How Long is Too Long? Excessive Pause Duration in Voice User Interfaces

Lucas J. Hess, Daniela Barron

Table of Contents

Executive Summary

The purpose of this paper was to answer this question: at what point is too long of a pause between when a human speaker provides input and a Voice User Interface (VUI) responds before negative user experience effects occur? When interacting with VUI's, humans have temporal perceptual expectations of this interaction which result from experiences such as human-to-human conversation (Cohen et al., 2004). When these expectations are violated, this may result in interfaces that are perceived as less comfortable, have less flow, and are more difficult to interact with and comprehend, which may result in more errors (Cohen et al., 2004).

During this investigation, it was revealed that the research on human temporal perception of voice feedback in VUI's was scarce. Fortunately, humans have perceptual expectations when interacting with VUI's which result from experiences, such as human-to-human conversation. Thus, by using this scarce VUI literature and utilizing researched norms in human-to-human conversation, an understanding into the perceptual expectations of the users related to VUI interactions was achieved (Gravano & Hirschberg, 2011).

Before investigating excessive durations of inter-speaker pauses, we reviewed research related to average durations of pauses in human and VUI interactions. By utilizing this research, we can estimate a normal conversational experience and thus, compare temporal feedback that is past this norm, which may be perceived as undesirable in VUI design (Gravano & Hirschberg, 2011). This overview of research demonstrated a range of 100-500 ms for inter-speaker pause durations and appeared that the general average duration of inter-speaker pauses indicated by this research is approximately 300-350 ms.

Following this analysis, we investigated the max duration of a pause between two speakers. This review revealed a range of approximately 500-1300 ms for max duration of an inter-speaker pause before negative effects to the user's experience ensue. The general average duration of excessive pause time is deduced to be around 1000 ms, or 1 second, ± 100 ms. The high variability in reported average and max durations for inter-speaker pauses may be due to many factors, such as if the conversation is task oriented or if a different language than English is spoken (ten Bosch, 2005; Stivers, 2009).

Although there is some variability in these results, this paper provides an estimate of excessive pause durations and reviews the negative effects associated with late conversational feedback which can increase errors and be at the detriment of the user's experience in VUI interactions (Gravano 2009; Wilson and Wilson 2005). Thus, the importance of implementation of aduration cap of 1000 ms, or 1 second (± 100 ms), into operational settings is vital for the user's experience, general function of the VUI, and consequently, the success of the whole business.

Introduction

A Voice User Interface (VUI) is what an individual interacts with in a spoken language application in order to accomplish a task or receive assistance, such as Apple's "Siri" or automated attendants in information technology support (Cohen et al., 2004). In the design of VUI's, there are many factors that are important for proper design, such as prompts, prosody, grammar, and call flow (Cohen et al., 2004). When interacting with VUI's, humans have perceptual expectations of this interaction which result from experiences, such as human-to-human conversation (Cohen et al., 2004). Within these expectations, individuals have expectations about the temporal characteristics of VUI conversation interactions, such as inter-speaker gaps, overlaps, and intra-speaker pauses (Levinson & Torreira, 2015).

When these expectations are violated, this results in interfaces that are perceived as less comfortable, have less flow, and are more difficult to interact with and comprehend, which may result in more errors (Cohen et al., 2004). Specifically, a human's temporal perceptual expectations as to when a VUI should provide feedback and respond to their prompts may result in these previously mentioned negative effects (Commarford & Lewis, 2005). Thus, the temporal design of VUI's is important for the user's experience of the VUI. The purpose of this paper was to investigate the human temporal perception related to VUI feedback. Our main goal was to answer this question: at what point is too long of a pause between when a human speaker provides input and a VUI response, before a negative user experience effects occur?

Research Overview

During the investigation, it was revealed that the research on human temporal perception of a VUIs voice feedback was scarce. Consequently, our research also incorporated research in human-human conversation. By utilizing norms in human-human conversation, we can understand the perceptual expectations of the users related to VUI interactions (Gravano & Hirschberg, 2011). Since non-speech and non-verbal cues can also be utilized in conversation more likely during face-to-face conversation than over the phone, such as breathing preparation cues (Torreira et al., 2015), the amount of face-to-face literature utilized in this paper was reserved and the goal was to focus on research which utilized telephone and no eye contact interactions. Research is described in the "Relevant Summary" sections of the tables for an understanding of the breadth of research included in the following review tables.

