Survival Analysis: Objective assessment of Wait Time in HCI

Siddhartha Asthana
Indraprastha Institute of Information Technology, Delhi
siddharthaa@iiitd.ac.in

Pushpendra Singh
Indraprastha Institute of Information Technology, Delhi
psingh@iiitd.ac.in

Parul Gupta
Indraprastha Institute of Information Technology, Delhi
parul1370@iiitd.ac.in

ABSTRACT

Waiting for the completion of a system process is an everyday experience. While waiting, system provides feedback to the user about ongoing process through temporal metaphors (Progress bar, Busy icons, etc.). One of the key performance requirement for temporal metaphors is to retain the user till the process completes. Researchers have evaluated these metaphors through subjective means, and objective assessment has not been well explored. In this paper, we present *survival analysis* as objective assessment method to evaluate temporal metaphors. Through a field experiment, we demonstrate the application of survival analysis and empirically establish that auditory progress bar (temporal metaphor for audio interfaces) works for callers of a distress helpline. To the best of our knowledge, it is the first study on distress callers. The paper further discusses the applicability of survival analysis for evaluating temporal metaphors and wait time experiments for other applications, tasks, and settings.

Author Keywords

Temporal metaphors, objective assessment, wait-time, time-perception

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces - Evaluation/methodology

INTRODUCTION

Waiting Time in HCI is an important aspect of system design to enhance User eXperience (UX) when the interaction between the system and its user is temporarily interrupted (e.g., loading of web page, file download, or waiting for the arrival of live agents at call centers) [27]. Literature suggests that providing feedback to a user on the system state, especially during these waiting periods, increases system usability [37]. Depending on the type of system, several temporal metaphors (visual progress bar [22], auditory progress bar [26], non-speech waiting cues [21], etc.) have been designed to provide such feedback to the user. The objective of introducing a temporal metaphor in the system is to mitigate the adverse effect of wait time so that the system can successfully hold the user untill the waiting process completes. For a researcher, it is imperative to measure the effectiveness of such metaphors in holding users of the system. However, there are no standard guidelines to measure the effectiveness of a proposed metaphor. Traditionally, researchers have studied the effectiveness of these metaphors through several subjective measures including user preference [26, 14], time perception [22, 21], satisfaction [23], acceptability [14], and appropriateness [17]. At the same time, the use of objective measures has not been explored much to evaluate the effectiveness of such systems.

In this paper, we discuss the implication of waiting time on HCI with a focus on the evaluation of temporal metaphors using objective measures. We discuss the observation that existing subjective measures are susceptible to several evaluation conditions (like cognitive load, the paradigm used for estimation [12], etc.) that can bring variations in the outcomes of the experiments. Further, there are several important objective outcomes that cannot be inferred from subjective measures like *"for what time duration proposed temporal metaphor can hold users, on average"*. Thus, we emphasize the need for objective measures to generate objective outcomes in the study. Towards this, we propose the use of survival analysis of actual wait time as an objective metric to evaluate temporal metaphors for holding the user. Survival analysis (or Life Testing) is an established method to deal with time-based data in the areas of medical science and reliability engineering. In the current context, for each temporal metaphor survival analysis generates the probability of holding a user for a given time and inferential and descriptive statistics to compare two or more metaphors.

We demonstrate the use of survival analysis through the assessment of an auditory progress bar (*a temporal metaphor for audio interfaces*) in a real-world system. Designing an auditory progress bar (APB) is a challenging task for researchers as system interfaces do not support any visual feedback, and it requires sound to carry the entire information load to convey the temporal information. APBs have been extensively studied in the context of *telephone hold environment*, where callers have to wait for the arrival of a live agent [26]. With the increasing use of call centers/helplines in daily life, the use of APB is becoming critical in engaging users of such systems so that they continue with the call and do not hang up. Lab studies [26] and field experiments, at bank call centers [16], have shown that APB can encour-

CHI 2015, April 18 - 23, 2015, Seoul, Republic of Korea

ACM 978-1-4503-3145-6/15/04...$15.00.
http://dx.doi.org/10.1145/2702123.2702428

age callers to wait longer. To the best of our knowledge, ours is the first study which investigates the effectiveness of APB for distress-helpline callers who are more impatient than normal callers due to stress. Through objective measures, the outcome of the field experiment empirically establishes that a speech based auditory progress bar can hold helpline callers for longer.

IMPLICATIONS OF WAITING TIME ON HCI

Despite technological improvements, the waiting time in interactive systems is unavoidable. Strong implications of waiting time on Human Computer Interaction has made researchers investigate acceptable waiting time, the significance of giving feedback on the waiting time to the user, and the design of temporal metaphors to provide better feedback.

Acceptable waiting time: The HCI community on acceptable waiting time has almost agreed on a 10 seconds threshold, beyond which users do not effectively focus on the task anymore [32]. In a study conducted on web users, Bouch et al., in agreement with Nielsen [32], demonstrated that 10 seconds is unsatisfactory for users [13]. The 10-second threshold holds for the systems with no visual representation (i.e. audio only) as well; it [17] showed that for waits more than eight seconds, users should be given speech-based delay information.

Feedback on waiting: It is now widely accepted that giving feedback on waiting time is considered satisfactory, and it increases the usability of the system [32] and the time the user is willing to wait [31, 13]. However, the temptation to give users a detailed feedback can be counterproductive. Giving a detailed feedback focuses the user's attention on the passing time, which makes the wait seem longer [14].

Temporal Metaphors: Feedback on waiting can be conveyed through several metaphors like icons [⧖], text messages [*Loading, please wait..*], audio messages [🔊 *Your Call is important to us, please wait*], progress bars [▭], etc. Among the various metaphors, progress bars [30] are dominant in current user interfaces. Research shows that, among the different types of feedback given to users, progress bars performed best in terms of acceptability for user attention and user preference [14, 21, 23]. Analogous to a visual progress bar, for interactive speech applications that lack visual display, researchers have designed auditory progress bars that include speech-based cues [29], non-speech-based cues [34, 26], etc.

