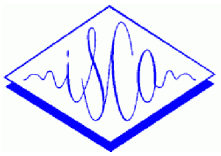
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Caller Response Timing Patterns in Spoken Dialog Systems

Silke M. Witt

Fluential, Inc.
1153 Bordeaux Dr, Sunnyvale, CA 94087, USA
switt@fluentialinc.com

Abstract

This paper contains an analysis of caller response timing patterns in spoken dialog systems. The findings presented here are based on data from live commercial dialog systems. It is shown that caller responses after a system finished playing the prompt resemble a uni-modal distribution and can be modeled with a rational distribution function. This finding allows understanding when callers tend to respond to a system depending on the dialog state type. Based on these findings a no-speech timeout optimization method is being proposed.

Index Terms: spoken dialog systems, turn-taking behavior, caller response timing, timeout optimization

10.21437/Interspeech.2012-81

1. Introduction

One of the biggest challenges in commercial spoken dialog applications is optimizing the turn-taking behavior of a system. One important component of such behavior are the pauses and no-speech timeout settings used by a system to determine when to take back a turn from the user in the case that the user is silent.

There exists extensive research on human-human turn-taking behavior, see, for example, [1]. Gravano et al. [2], studied the acoustics and prosody associated with human-human turn-taking in a gaming environment in order to learn what kind of cues indicate yielding a turn in a conversation. Likewise, substantial research has been conducted in the area of turn-taking in human-computer dialog systems. A number of studies have focused on developing systems that analyze prosodic and semantic cues to determine when a user has finished speaking versus just pausing during his/her utterance; for details see Raux et al. [3], Edlund et al. [4], and Schlangen [5]. Such research has resulted in methods to estimate when a caller is finished with speaking so that a system can take back the turn without interrupting the user.

However – at least to the knowledge of the author – there has been minimal work on optimizing the turn-taking scenario when a system gives the user the turn but the user does not take the offered turn, i.e. the user is silent. There has been some mentioning of this issue by Margulies [6] who concludes that "accuracy and success in IVR-based dialogues is closely correlated to more simplistic coupling between silence and syntax for touch-tone based systems and silence and prosody for speech-based systems." Generally, it has been observed that if the no-speech timeout setting is too short, a system tends to interrupt a caller when he/she is just pausing in a turn, see also Witt et al. [7]. Margulies [6] also points out this challenge between too-short and too-long timeout settings. He adds that the majority of turn-taking failures due to non-optimally chosen timeout settings were cases of too short timeouts. In order to optimize this particular turn-taking scenario when the user is silent and the system needs to take back the turn, this paper will focus on understanding the patterns of the elapsed time between the end of a system prompt and the onset of the user's response. The intent is to obtain an understanding of caller response timings across different spoken dialog systems. It is hoped that such information will prove useful in the design of efficient user interfaces as well as in the optimization of turn-taking via timeout settings.

2. Prior work

Bull et al. [8] studied inter-speaker intervals in human-human dialogues and found that these intervals between different speakers' utterances depend on the three groups of factors: 1) speaker differences; 2) task related factors and 3) dialogue structure related factors. The question that is to be studied in this work is how such findings relate to human-computer dialogs.

Levow [9] is one of the first to discuss the difficulties of optimal timings in spoken dialog systems. She describes how on the one hand callers tend to complain about a system's slowness while on the other hand, if the system responds too fast, turn-taking issues - such as the system and the user speaking at the same time - are likely to occur. Williams et al. [10] measured response timing differences between novices and expert users to two different dialog states and found that novice users actually tended to respond faster, sometimes even barging in on the system. Cummerford et al. [11] conducted a usability experiment with 6 participants measuring the response timings in a menu dialog state with and without help commands being offered and found that the response time a) depended on whether the option the user wanted was offered in the menu and b) that if the desired menu option existed, more than 90% of the users responded within 2.5sec.

3. Data Collection and Method

Since the aim of this work was to obtain a generic understanding of caller response timing patterns, we analyzed data taken from several live commercial deployments in form of different types of log files. These live deployments cover the following system domains: (1) Technical Support (for a computer manufacturer), (2) Energy company service, (3) Cable television call routing and (4) Banking self-service. The response timings for these applications were extracted from the recognizer log and application logs. The advantage of using data from live systems is that it provides realistic, un-biased data. The downside is that it sometimes is not possible to extract data for a particular variable while all other variables stay constant. For this reason data presented in section 4.2 is taken from experimental data; for details on that data see [7].