Average Duration of Inter-Speaker Pauses

Table 1 demonstrates a comprehensive review of the general averages in duration of inter-speaker pauses during human-human conversations (face-to-face and not face-to-face), and during interactions with VUIs.

Table 1: Average Durations of Inter-Speaker Pauses Indicated by Past Research

Authors	Duration (ms)	Relevant Summary
Baumann (2008)	331-363	Investigated turn-taking strategies in a simulated environment. Participants exchanged audio streams in real-time, and autonomously judge turn-taking behavior.
Beattie & Barnard (1979)	250	Investigated timing of turn taking during American English service-based conversations over the phone.
Brady (1968)	345-456	Investigated gaps in sixteen phone calls between friends in the USA.
Gravano (2009)	100-200	Investigated the final and initial utterances of turns in a conversation using task-oriented dialog, and ways to potentially predict what kind of turn-yielding cues someone might be using for applications into VUI systems.
Holler et al. (2016)	100-500	Analyzed human-to-human conversational structure, and presented an overview of the research and literature surrounding turn-taking in conversations.
Kendrick & Torreira (2015)	300	Indicated from corpus analysis that gaps longer a norm of 300 ms decrease likelihood of an unqualified acceptance and dispreferred turn format.
Norwine & Murphy (1938)	410	Investigated pauses in calls on a New York-Chicago telephone circuit used for Bell System business.
Sellen (1995)	480	Videoconferencing systems were evaluated experimentally and differed based on if participants were visible and if they were in the same room. The duration included is when speakers were not visible to each other.
Stivers (2009)	200	Investigated universal basis for turn-taking behavior demonstrated between all languages studied.

Weilhammer & Rabold (2003)	380	Investigated task-oriented telephone conversation and pauses between English, German, and Japanese. The mean reported is English speakers.
Wilson & Wilson (2005)	110-400	Investigated using brain oscillation as a technique to understand turn-taking in conversation and delves into how a speaker and a listener can become entrained by identifying rate of speech and syllable production.

Table 1 is to serve as reference for how long typical pause durations tend to be in VUIs, so that we can understand the difference between average pause durations compared to max durations. Using this research, we can estimate a normal conversational experience and thus, understand temporal feedback past this norm which may be perceived as undesirable in VUI design (Gravano & Hirschberg, 2011). Understanding the average durations indicated in this section allow us to understand where a noticeable difference in temporal feedback is perceptually apparent or undesirable to the user.

The research summarized in Table 1 demonstrates a large range of 100-500 ms for inter-speaker pause durations. It appears that the general average duration of inter-speaker pauses indicated by this research is approximately 300-350 ms. This large amount of variability in the range may be due to different factors, such as conversational contexts and characteristics of the speakers (discussed later in limitations; Gravano, 2009).

Max Duration of Inter-Speaker Pauses

Table 2: Max Durations of Inter-Speaker Pauses Indicated by Past Research

Author	Duration (ms)	Relevant Summary
Beattie & Barnard (1979)	1250	Investigated temporal characteristics of speaker transitions in natural telephone conversation.
Commarford & Lewis (2005)	1300	Presented analysis on optimal pause duration between menu presentation and global navigation commands in a VUI system.
Heldner and Edlund (2010)	500-1000	Explored durational aspects of pauses, gaps, and overlaps in conversational corpora for use in speech technology design.
Kendrick & Torreira (2015)	700	Corpus Analysis demonstrated that gaps longer than 700 ms indicated negative effects.

Roberts et al. (2011)	600	Universal temporal mechanisms of spoken language were investigated using telephone conversations between friends.
Wilson and Wilson (2005)	910-1,000	Investigates using brain oscillation as a technique to understand turn-taking in conversation, and delves into how a speaker and a listener can become "mutually entrained through recognizing rate of speech and syllable production."