EVALUATION OF WAIT TIME: SOURCE OF VARIATION

An experiment on the evaluation of waiting time involves a users' reporting about the time perceived by them (also known as subjective time). Subjective time is a complex cognitive process, and user assessment depends on several factors. Lallemand et al. have discussed the contribution of cognitive psychology to better understand the subjective time [27]. Here, we discuss how these different cognitive processes and assumptions can bring variations in results and their interpretations for a waiting time experiment.

Mode of Experiment

There are two paradigms to conduct a waiting time experiment for subjective assessment, viz., prospective and retrospective paradigms. In a prospective paradigm, users are informed at the beginning of the experiment that they have to estimate the duration of a given time interval. In the retrospective approach, users are not aware of the time estimation task until the end of the experiment. Studies have shown that retrospective and prospective paradigms have opposite effects on the duration judgment [2, 10]. Block et al. have tried to investigate the reason for performance differences due to different paradigms used for subjective assessment and concluded that "in prospective paradigm, a greater cognitive load on user decreases the subjective-to-objective time duration ratio whereas in retrospective paradigm a greater cognitive load increases the subjective-to-objective time duration ratio" [12]. This performance difference due to the choice of paradigm used for conducting an experiment and the lack of standard methods to translate results from one paradigm to another makes it difficult to compare studies conducted in the past.

Cognitive Workload on the User

The subjective assessment of the time depends a lot on the attention given by a user on the passage of time. Thus, time is perceived as passing slowly when attention is focused primarily on time [11]. The amount of attention on temporal information is less if attention on processing non-temporal information is more, due to a high cognitive demand of a specific task [9, 40]. Thus, the subjective time judgment depends a lot on cognitive workload on the user. Further, users differ from each other in terms of cognitive capabilities, which means that for the same task, different users may perceive different cognitive loads. Thus, for the same experiment, different users report different perceived (subjective) times. Thus comparing results for performance difference of temporal metaphors based on different studies in literature is difficult, as it is not certain whether the difference in results is achieved due to a different system or different cognitive capabilities of their respective user set. One way to avoid such problems is by having a properly sampled large number of users in each experiment, but that is costly and not always feasible. Another approach may be to report cognitive load on the users during the experiment along with other experimental results. There are different psychometric scales to measure cognitive load on the user, such as the Subjective Mental Effort Questionnaire (SMEQ), also referred to as the Rating Scale for Mental Effort (RSME) [42], and NASA-TLX [1]. However, little is known about the representativeness of these scales in terms of the actual cognitive load on the users and the distribution of observed load on the user again in terms of how much is due to the system and how much is due to the difference in cognitive capabilities of a user.

Retention Delay

Retention delay is the duration of time beyond which information cannot be retained by the short-term memory of the human brain. A waiting time experiment conducted with or without retention delay may produce different results. As the human brain does not remember past events in a consistent and linear manner [3, 13], event recall happens with selectivity and bias, where it is easier to remember the first or last moments of an event than those happening in between. Initial moments are stored in long-term memory, and the last or recent moments are stored in short-term memory. Short-term memory is less stable and can be affected by a delay of

more than 15-30 seconds (or retention delay); it can also be affected by an interfering activity [18]. Results for a retrospective evaluation after retention delay will solely be based on the long-term memory [27]. However, if the evaluation is done without retention delay both long-term and short-term memory affect the subjective assessment.

Psychological Models for Subjective Time

Most of the literature on waiting time is focused around explaining the distortion in subjective time perception (under or overestimation). Several models have been proposed to study the perceived time based on the existence of the internal clock [4, 15, 39]. The first model, proposed by Triesman, consists of a pacemaker (internal clock) and an accumulator connected by a switch [38]. The pacemaker continuously emits a pulse, counted by the accumulator if the switch between pacemaker and accumulator is turned on. If the subject pays attention to the time, switch turns to on-state; otherwise, it is in off-state. The subjective time relies on the number of pulses counted by the accumulator. The second model is presented by Gibbon (see scalar timing theory [19]) and has two different assumptions from the model proposed by Triesman. Gibbons model has three interconnected levels instead of two: clock (pacemaker), memory, and decision and according to the second assumption, the clock does not emit a regular pulse but a distribution under Weber's law. The attentional gate model was proposed by Zakay et al. [41]: it is considered the best to explain subjective time in HCI [27]. This model assumes an additional gate between the pacemaker and switch. The subjective time depends upon the way a person divides his/her attention between information encoded by a temporal information processor and non-temporal information processor. The gate takes temporal information into account and co-ordinates an activation of the switch. Thus, different researchers may make use of different models to explain their results and develop different theories. For a comprehensive literature review of the models, readers are referred to [11, 20, 28].

THE WAY FORWARD

The same experiment, when conducted by different researchers, may produce different results based upon different configurations that they choose for evaluation. We do not argue which one is better, but we emphasize that even when researchers explicitly mention the evaluation methodology they have chosen, it is not easy to compare the results with other experiments conducted with different evaluation methodology. As an implication of this, the literature may seem to have inconsistent findings and conflicting theories. A similar observation was made by Kortum et al. while designing an auditory progress bar (APB) for a telephonic system [26]. They mentioned that ten different studies conducted for an auditory progress bar had mixed performances (over and underestimation of time) and listener preferences. We believe that subjective analysis is a good method for understanding the attitude and preference of a user, but researchers also need to incorporate an objective analysis method for a variety of reasons. First, objective methods are less influenced by individual users, i.e., they can capture the actual behavior of a user with the system rather than user perception about the system. Second, it is less time-consuming to collect and analyze objective data for a large set of users. Third, there may be some system aspects that users cannot assess or report–for instance, *how long a system is able to hold users to wait*–and various probabilistic and statistical insights, like *the probability that a user will wait for y seconds*, cannot be assessed by users. Hence, an objective evaluation of a system needs to be integrated with the traditional subjective usability assessment methods. In this paper, we address this gap by introducing an objective evaluation technique that can generate such insights.