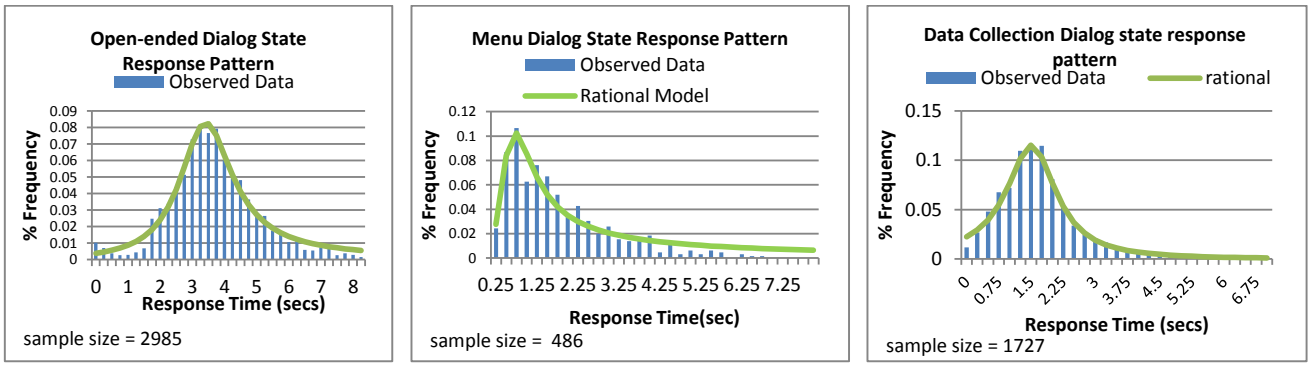


Figure 1: Response timing patterns for three dialog state types: (a) Open-ended dialog state , (b) Menu-style dialog state, (c) Data collection dialog state

4. Caller Response Timings

Based on the live data as described above, Figure 1 depicts the response timing patterns for three dialog state types: Open-ended ('How may I help you"), menu–style question and a data collection-style question (e.g. asking for an account number), starting at prompt end. In all three cases it can be seen that the distribution of responses from the end of a prompt onwards over time resembles a uni-modal distribution. Note also that especially in Figures 1(b) and 1(c) caller had the opportunity to respond anytime up to 7 seconds but there is no additional increase of responses later on, i.e. no secondary smaller curve. Note also that the data was taken from different dialog systems with different caller populations, yet they all display the same pattern. These observations lead to the hypothesis that caller response timings after prompt end tend to have a curved shape that can be modeled with a uni-modal distribution independent of system type and dialog state type.

In order to find a good model for this observed uni-modal distribution, we experimented with a several different common probability distributions such as Beta, Gaussian, Lorentzian, reciprocal and rational distributions. The best match was found with a rational distribution:

$$(1) \qquad G_{Rational}(x) = \frac{a + bx}{1 + cx + dx^2}$$

Figure 1 also depicts the estimated rational distribution in addition to the actual data. The parameters for the rational distribution can be estimated with standard, iterative methods, see [12]. The parameters of the rational models plotted in Figure 1, were estimated with the Levenberg-Marquart method.

Table 1 gives the goodness of fit for the example response patterns in Figure 1, where the "Goodness of fit" was measured via $\chi 2$ (based on Pearson's chi-square test) and r^2 (as the coefficient of determination).

In order to validate the hypothesis that the response time pattern after prompt end tends to resemble a uni-modal distribution with a single hump for most dialog states independent of dialog state type and system domain, Figures 2 and 3 shows the response time pattern for larger number of dialog states from several different spoken dialog systems. Figure 2 shows an overlay of the response time patterns for four different open-ended dialog states from four different systems. While there exists an expected variation in the exact shape of the curves, all curves share the same general shape.

Dialog state type	χ^2 Measure	r^2 Measure
Open-ended	0.000525	0.988851
Menu	0.001957	0.96746
Data Collection	0.000522	0.992456

Table 1:Goodness of Fit of the rational distribution for the response patterns from Figure 1

Likewise, Figure 3 shows the response timings for 18 different confirmation states. Again the overall pattern of a uni-modal distribution holds true with some statistical variations. Note that all responses happen within the span of the curve, for example there are no cases of two humps of different height following each other. Note also that the logfiles that these graphs have been derived from only logged data with a granularity of seconds, not millisecond which is the reason for the graph not being more granular.