Table 2 demonstrates the average duration of studies investigating the max duration of inter-speaker pauses within human to human conversations and VUI interaction. Consequently, Table 2 summarizes the research available that can indicate what is an undesirable duration between when a user speaks and when a VUI responds in order to avoid a displeasing user experience.

The research in Table 2 demonstrates a max duration of inter-speaker pauses before negative effects ensue as being at a range of approximately 500-1300 ms. The general average duration of excessive (max) pause time seems to be around 1000 ms, or 1 second, ± 100 ms. The high variability indicated in this large range of 500-1300 ms by this research may be due to many different factors such as the duration of the response and whether the conversation was task-oriented (discussed later in limitations; ten Bosch et al. 2005; Gravano 2009).

Effects of Excessive Pause Durations

As a consequence of excessive pause durations, negative effects to the perception of the user can result. For example, a user interacting with a VUI may want to speak again at an excess of 1250 ms (Beattie & Barnard, 1979; Commarford & Lewis, 2005). Roberts and colleagues (2011) demonstrated that a pause of excess of 600 ms generates negative inferences about responses as the user perceives the likelihood of a dispreferred response to be greater. Kendrick and Torreira (2015) similarly indicated that longer pauses, excess of 700 ms, are perceived as a decreased likelihood of a preferred response. Kendrick and Torreira (2015) also indicated that anything above the norm duration of 300 ms, or average duration (summarized for all studies in Table 1), decreased the likelihood of a general acceptance and increased the likelihood that a response will possess a dispreferred turn format (e.g. more likely to result in: “Yes, but…”). Given this research, excessive pauses do not just make users uncomfortable (Cohen et al., 2004), but also can result in users making inferences about the upcoming responses and respond out of turn (Beattie & Barnard, 1979; Commarford & Lewis, 2005; Roberts et al., 2011). If a user takes the floor, or speaks, when the system also takes the floor, then this may result in

errors in the VUI system or be at the detriment of the user's experience (e,g, user being confused by VUI's turn-taking pause behavior). The importance of ensuring that excessive pauses are not implemented into IVR and VUI systems is thus imperative for the general functioning of the system and to the user's experience.

Factors Contributing to Variability in Pause Duration

There are many factors that can influence human perceptions and experiences of pause durations with VUI's. One limitation is that inter-conversational pause durations vary for different languages, such as English (380 ms), German (363 ms), and Japanese (389 ms) (Weilhammer & Rabbold, 2003). However, other research has argued that this is negligible difference and noted that the factors that affect response times are often similar, regardless of language or culture (De Ruiter et al. 2006; Levinson et al., 2015; Norwine and Murphy 1938; Sellen 1995; Stivers et. al. 2009). Despite the controversial evidence, there appears to be a dearth of research on the differences of turn-taking conversational gaps amongst different languages; thus, more knowledge about the area may be beneficial in potentially better understanding user pain-points of foreign language speakers, and lead to better accommodations and user experience for that particular audience.

Another limitation in this research is providing an exact inter-speaker duration threshold, as individual differences have been demonstrated in this domain (Gravano, 2009; Wilson & Wilson, 2005). Brady (1968) studied a corpus of 16 phone calls between friends in the USA, and reported that average pause duration in conversations depends mostly on the threshold used for the automatic speech detection (speech detectors can have thresholds set for rejecting noise or recognizing potentially extraneous noise that may occur during a conversational gap (Brady 1968). Research has also demonstrated that individuals tend to match their new conversational partners in terms of pause durations (ten Bosch et al., 2004, 2005; Wilson and Wilson 2005; Gravano 2009). Gravano (2009) argued that there are several ways a conversational partner may signify they are about to stop speaking (turn-yielding cues) and that each different turn-yielding cue combination typically warrants a different turn-taking interval. Recent research has also argued that the longer a speaker plans to speak, more cognitive preparation is needed to produce longer responses, thus warranting longer pauses (Torreira et al., 2015). From these studies, we can deduce that there is no exact, set time interval between conversational turns between users, and that this threshold may be flexible depending on a number of factors. However, the most common time intervals between conversational turns are typically within a hundred milliseconds of each other, creating the potential for the use of a mean time a VUI may wait after a user stops speaking before beginning its turn.