We propose the use of actual wait time rather than perceived wait time as an evaluation metric for objective assessment. Subjective studies were largely focused around the perceived wait time, and the use of actual wait time is limited to calculate the mis-estimation (over or underestimation) in perceived wait time [26]. In subjective studies, low perceived time is desirable which is inferred as the system can hold a user for longer. But by merely inspecting the perceived wait time, it is difficult to guess how long a user will actually wait because perceived wait time is dependent on factors other than actual wait time like the cognitive load and paradigm used (retrospective or prospective) for time estimation [12]. Hence, in our proposed objective assessment we tend to measure the actual wait time of the caller directly instead of estimating it from perceived wait time or other perceptual variables. Based on the requirement, we propose the use of survival analysis (a branch of statistics) as an evaluation technique for objective assessment of the wait time in HCI. It is a statistical tool for time-based analysis and has the capability to deal with censored data. With survival analysis, we can quantify the effect of temporal metaphors, judge its statistical significance, and generate probabilistic estimates of holding a user with the passage of time.

Conceptual Foundation

Here, we describe the important terms and concepts which will showcase the appropriateness and applicability of survival analysis. For a detailed description of survival analysis, readers are referred to [25].

Time Based Events and Censored Data

A time-based event is defined as an event whose time of ocurrence is the interest of study. The case in which the time of an event is known below or above an observed value but the exact time of the event is unknown is called **censored observation** [25]. For instance, in the case of temporal metaphors, the outcome of interest is to find out the tolerable wait time before a user quits or cancels an ongoing process or transaction. Here, the event to observe is the canceling or aborting of an ongoing process by the user. Different users will quit or cancel the process after waiting for a certain time, which may differ among users. The wait time of different users can be used to statistically compute the tolerable wait time for the users with respect to a temporal metaphor. However, for certain users we will not observe any event if they wait till the completion of the process. In this scenario, we know only that the user waits till a certain time but do not know exactly how long she is willing to wait. In statistical analysis, such data points are termed as the **Right censored data** as the tolerable

waiting time for a user is more than the observed time by an unknown value.

There are two challenges in analyzing data pertaining to a time-based event. The first challenge is to deal with censored observations and accommodate them with normal observations in the analysis. The second challenge is that we cannot expect data to be distributed normally as with increasing wait time lesser number of users will be willing to wait.

Survival Function and Probability Estimation

Survival function is the function that relates the probability of an event with the time estimated from time-based data discussed above. Conventionally, a Survival function F(x) is defined as,

$$F(x) = Pr(X > x) \tag{1}$$

where, X is a random variable denoting the time of event occurrence and Pr(X>x) is the probability that the event will occur later than the specified time x. For estimating such survival function, we will be using Kaplan-Meir [25], a non-parametric maximum likelihood estimator, which is used in several fields of science, like medical science for estimating the life of a patient, in reliability engineering to measure the time of part failure, etc. [25]. To estimate the probability for plotting the survival function, the following procedure is followed:

- Arrange the time-based data in ascending order and denote the time value as $x_0, x_1, ..., x_n$.
- At every event, calculate the probability using following recurrence relation:

$$F(x_i) = \frac{r_i - n_i}{r_i} F(x_{i-1}) \tag{2}$$

  where r_i is the total number of events (including censored events) having time value $\geq x_i$, and n_i is the number of events (non-censored) $> x_{i-1}$ but smaller than x_i.

Size of Effect

To quantify the size of effect of temporal metaphors on wait time behavior of users, we can calculate mean and median survival time from survival analysis (see descriptive measures of survival experience [25]).These terms are defined as:

- *Mean survival time:* It is calculated as the area under the survival curve [25]. It is one of the measurements to estimate the central tendency of the data distribution of wait time and represents how long a user waits on the system on an average.
- *Median survival time:* It is calculated as the smallest survival time for which the probability of the survival function is less than or equal to 0.5 [25]. Similar to mean, it is another factor that measures the central tendency of data distribution but is less susceptible to extreme values.

Statistical Significance

Logrank [8], a non-parametric test, is widely used in survival analysis to estimate whether the effect has statistical significance or not. It is a form of the Chi-square test and the hypothesis testing tool to compare two survival distributions. It calculates a test statistic for a null hypothesis (*survival curves for different groups are the same*) based on the following relation:

$$Statistic(Logrank) = \frac{\sum(Expected - Observed)^2}{\sigma^2(Expected - Observed)} \tag{3}$$

where $\sum(Expected - Observed)^2$ is the summation of the square of the difference between the expected number of events and observed number of events for each time point, and $\sigma^2(Expected - Observed)$ is the variance of difference between expected and observed events.

FIELD EXPERIMENT: OBJECTIVE ASSESSMENT OF AUDITORY PROGRESS BAR

In this section, we demonstrate the proposed objective evaluation technique using survivability analysis as applied on the actual wait time to generate statistical and probabilistic insight about the Auditory Progress Bar designed to handle waiting time in a telephonic queue.

Auditory Progress Bar

The Auditory Progress Bar (APB) is analogous to the Visual Progress Bar for systems with no visual representations. APB is a temporal metaphor where sound must carry the entire information load to convey temporal information. The use of the Auditory Progress Bar (APB) can be seen in telephonic queues at a call center in the form of background music, apology messages or estimated delay announcements. APBs can be broadly divided into speech based and non-speech based cues. Researchers have studied and compared various speech and non-speech based APB designs:

Non-speech cues: Peres et al. investigated building an auditory progress bar that has pleasing aesthetic and good temporal qualities [34]. In a lab study, they experimented with different audio tones like sine, cello, and songs of varying pitch and duration and measured user preference on a seven-point Likert scale and system performance based on perceived waiting time and actual waiting time. In a similar experiment, Harrison et al. explored nine different combinations of linear and non-linear tones compared against each other to determine the tone perceived as fast [21]. The study concluded that, among each pair of tones, whichever tone users encounter first is perceived as faster. Fröhlich suggested that non-speech cues must be used to replace a silent wait of short duration (four to eight seconds) [17]. The study showed that for high waiting time, continuous indication by speech cues was more pleasant and appropriate than non-speech cues.

