U-Test for independent samples, N = 13 and 14; mean diff = -1,16, Z = -3,61, $p < .01$) and led to a significantly lower desire for waiting cues (mean diff = -,81, Z = -2,288, $p < .05$). Since it seems arguable that input confirmation should be given before waiting times, we only took the data of the respective user group into account for further analysis.

Figure 1 shows the perceived pleasantness of silence compared with waiting cues in the tested SRT durations. Until a 4 second duration, both silence and waiting cues were rated as very pleasant. For durations longer than 4 seconds, silence was not clearly perceived as pleasant any more (i.e. the 95% confidence interval does not any more fully cover rating scale values for "pleasant") and received significantly lower pleasantness scores than the waiting cue (Wilcoxon-test for paired samples, $p < .05$).

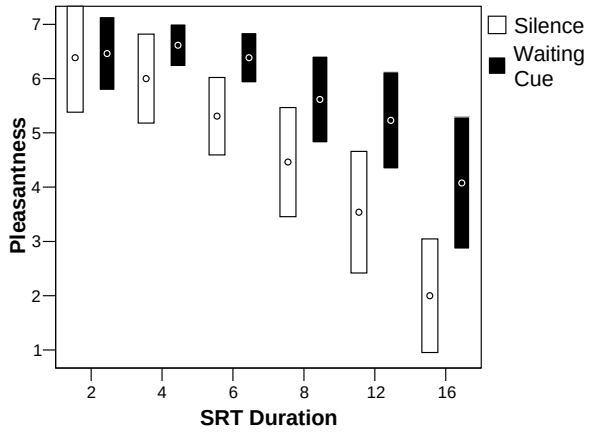

Figure 1. Subjective mean ratings and 95% confidence intervals for "pleasantness of waiting time" (N = 13; 1 = very unpleasant, 7 = very pleasant)

Similarly, for silent durations longer than 4 seconds, waiting cues were not perceived as clearly unnecessary any more (i.e. the mean value is already at the neutral point of the rating scale and the 95% confidence interval does not fully cover the rating scale values for "necessary").

Indication of Remaining Time

The indication of the remaining time received very positive pleasantness and appropriateness ratings when given by speech, but negative ratings when given by the non-speech audio progress meter (see bottom bars of Figure 2). The difference is significant, Mann-Whitney-U-Test for independent samples, N=8 and N=13, Z = -3,94 and -3,85, both $p < .001$. Thus, it seems that the good evaluation results for a non-speech audio progress meter reported by [3] are not applicable to telephony speech applications.

Comparison of Waiting Sounds

Comparing the subjective preferences for musical, natural and synthetic waiting sounds, the "appropriateness of the waiting sound" ratings (see Figure 2) were more informative than those for the "pleasantness of waiting time", which had the same tendency, but were almost all on the positive side given the relatively short waiting time interest.

All musical sequences, i.e. the instrumental groove, the drum-only loop and even the sometimes irritating minimal percussive loop, received significantly higher appropriateness scores than the natural sounds (clicking clock and pouring water) and the synthetic tick sound (Wilcoxon-tests for paired samples, N=11, $p < .05$).

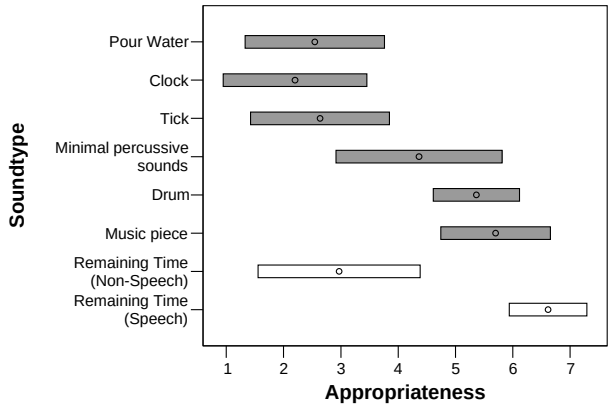

Figure 2. Subjective mean ratings and 95% confidence intervals for non-speech waiting cues and remaining time indicators. 1 = very inappropriate, 7 = very appropriate.

Commercial Announcements and Navigational Advice

During the 2 waiting times with commercial announcements the subjects encountered in the observation phase, impatient behaviour in 8 of the 26 cases (30%) and only one retry of a command was observed. In the interview after the observation phase, several test persons said the commercial information was confusing and irrelevant to their usage situation.

The suspicion of many subjects that the usage progress was retarded by the commercial provides empirical substantiation of expert-based statements made in voice user interface guidebooks mentioned above.

During the 2 navigational advices, there were slightly less (6) impatient behaviours (23%). However, in 8 cases (30%) users were irritated thinking that the system had not understood and retried their command.

Patience Threshold

The mean duration subjects were willing to wait was 45 seconds for news items (95% confidence interval between 39 and 52 seconds), and 35 seconds for information service items (95% confidence interval between 29 and 41 seconds).

Considering the lower edge of the confidence interval, one could argue that waiting times longer than 30 seconds risk a potential hang-up and should be prevented.

CONCLUSION

Design implications indicated by this empirical user study are that in telephony speech applications, silence should be replaced by waiting cues at durations from 4 seconds (due to user preference) and definitely at durations longer than 8 seconds (due to observed user behaviour and impatience). If technically feasible, the remaining time should be indicated, preferably by speech, rather than by an audio progress meter. Commercial announcements and navigational advice during waiting times should be avoided. Users should be kept waiting for an absolute maximum of 30 seconds.

Furthermore, our empirical results suggest that waiting cues containing musical sequences rather than natural or synthetic tick sounds should be used for waiting cues. Due to the high preference for music as waiting cues, we are currently investigating the relevant characteristics for user satisfaction in another user study.

ACKNOWLEDGEMENTS

This work was funded by Kapsch Carrier-Com AG and by Mobilkom Austria AG, together with the Austrian competence centre program Kplus.

REFERENCES

1. Balentine, B, and Morgan, D. *How to Build a Speech Recognition Applicatio*n: *A Style Guide for Telephony Dialogues.* San Ramon, CA, Enterprise Integration Group. 2004.
2. Cohen, M.H., Giangola, J.P., and Balogh, J. Voice User Interface Design. *Voice User Interface Design*, Boston: Addison-Wesley, 2004.
3. Crease, M.G. and Brewster, S.A. Making Progress With Sounds - The Design and Evaluation Of An Audio Progress Bar. *Proc. of ICAD'98*, British Computer Society (1998).
4. Larson, J. *The voice XML Guide, Lesson 12: Resource Management*. http://www.larson-tech.com/U12/12.html
5. Myers, B. A. The importance of percent-done progress indicators for computer-human interfaces. P*roc. ACM CHI'85 Conf.*, ACM Press (1985).11-17.
6. Nielsen, J. *Usability Engineering*. Boston, MA, Academic Press, 1993, 135-137.
7. Polkosky, M.D. User preference for system processing tones (Tech. Rep. No. 29.3436). Raleigh, NC: IBM, 2001.
8. Polkosky, M.D., Lewis, J.R. Effect of Auditory Waiting Cues on Time Estimation in Speech Recognition Telephony Applications. *International Journal of Human-Computer Interaction*, 14, 3-4 (2002), 275-278.
9. VoiceAward 2004, http://www.voiceaward.de/
10. Von Niman, B. Mobile Communication: Simplifying the Complexity, *Business Briefings: Wireless Technology* (2004). www.bbriefings.com