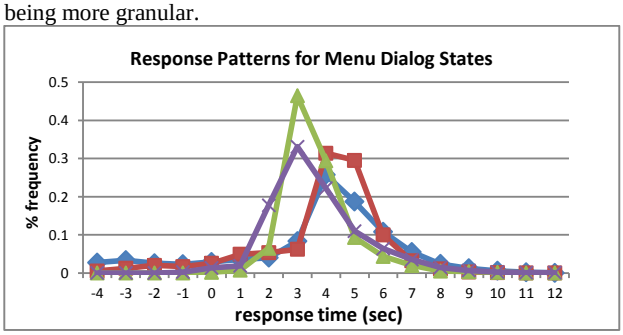


Figure 2: Response patterns for 4 different menu dialog states

Comparing Figures 1,2 and 3 shows that while the overall shape of the response pattern is very similar for all dialog states, the height and width of the curve depends on the individual dialog state.

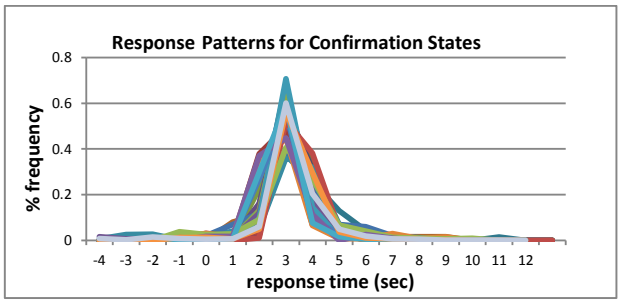


Figure 3: Response timings of 18 different confirmation dialog states

The question type (i.e. open-ended, menu, etc.) as well as other circumstances (such as application domain or caller population) tend to determine the exact shape of these distributions. For example, in dialog states where callers tend to know the answer to the system's question, callers tend to respond quickly and thus the mean of the frequency distribution will be smaller compared with a dialog state where callers have to think before answering a question. However, the common pattern between all curves is that they can be fitted with a rational distribution. An alternative way to examine the caller response timings for different dialog state types is to look at the cumulative frequency distribution for each dialog state type. The cumulative frequency distribution describes the percentage of callers that responded by a certain moment in time. For example, in Figure 4, only about 18% of the callers would have responded 1 second after prompt end in the case of the open-ended dialog state, whereas 60% would have responded in the case of the confirmation state.

Moreover, in the case of the confirmation state, close to 100% of all responses are complete by 3secs, whereas in the case of the open-ended state this does not happen until 6secs have elapsed.

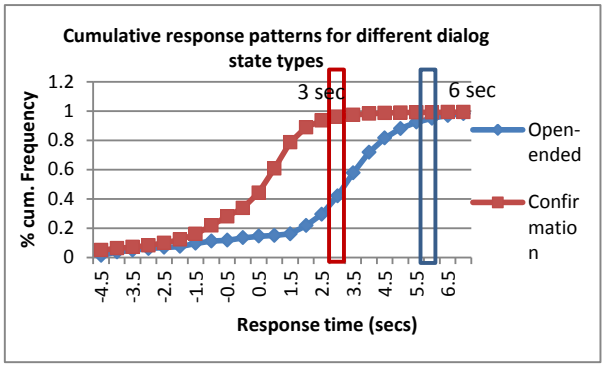


Figure 4: Difference in cumulative response time for open-ended question versus confirmation prompt

4.1. Response times dependency on context

Two other variables that influence the response timing patterns are: (1) the context of a dialog state and (2) the events preceding a dialog state. This section now shows an example for both cases. Answers to almost the same question and same type of callers, can result in very different response timing patterns, if the context is different.

The example presented here is from a deployed system for an energy company. In this application there exist two dialog states that are essentially the same, i.e. they ask for a date, but in one case, the date is for when to stop energy service to a home, whereas in the other case, it is when to start service. In other words, the difference between the two dialog states lies in their context. Figure 5 below shows the difference in the response patterns. A t-test showed that the difference between the two response patterns is significant with p=0.16 (α=0.05). Callers seem to be more prepared to provide a start date for energy service as opposed to when stopping energy service.