Speech cues: Feigin conducted an experiment that announced an expected waiting time (<1 min, 1 min, 2 min, etc.) at a high level of congestion in the call center of a US bank and recommended the callers to return to IVR for self-service [16]. It was shown that approximately 1.3% of the total callers returned to IVR and the others opted to join the agent queue. The announcement resulted in the lowering of the initial high abandonment rate, maintaining the callers beyond the 70 to 80-second mark, and a new trend of abandonment around 170-180 seconds. Armony et al. have studied the equilibrium associated with the delay announcement [7]. The equilibrium exists in the sense of the cyclic dependency

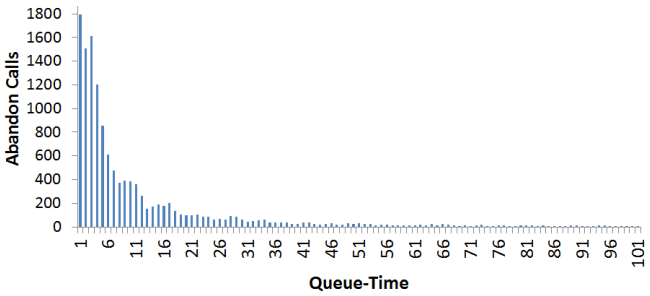


Figure 1. Distribution of Queue-time (in seconds) of abandoned call in pre-experiment data collection

of three factors, viz., the caller's response to the delay announcement, the system performance depending upon the caller's response, and the announcement depending upon system performance. Several other researchers have shown keen interest in developing ideas around delay announcements. Allon et al. have studied the impact of postponing the delay announcement in comparison to providing the delay announcement immediately [5]. They have shown that this postponement can help the call center maintain more credibility among callers and augment the resulting equilibrium in certain settings. In all the studies, the delay announcement from the call center is assumed to be of high credibility on the caller's side. However, theoretical research has also explored the effects on the queuing model if call centers provide vague and unverifiable information about delays to lure customers [6].

Speech vs. non-speech cues: Rafeli et al. explore the psychological aspects of different time fillers and their effect on the cognitive processes of the waiting caller by comparing delay announcements as a filler with two other time fillers - music and apology message [29]. Niida et al. measure the effectiveness of different time fillers on a visual display [33].

In this paper, we investigate the effect of speech-based APB on a distressed caller helpline. Next, we describe the distressed caller-helpline and our experimental design. Then we analyze the APB using survival analysis.

Helpline Background

In this section, we describe the background of the helpline to provide a better understanding of the type of callers and the context of the calls. Then, we report on the initial inspection of the data that we collected before the experiment.

181-Women in Distress Helpline

The state of Delhi is the national capital of India. To address women's concerns, the state of Delhi has established helplines such as 1091 (Delhi Police Women Helpline) and 23379181 (Delhi Commission for Women). Despite these helplines, however, women continue to find it difficult to gain quick access to information regarding police or other law enforcement agencies. To handle these issues, the state government of Delhi decided to start its supplementary helpline, named the *181 - Women in distress* helpline, to provide assistance to women in distress. The call center is managed by women call executives who are trained in handling women's issues and some legal advisors. The calls vary from seeking counseling over petty matters (e.g., verbal arguments within the family) to criminal assaults (dowry death, rape, etc.) as told to us by agents attending to the calls. The helpline provides advice and consultation to women and coordinates with other state agencies (police, legal aid, medical service, etc.) whenever needed for the fast resolution of women's complaints.

Helpline inspection

To understand the waiting behavior of callers, we analysed the helpline prior to experiment design. We analyzed 29,548 calls that the helpline had received in the past three months before the experiment. Out of the total calls, 13,428 were abandoned calls, i.e., calls that were not answered by any agent, and 16,120 were answered calls. We also observed that most of the callers try again, so the actual number of unanswered callers is not as high as it looks prima facie. To understand the system dynamics, in order to design our experiment better, we performed some basic analysis.

We analyzed waiting time (or queue-Time) for the abandoned calls and the answered calls separately. Figure 1 shows the distribution of abandoned calls based on queue time. For abandoned calls, we found that 95% of the calls had a queue time (waiting time) of less than 97 seconds. It shows that 95% callers on such helplines do not even wait 100 seconds, which reflects the impatience of callers due to stress. On the other hand, the waiting time in normal call centers could be as high as 40 to 50 minutes (as reported in a leading newspaper of the UK [1]). This shows that the nature of helpline callers is different from a normal call center. We want to reduce the number of abandoned calls as these calls are important for the purpose of the helpline. Though many callers return, it will still be better to serve them at the first instant. One solution is to increase the number of call executives at the time of rush hour, but due to several reasons, the number of executives cannot be increased. Thus, we would like callers to wait more so that they can talk with a helpline executive and get their problems resolved. Literature suggests that the speech based auditory progress bar announcing the estimated wait time can encourage callers to wait more in call centers of the bank [16]. However, to the best of our knowledge, this is the first study that investigates whether such approaches work for callers of helplines who may be impatient due to stress.

Research Ethics: We took the appropriate permissions from the concerned authority managing the helpline and followed ethical practices in conducting this research. Authors are familiar with US-based IRB approval. Though such approval is not required for conducting research in India, we were very much aware of the sensitivity of the data and implications of mishandling it; all the precautions were taken to ensure that the privacy of the callers remained intact. We did not access case information details, audio recordings related to the call, or any personally identifiable information about the callers. Except for the delay message, we did not change anything in the helpline setup, and our experiment did not affect the normal working of the helpline.