Finally, some research included in this paper used methodologies of face-to-face conversations, while the majority focused only on the exchange of audio. Face-to-face conversation methods may have other factors influencing conversational turn-taking besides pitch, prosody, intonation, or respiratory cues, and thus may have a different turn taking interval than normal phone conversations. For instance, Bosch et. al. and Levinson noted that transition speeds are higher on the phone than face-to-face (Levinson, 1983; ten Bosch et al., 2005).

Conclusions

Even though there is no exact duration threshold that will be passed which results in negative effects, we can estimate a duration that should not be exceeded in VUI design based on the research compiled in this paper. For reference, we first examined the average duration of inter-speaker pauses and observed a range of 100-500 ms. Although, the average durations were generally concentrated at approximately 300-350 ms. Thus, an average duration of 300-350 ms can be used for comparison against the main goal of this paper, identifying the duration of an excessive inter-speaker pause in VUI design.

The excessive inter-speaker pause duration according to past research examined in this paper was approximately 1000 ms, or 1 second (± 100 ms). Although, the variability of this research was rather high, specifically with a range of 500-1300 ms. The variability in average pause length was also quite high (range = 100-500 ms). This variability may be the result of a number of factors that can affect temporal perception, including whether conversations were in English or another language, whether the conversation took place on the phone or face-to-face, what turn-yielding cues someone uses, as well as the frequency of a particular turn-yielding cue (Gravano 2009; Wilson and Wilson 2005).

Given the limitations of the research reviewed in this paper for application to VUI design, future research should investigate a possible more exact, but probably still flexible, duration threshold for excessive pauses in human-machine voice interaction and VUI design. In the meantime, this paper provides an estimate of excessive pause durations and the negative effects associated with late conversational feedback by utilizing a broad breadth of research from VUI interaction to human-to-human conversational behavior. The research reviewed in this paper demonstrates that the negative effects of late conversational feedback can increase errors and be at the detriment of the user's experience (Gravano 2009; Wilson and Wilson 2005). Thus, the importance of implementation of the duration cap of 1000 ms, or 1 second (± 100 ms) into operational settings is vital for the user's experience, general function of the VUI, and consequently, the success of the business.

References

Note: All available references are linked; however, some may require a journal subsciption to view the article

Baumann, T. (2008). Simulating spoken dialogue with a focus on realistic turn-taking. Proceedings of the 5th International Workshop on Constraints and Language Processing, 1-8.

Beattie, G. W., & Barnard, P. J. (1979). The temporal structure of natural telephone conversations (directory enquiry calls). *Linguistics*, *17*(3-4), 213-230.

Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. Speech Communication, 47(1-2), 80-86.

Brady, P. T. (1968). A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, *47*(1), 73-91.

Cohen, M. H., Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional. Print.

Commarford, P. M., & Lewis, J. R. (2005). Optimizing the pause length before presentation of global navigation commands. In Proceedings *of HCI*, 2, 1-7.

Gravano, A. (2009). Turn-taking and affirmative cue words in task-oriented dialogue. Columbia University Computer Science Technical Reports. *Department of Computer Science, Columbia University,* CUCS-009-09, 1-219.

Gravano A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language,* 25 (3), 601-634.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555-568.

Holler, J., Kendrick, K. H., Casillas, M., & Levinson, S. C. (2016). *Turn-taking in human communicative interaction*. Frontiers Media.

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, *52*(4), 255-289.

Levinson, S. (1983). *Pragmatics.* Cambridge: Cambridge University Press. Print.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, *6*, 731.

Norwine, A. C., & Murphy, O. J. (1938). Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, *17*(2), 281-291.

Roberts, F., Margutti, P., & Takano, S. (2011). Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese. *Discourse Processes*, *48*(5), 331-354.

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human-computer interaction*, *10*(4), 401-444.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.

Torreira, F. J., Bögels, S., & Levinson, S. C. (2016). Breathing for answering. The time course of response planning in conversation. *Frontiers in Psychology: Turn-Taking in Human Communicative Interaction, 135-145.*

Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. In *Proceedings of the International Conference of Phonetic Sciences*, 2145-2148.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic bulletin & review*, *12*(6), 957-968.