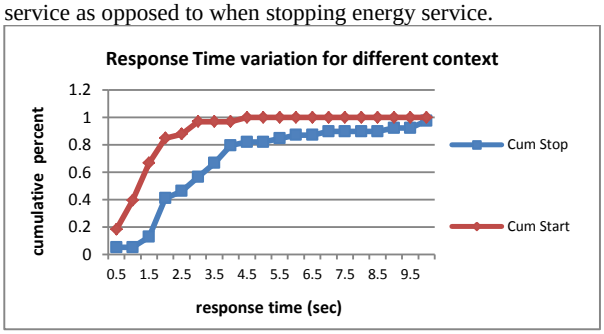


Figure 5: Dependency of the pattern on events in the previous dialog state

4.2. Response times dependency on preceding events

In Figure 5 we show the impact on the response pattern as a function of what happened in the preceding dialog state. The data is based on 696 data points and the difference between the three groups is significant with p=*0.00002 (α=0.05)*. The lower line indicates the response pattern in the case where there was no turn-taking issue in the preceding dialog state, i.e. there was a successful recognition event. 'Impatient barge-in' indicates that in the previous turn the user barged-in on the system. Note that this data was not available from live data and thus taken from an experimental study, see [7].

Clearly, those callers tend to respond faster as can be seen by the steeper rise of the blue curve when compared to the green line where the user had a successful turn before. Lastly, the steepest increase can be found in the case where the user had a turn-taking error in the preceding dialog state, i.e. the system and user were talking at the same and/or talking over each other. In this case, users seem to have learnt that they better respond fast in order to avoid problems in this turn in the dialog.

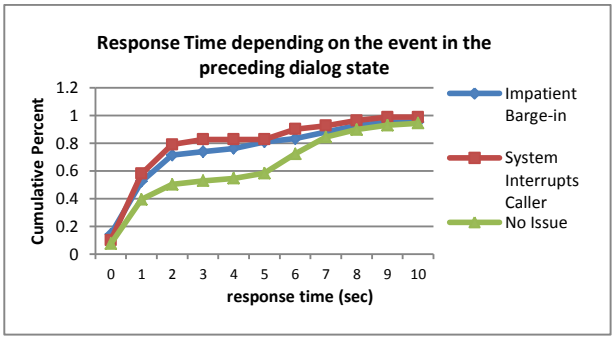


Figure 6: Dependency of the response pattern on the context around a dialog state

5. Timeout Setting Optimization

Typically, commercial spoken dialog systems only use a global timeout setting for all dialog states in a system. In other words, dialog systems do not take into account the differences in the response patterns for each dialog state. However, if a timeout is too short or too long in a particular dialog state, this can lead to serious turn-taking issues. Thus, it is proposed to use optimized local timeout values which minimize for a particular dialog state the duration of silence for those callers who decide not to response while at the same time ensuring that the 'almost all' callers who will respond would have responded. This can be defined as 95% or 98% of all callers. Therefore, an optimized timeout is defined as the time when for example 98% percent of all callers would have responded. Such timeout choices would be optimal in the sense that the system would give the overwhelming majority of all callers an opportunity to respond without unnecessarily waiting for those callers that will not respond, i.e. cause a timeout error.

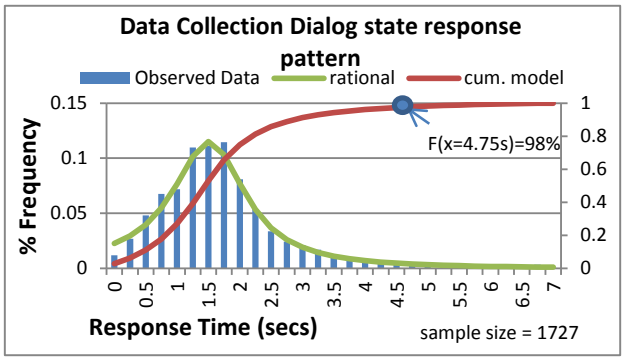


Figure 7: Estimated cumulative probability distribution

Let $F_{D_i}(t) = P_{D_i}(X \leq t)$ describe the cumulative response distribution for dialog state D_i with t denoting a response time. Thus, in order to find the optimized timeout setting for each dialog state, $F_{D_i}(t)$ has to be estimated. $F_{D_i}(t)$ can be modeled using the rational distribution function G(x). Since we are looking at the time period from the end of the prompt until the maximum observed response time, t_{max}, the distribution needs to be bounded at the lower end by 0 and t_{max} at the upper end. The cumulative distribution function also gets normalized by the preceding coefficient. With this, $F_{D_i}(t)$ becomes:

$$(2) \quad F_{D_i}(t) = \frac{1}{\int_0^{t_{max}} \frac{a+bx}{1+cx+dx^2}} \int_0^x \frac{a+bx}{1+cx+dx^2}$$

Table 5 depicts a list of example dialog states for which the optimal timeout has been calculated. As can be seen, the optimal timeouts where 98% of all callers will have responded vary significantly by dialog state. These variations match the observations about response time patterns reported in section 4. For example, for a confirmation state $t_{0.98}$ = 3.5s is a much smaller value than $t_{0.98}$ = 7.5s for an open-ended question.

Such individual timeouts have been partially implemented in a deployed dialog system and initial, anecdotal results indicate that this approach helps with shortening call durations and overall system success, however a detailed analysis is still in progress.

Dialog State ID	Type	χ^2	$t_{0.98}$
1	Open Ended	0.0005	7.5s
2	Menu	0.002	8s
3	Menu	0.015	4.5s
4	Menu	0.003	4s
5	Menu	0.004	4s
6	Data Collection	0.001	4.75s
7	Confirmation	0.007	3.5s

Table 2: Optimized timeouts for example dialog states

6. Conclusions and Future work

This paper presented results from an investigation on the response timing patterns of users of spoken dialog systems. Data from about over 30 dialog states and 6 different dialog systems was analyzed in order to derive a general hypothesis of response timing patterns. It was found that the histogram of caller responses tends to resemble a uni-modal distribution and can be modeled with a rational distribution. It was also shown that while the resemblance to a rational distribution is general, the shape of the distribution depends on the question type, the system context and other factors. Future work will focus on measuring the impact of the proposed no-speech timeout method on overall spoken dialog system success and call duration.

7. References

[1] Cowley, S.J.(1998). Of timing, Turn-Taking, and conversations. Journal of Psycholinguistic Research, 27(5):541–571, September 1998.

[2] Gravano A. and Hirschberg J. (2010). Turn-taking cues in task-oriented dialogue. Computer Speech & Language.

[3] Raux A. and Eskenazi M. (2010). Optimizing end-of-turn detection for spoken dialog systems. In Workshop on Modeling Human Communication Dynamics at NIPS 2010.

[4] Edlund, J., Heldner, Mattias, & Gustafson, J. (2005): Utterance segmentation and turn-taking in spoken dialogue systems. In Fisseni, B., Schmitz, H-C., Schröder, B., & Wagner, P. (Eds.), Computer Studies in Language and Speech (pp. 576-587). Frankfurt, Germany: Peter Lang.

[5] Schlangen, D. (2006). From reaction to prediction: Experiments with computational models of Turn-Taking. Interspeech, Pittsburgh, USA.

[6] Marguilies, E. (2004). Adventures in Turn-Taking. Notes on Success and Failure in Turn Cue Coupling. Sterling Augits and Cosulting Inc.

[7] Witt S.M., Loose R., Rolandi W., Master A., Zuber E., Brooks T. (2010b). Optimizing Successful Turn-taking in Spoken Dialog Systems, HFE 2010, San Francisco, USA.

[8] Bull. M. and Aylett, M., An analysis of the timing of Turn-Taking in a corpus of Goal-Oriented dialogue.

[9] Levow, G.-A. (1997). Making sense of silence in speech User Interfaces. CHI'97.

[10] Williams, D. and Cheepen, C. (1998). 'The sound of silence': A preliminary Experiment Investigating Non-Verbal auditory representations in Telephone-based automated spoken dialogues. ICAD.

[11] Commarford P.M., Lewis J.R, (2005), Optimizing the Pause Length before Presentation of Global Navigation Commands, Proc. of HCI 2005, vol 2, p.1-7, St Louis, USA.

[12] Numerical Recipes. The Art of Scientific Computing, 3rd Edition, 2007, Cambridge University Press, www.nr.com.