Experiment Design

We designed one APB that gives a periodic delay announcement and compared it with the base line system (BL).

[1] `http://www.mirror.co.uk/news/uk-news/big-six-hold-times-energy-3168050`

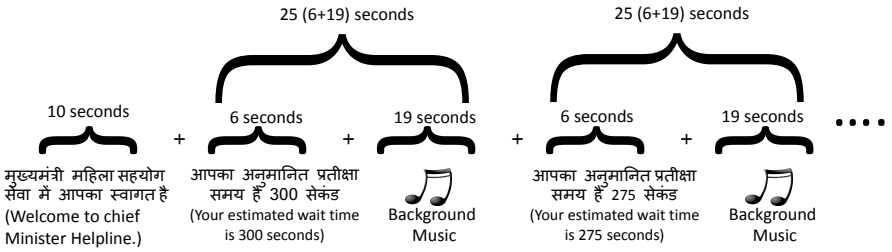


Figure 2. Functioning of APB: First callers are greeted with a 10 seconds welcome message followed by first message about an estimated delay of 300 seconds. Each message about estimated delay takes 6 seconds. Thus, caller with Queue-time greater than 16 seconds receive at least one delay announcement. Every estimated delay message is followed by background music of 19 seconds. The repeated announcement of delay messages by subsequently deleting 25 seconds from the previous estimate continues till estimate reaches zero. After that caller only hears the background music.

1. *Baseline (BL):* This system is the existing system of the 181 helpline, which does not provide any delay announcement to the caller. It has a pre-recorded audio message that repeatedly plays a welcome message to the caller in the Hindi language. The message translates to ***"Welcome to the Chief Minister's Women Helpline!!"***

2. *APB:* This is our proposed system, which announces an expected initial waiting time of 300 seconds to each caller. System keeps updating waiting time after every 25 seconds. The value of each announcement against actual wait of caller is shown in Figure 2. It is a simple system which deducts 25 seconds from the last announced time and plays it in subsequent announcements as the time left for the caller. After the 12 ($300/25 = 12$) announcements, actual wait time becomes zero and hence system stops announcing any delay. The time for duration (300 seconds) and interval (25 seconds) of APB design is based on following considerations.
   - Duration: Literature suggests to estimate the wait time for an announcement based on coverage probability (β) so that actual wait time never exceeds the announced time with probability β [24]. To ensure that the actual wait never crosses the announced wait, we would have to announce wait-time corresponding to $\beta = 1$. However, due to the exponential distribution of wait-time, the value corresponding to $\beta = 1$ will lie at infinity. We did analysis of 29,548 calls that are described in previous subsection (*Helpline inspection*) to estimate the wait time for announcement so that value of β is close to 1. The minimum value which covers sufficiently large number of callers in our system is 300 seconds ($\beta = 0.999$).
   - Interval: Choice of interval is influenced by two aspects covered in [14] and [17]. Literature suggests constant attention towards the passage of time makes the wait time perceived as longer [14]. Hence, the value of intervals between the announcements should be high as every announcement draws the caller's attention towards the passage of time. On the other side, Fröhlitch et al. indicates that the wait time greater than 30 seconds risks a potential hang-up [17]. Nah et al. quoted another study that also indicates a 30 second threshold in another context [31]. Thus, we chose a high value of interval [14] that was less than 30 seconds [17, 31] and perfectly divided the duration (300s).

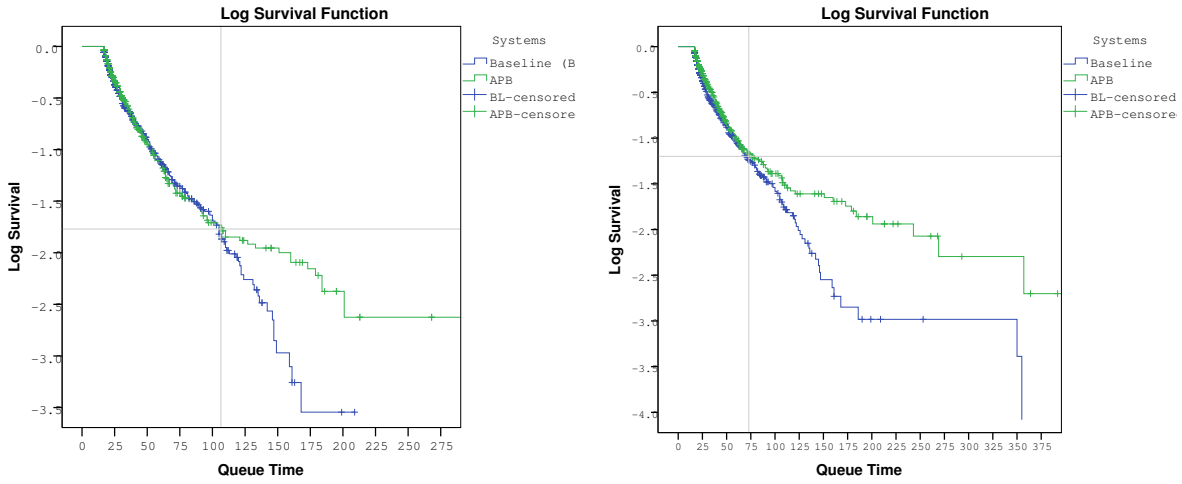
We designed to route every alternate call to the same system. For each new call, the system generates a numerical call ID in an incremental fashion. We took each new call as a different call even if it came from some previous number logged in our database and assigned it to one of the systems alternatively. To preserve the privacy of the caller, the system anonymized the caller-ID (telephone number) to a different numerical ID before storing it in the database. Thus, repeated calls from the same number could be identified but reverse mapping to the phone number was not possible.

Data Collection

For the analysis, we collected data for 1 month (6^{th} August - 5^{th} September 2014). The system logged the following call attributes:

- Caller ID: For each call connected to our system, we logged the anonymized caller id of the telephone number.
- Time stamps: For each call, we logged the time stamps of the call's start and end.
- System allocation: For each call, we logged the system name that was allocated, i.e., BL or APB.
- Announcement Details: We logged the number of announcement messages listened to during each call and the corresponding value announced in each announcement along with the time stamp of each announcement.
- Queue Time: We explicitly logged the time spent by each caller on the system waiting for an agent to answer her calls.
- Call Status: For each call, we logged whether it was answered by an agent or abandoned by the caller.

In one month, we received a total of 8,748 calls. There was a large number of calls that got answered or disconnected with waiting time (or Queue Time) $<= 16$ seconds and hence did not receive any delay announcement (calls with Queue Time $>= 16$ seconds received at least one delay announcement, see Figure 2). As we want to compare the performance of the system when callers got a delay announcement to its performance when callers did not get any announcement, we removed calls with Queue Time $<= 16$ seconds. For consistency in analysis and comparability between the systems, we removed such calls from both the systems and were left with 1,353 calls.



(a) Total Calls (b) First Calls

Figure 3. Survival Curve: A comparison of survivability of callers on Baseline, and APB. Markers on each curve show the corresponding censored (Answered calls) data.

Results and Analysis

Here, we present the analysis of 1,353 calls. Table 1 represents the distribution of calls between the two systems and the number of calls answered and abandoned on each system. These 1,353 calls represent the total calls from 1,082 callers, where some callers called more than once. The subsequent calls from the same caller may have a priming effect on their waiting behavior, based on their experience of the first call. To avoid any priming effect on the calls analysis, we have also presented the analysis for the data where only first call of each caller is considered. The distribution of the 1,082 remaining calls (i.e., considering only first call of each caller) is shown in Table 2. In the rest of the section, we will present the analysis for total calls and first callers separately.

System	Calls		
	Answered	Abandon	Total
Baseline	255	471	726
APB	241	386	627
Total	496	857	1353

Table 1. Total Calls

System	Calls		
	Answered	Abandon	Total
Baseline	208	383	591
APB	207	284	491
Total	415	667	1082

Table 2. First Call

Abandon Rate

A caller who waits longer has more of a chance of her call getting answered than a caller who waits for a shorter duration of time. To observe which system (APB or Baseline) can hold a caller longer, Abandon Rate is defined as the percentage of total calls that were unanswered. Hence, a better system will observe an abandon rate lower than the other system. Based on the analysis of total calls (see Table 1), the abandon rate of APB (61.56% i.e., $386/627$) is less than the Baseline (64.87% i.e. $471/726$) by approximately 3%, but results were not statistically significant ($\chi^2(1, N = 1353) = 1.59$, $p = 0.2$). On analyzing the first call only, we found the abandon rate of APB (57.8% i.e., $284/491$) was much lower than the Baseline (64.8% i.e. $383/591$) ,and the results were statistically significant ($\chi^2(1, N = 1082) = 5.5$, $p = 0.019$). This indicates that APB performed better than Baseline in holding the callers. However, a caller waiting time is just one of the factors, and there are other factors, such as congestion in the line and number of agents which also affect the abandon rate of the system. Thus, we will further analyze the system for more insights.

Survival Analysis

In this section, we demonstrate the use of survival analysis techniques as applied to wait time (Queue-Time) data collected in the experiment.

Probability Estimation

We did a survival analysis (Kaplan-Meier) to analyze the probability of callers waiting X seconds or more in the queue across different systems. In this analysis, the *queue time* is the *survival time* of the caller, *call abandonment* is the *observed event*, and *answered calls* are treated as *right censored data*[2]. We did a survival analysis of total calls and calls that were left after removing repeated calls from the same telephone number.

Figure 3a shows the estimated probability of callers waiting for a particular time (in seconds) on two different systems using total calls. The performance difference can be seen between the two systems for Queue Time > 100 seconds where the probability that callers will wait on APB is more than the Baseline. The same performance difference can be seen much earlier (around Queue Time = 60 seconds) when only first calls are considered (see Figure 3b).

[2]In this experiment, the queue time of answered calls represents right censored data as caller could have waited more than the queue time but unknown by how much.

	Mean[a]				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Baseline	63.318	4.199	55.087	71.548	39.000	2.187	34.714	43.286
APB	76.522	5.910	64.938	88.106	39.000	1.595	35.874	42.126
Overall	69.737	3.623	62.637	76.837	39.000	1.379	36.296	41.704

Table 3. Means and Medians for Survival Time based on total calls

	Mean[a]				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
Baseline	74.100	7.364	59.667	88.533	40.000	2.712	34.684	45.316
APB	94.063	7.876	78.627	109.499	43.000	2.582	37.938	48.062
Overall	90.125	7.783	74.870	105.379	42.000	1.661	38.744	45.256

a. Estimation is limited to the largest survival time if it is censored.

Table 4. Means and Medians for Survival Time based on first calls

Size of Effect

We used the mean[3] and median[4] of probabilistic data generated through survival function to quantify the effect of different systems on caller waiting time. Table 3 shows the mean and median estimates of survival time for each system using total calls. The mean of APB (76.52 seconds) is approximately 13 seconds more than the Baseline (63.13 seconds). This shows that on an average, APB can hold each caller 13 seconds more than the Baseline. We observed that both the systems have the same median of 39 seconds that is evident as the left half of the survivability curve is almost the same (see Figure 3). This can also be interpreted as if the helpline has sufficient agents where the waiting time of callers is below 39 seconds, then we will not observe any performance difference between the two systems.

We did the same analysis for *first calls*, and estimated mean and median are shown in Table 4. We can see a wider gap in the performance. The mean of APB (94.06 seconds) is approximately 20 seconds more than the Baseline (74.1 seconds). Similar to the analysis on total calls, median values are close to each other.

A performance difference of 13 seconds or 20 seconds (considering only first calls) may not impact full performance difference for a normal call center, but it is a considerable improvement for helpline callers where the average waiting time is less than 64 seconds (see mean of BL in Table 3) or 75 seconds (considering first call, see mean of BL in Table 4). It is also evident from the abandon rate, which was reduced to 3-7% (see previous subsection). Further, optimization on the APB design may bring better performance.

Statistical Significance

We did further analysis to check the statistical significance of our observed performance difference through Chi-Square statistics, i.e., Logrank (Mantel-Cox). While testing total calls for statistical significance, we did not observe any difference (χ^2 (1, 1353)=0.945, p=0.33). However, on analyzing the data of first calls, we find the results are statistically significant (χ^2 (1, 1082)=5.058, $\mathbf{p<0.05}$).

We further analyzed the reason for not achieving statistical significance with total calls. We believe one of the reasons could be the repeated calls in the data set. We hypothesize that a caller can estimate the waiting time in subsequent calls based on one's experience of the first call. Thus, irrespective of the system, callers have sufficient information on the waiting time. To check this, we analyzed 271 repeated calls (i.e. first calls subtracted from total calls). We did not find any statistically significant differences (χ^2 (1, 271)=0.003, p>0.05) in the two systems. The repeated calls on APB had a lower mean (83.38 seconds, 95% CI[52.25, 114.51]) and median (40 seconds, 95% CI[34.47, 45.53]) than the first calls.

DISCUSSION

This section reflects on our results and discusses possible interpretations and limitations of it. We discuss how our proposed method can be generalized to other tasks, application types, and settings.

We showed that the user wait time data can be used objectively to evaluate the performance of temporal metaphors for their ability to hold callers. We showed survival analysis, a lifetime estimation technique, can be used for such purposes to generate probabilistic and statistical insights.

The first concept that we discussed is a survival curve. It is a continuous time-based evaluation of temporal metaphors. It denotes the probability a user will wait for a given time. Several insights can be drawn from this while comparing temporal metaphors. For instance, if curves are non-overlapping [ ] then the temporal metaphor with a curve on a higher side along ordinate (Y-axis) represents better performance. In the case of overlapping curves [], let's say intersecting at time **t** on the X-axis can be used to make a thoughtful decision on which temporal metaphor should be used based on the average wait time of task for which that temporal metaphor is employed. If average wait time of task is less than **t**, then the metaphor whose survival curve is at the higher side for wait time less than **t** should be used. Otherwise, if average wait time of task is greater than **t**, then the metaphor whose survival curve is at the higher side for wait time greater than

[3]Mean survival time is estimated as the area under the survival curve

[4]The median survival time is calculated as the smallest survival time for which the survivor function is less than or equal to 0.5

t should be used. An important underlying assumption for statistical evaluation of the survival curve that overlapping curves violate is proportional hazard. If survival curves cross each other, then this is the evidence that hazards are not proportional. For such overlapping curves, the analysis should take into account the interaction effect and separate the analysis before and after the crossing. We refer the reader to [25] for hazard function and assumption of proportional hazards.

Second, we discussed quantifying the size of the effect of temporal metaphors on the wait time behavior of users. If a particular metaphor has performed well then quantifying it for the achieved difference in performance will further enhance the understanding of temporal metaphors. This may be valuable for practitioners and researchers. Such data needs to be reported with both the mean and median with their respective confidence intervals to have a better insight about the central tendency of the data. Medians may be good to report in non-Gaussian or positively skewed data (as in task time data in usability studies [36]), but sometimes due to resistance to extreme values, they do not capture the performance difference adequately. This is evident from our results where the medians are close to each other even when systems have a significant difference in performance statistically. Thus, the mean can be helpful in quantifying the effect size.

We also discussed assessing the statistical significance of the performance difference. We employed the Logrank test for this purpose, which is a form of the Chi-square test and widely used in survival analysis and life estimation techniques [8]. There are other similar tests like the Wilcoxon-test and Tarone-ware test which can be used for different applications based on the need. These tests differ in sensitivity to the survival curve and put different weights to initial events and events that are happening later in the time. For more discussion on the comparison of Logrank and other applicable tests in survival analysis and the effect of right-censoring, please refer [35].

This objective evaluation method is applicable to other application types and is well suited for large-scale, automatic usability evaluations. Along with the event of interest, user wait time behavior can be automatically logged in system and software products. Automatic system logging enables a large-scale data collection for usability assessments without relying explicitly on user response through a form or survey.

CONCLUSION AND FUTURE WORK

This paper, in which we discuss waiting time in HCI with a focus on evaluation of temporal metaphors makes a threefold contribution. First, we identify how evaluation based on subjective measures can be influenced by contextual variables like cognitive load, retention delay, paradigm used, etc. The research on wait time has not explored objective assessment techniques. We identified the need for an objective assessment technique to minimize the influence of any contextual variables, generating objective outcomes like *How long a user will wait*. Second, we developed a method for analyzing the wait time data using survival analysis, which can be the basis for an objective assessment based user's wait time rather than perceived time. Survival analysis or other lifetime estimation techniques are popular among reliability engineers and medical scientists, and we have indicated its potential use within HCI for wait time experiment. Third, we demonstrated the use of the proposed objective assessment through a field experiment. We conducted this experiment at a women's helpline where callers are more impatient than callers at a standard call center probably due to stress (as indicated by data inspection of 29,548 calls). Based on our proposed objective assessment, we showed that temporal metaphors are useful in holding the callers longer, and the results were statistically significant. One of the limitations of our findings is that the majority of callers to 181-helpline are women who live in Delhi, and may resemble callers in a similar urban demography in terms of technology exposure and familiarity with call centers. Future studies will compare our approach to how they relate to subjective measures and test with temporal metaphors other than APB.

ACKNOWLEDGMENTS

Authors will like to acknowledge the support provided by TCS Research and ITRA project, funded by DEiTY, Government of India under grant with Ref. No. ITRA/15(57)/Mobile/HumanSense/01

REFERENCES

1. Development of nasa-tlx (task load index): Results of empirical and theoretical research.
2. In *Time, Action and Cognition*, vol. 66. 1992.
3. Allan, L. The perception of time. *Perception & Psychophysics 26*, 5 (1979), 340–354.
4. Allan, L. G. The influence of the scalar timing model on human timing research. *Behavioural Processes 44*, 2 (1998), 101 – 117.
5. Allon, G., and Bassamboo, A. The impact of delaying the delay announcements. *Operations research 59*, 5 (2011), 1198–1210.
6. Allon, G., Bassamboo, A., and Gurvich, I. "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations research 59*, 6 (2011), 1382–1394.
7. Armony, M., Shimkin, N., and Whitt, W. The impact of delay announcements in many-server queues with abandonment. *Operations Research 57*, 1 (2009), 66–81.
8. Bland, J. M., and Altman, D. G. The logrank test. *BMJ 328*, 7447 (2004), 1073.
9. Block, F., and Gellersen, H. The impact of cognitive load on the perception of time. In *Proc. of the 6th Nordic Conference on Human-Computer Interaction*, NordiCHI '10 (2010).
10. Block, R., and Zakay, D. Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review 4*, 2 (1997), 184–197.
11. Block, R. A. *Cognitive models of psychological time*. Psychology Press, 2014.

12. Block, R. A., Hancock, P. A., and Zakay, D. How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica 134*, 3 (2010), 330 – 343.

13. Bouch, A., Kuchinsky, A., and Bhatti, N. Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00 (2000).

14. Branaghan, R. J., and Sanchez, C. A. Feedback preferences and impressions of waiting. *Human Factors: The Journal of the Human Factors and Ergonomics Society 51*, 4 (2009), 528–538.

15. Church, R. M., and Broadbent, H. A. Alternative representations of time, number, and rate. *Cognition 37*, 12 (1990), 55 – 81. Special Issue Animal Cognition.

16. Feigin, P. D. Analysis of customer patience in a bank call center. Tech. rep., The Technion, Haifa, Israel, 2006.

17. Fröhlich, P. Dealing with system response times in interactive speech applications. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05 (2005).

18. Gaonac'h, D., and Larigauderie, P. *Mémoire et fonctionnement cognitif: la mémoire de travail*. Armand Colin, 2000.

19. Gibbon, J., Malapani, C., Dale, C. L., and Gallistel, C. Toward a neurobiology of temporal cognition: advances and challenges. *Current Opinion in Neurobiology 7*, 2 (1997), 170 – 184.

20. Grondin, S. Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics 72*, 3 (2010), 561–582.

21. Harrison, C., Amento, B., Kuznetsov, S., and Bell, R. Rethinking the progress bar. In *Proc. of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07 (2007).

22. Harrison, C., Yeo, Z., and Hudson, S. E. Faster progress bars: Manipulating perceived duration with visual augmentations. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10 (2010).

23. Hurter, C., Cowan, B. R., Girouard, A., and Riche, N. H. Active Progress Bar: Aiding the Switch to Temporary Activities. In *Proc. of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers*, British Computer Society (2012), 99–108.

24. Jouini, O., Akin, Z., and Dallery, Y. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management 13*, 4 (2011), 534–548.

25. Kleinbaum, D. G., and Klein, M. *Survival analysis*. Springer.

26. Kortum, P., Peres, S. C., and Stallmann, K. Extensible auditory progress bar design: Performance and aesthetics. *Int'l Journal of Human-Computer Interaction 27*, 9 (2011), 864–884.

27. Lallemand, C., and Gronier, G. Enhancing User eXperience During Waiting Time in HCI: Contributions of Cognitive Psychology. In *Proc. of the Designing Interactive Systems Conference*, DIS '12 (2012).

28. Lejeune, H. Switching or gating? the attentional challenge in cognitive models of psychological time. *Behavioural Processes 44*, 2 (1998), 127 – 145.

29. Munichor, N., and Rafaeli, A. Numbers or apologies? customer reactions to telephone waiting time fillers. *Journal of Applied Psychology 92*, 2 (2007), 511.

30. Myers, B. A. The importance of percent-done progress indicators for computer-human interfaces. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '85 (1985).

31. Nah, F. F.-H. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour & Information Technology 23*, 3 (2004), 153–163.

32. Nielsen, J. *Usability Engineering*. Morgan Kaufmann Publishers Inc., 1993.

33. Niida, S., Uemura, S., Nakamura, H., and Harada, E. Field study of a waiting-time filler delivery system. In *Proc. of the 13th Int'l Conference on Human Computer Interaction with Mobile Devices and Services*, ACM (2011), 177–180.

34. Peres, S., Kortum, P., and Stallmann, K. Auditory progress bars: Preference, performance and aesthetics. In *Proc. of the International Conference on Auditory Display (ICAD2007)* (2007).

35. Peto, R., and Peto, J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)* (1972), 185–207.

36. Sauro, J., and Lewis, J. R. Average task times in usability tests: What to report? In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, ACM (2010), 2347–2350.

37. Scapin, D. L., and Bastien, J. C. Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour & Information technology 16*, 4-5 (1997), 220–231.

38. Treisman, M. Temporal discrimination and the indifference interval: Implications for a model of the "internal clock". *Psychological Monographs: General and Applied 77*, 13 (1963), 1.

39. Zakay, D. Gating or switching? gating is a better model of prospective timing (a response to switching or gating? by lejeune). *Behavioural Processes 50*, 1 (2000), 1 – 7.

40. Zakay, D. Attention and duration judgment. *Psychologie Bulletin Francaise 50* (2005), 65–79.

41. Zakay, D., and Block, R. A. Temporal cognition. *Current Directions in Psychological Science* (1997), 12–16.

42. Zijlstra, F. R. H. *Efficiency in Work Behaviour: A Design Approach for Modern Tools*. Delft University Press, 1993.