A Large-Scale User Study of an Alexa Prize Chatbot: Effect of TTS Dynamism on Perceived Quality of Social Dialog

Michelle Cohn[1], **Chun-Yen Chen**[2], **Zhou Yu**[2]
[1]Department of Linguistics, [2]Department of Computer Science
University of California, Davis
{mdcohn, abtchen, joyu}@ucdavis.edu

Abstract

This study tests the effect of cognitive-emotional expression in an Alexa text-to-speech (TTS) voice on users' experience with a social dialog system. We systematically introduced emotionally expressive interjections (e.g., "Wow!") and filler words (e.g., "um", "mhmm") in an Amazon Alexa Prize socialbot, Gunrock. We tested whether these TTS manipulations improved users' ratings of their conversation across thousands of real user interactions (n=5,527). Results showed that interjections and fillers each improved users' holistic ratings, an improvement that further increased if the system used both manipulations. A separate perception experiment corroborated the findings from the user study, with improved social ratings for conversations including interjections; however, no positive effect was observed for fillers, suggesting that the role of the rater in the conversation—as active participant or external listener—is an important factor in assessing social dialogs.

1 Introduction

Dialog systems, despite recent improvements, still face a fundamental issue of how to convey interest and emotion via text to speech (TTS) synthesis. Many TTS voices have been described as "robotic" or "monotonous" by human listeners (Baker, 2015), an issue further exacerbated for generation of longer utterances (Németh et al., 2007). This is particularly relevant for non-task-oriented dialog systems, such as those that aim to engage users in social chitchat (Akasaki & Kaji, 2017; Liu et al., 2017); for example, Tokuhisa & Terashima (2009) found that affective (i.e., emotion conveying) productions relate to perceptions of speaker enthusiasm in non-task-oriented human-human conversation. In another study, adjustment of the prosodic features of computer TTS affects listeners' perceptions of the system's type of clarification request (Skantze et al., 2006), signaling its "cognitive state". Still, the ability to design a computer or robot system to convey cognitive-emotional expressiveness remains an area of rich study in the field of Affective Computing (AC) (cf. Tao & Tan, 2005). While prior approaches to model human-like expressiveness in various systems have involved manipulation of the overall TTS prosody, including pitch, rate, and volume (e.g., Gálvez et al., 2017; Henning & Chellali, 2012; Montero et al., 1998; Mustafa et al., 2010; Nass & Lee, 2001; Schröder, 2007), the present paper tests whether adding minimal and discrete emotional-cognitive expressions in a TTS voice impacts user experience with a social dialog system. More specifically, we examine whether a full "overhaul" of prosody is necessary to meaningfully improve a dialog system, or whether we can inject units of cognitive-emotional expression in carefully specified locations to produce a similar effect.

Yet, our understanding of what types of TTS modifications will result in believable and sincere expressions of emotion and cognitive states in a dialog system remains an open question; there have been mixed findings as to whether "human-like" TTS adjustments, such as adding filler words, result in improved user metrics (e.g., Syrdal et al., 2010; Pfeifer & Bickmore, 2009).

Critically, the vast majority of human-computer dialog studies have been run on a limited number of participants and conversations (e.g., n=96 in in Brave et al., 2005) and in a lab setting where users are recruited to interact with the systems (e.g., Brave et al., 2005; Cowan et al., 2015; Qvarfordt et al., 2005; Yu et al.,

Proceedings of the SIGDial 2019 Conference, pages 293–306
Stockholm, Sweden, 11-13 September 2019. ©2019 Association for Computational Linguistics

2016); that is, users may not be interacting with real intents. For one, the presence of an experimenter could impact the way users interact with the system (cf. Orne, 1962). This is also true for dialog systems; users may be less comfortable to engage in more naturalistic conversation, or may be more willing to accept errors or incongruencies by a computer system while in the lab. Additionally, having fewer observations, as well as a participant pool largely consisting of college age students (e.g., Cowan et al., 2015) may impact researchers' ability to generalize findings to other user demographic groups (cf. Henrich & Heine, 2010).

In this paper, we describe an experiment where we systematically manipulated the Amazon Alexa TTS generation in Gunrock, the 2018 Alexa Prize winner socialbot (Chen et al., 2018). Our participants included over 5,000 real users who engaged with the system from their own homes and devices. We targeted two types of TTS manipulations: interjections (e.g., "Awesome!") and filler words. We selected these two elements as they are ways humans communicate their cognitive-emotional states, but vary in their intensity: while interjections express enthusiasm and strong emotion, filler words communicate the speaker's cognitive states (e.g., "Um... let me think") in a more tempered fashion. Both interjections and fillers have also been proposed to serve as socio-affective "glue" between interlocutors, expressing emotional and cognitive states that serve to strengthen relational bonds between humans and computers (Auberge et al., 2013; Sasa & Auberge, 2014; 2017).

In addition to its scope, this study is novel in several regards. First, no prior work, to our knowledge, has explored how individuals respond to emotion generated by a voice-activated digital assistant (e.g., Amazon's Alexa, Apple's Siri); users may have a more personal connection with and may even show greater personification of these increasingly prevalent household devices (Lopatovska, & Williams, 2018). Additionally, this paper introduces a methodology for designing and inserting interjections and filler words, both in terms of their context as well as their acoustic adjustments using Speech Synthesis Markup Language (SSML). Furthermore, no prior experiments have parametrically tested the presence of these two elements in controlled studies; doing so allows us to test whether there is a cumulative effect of these cognitive-emotional insertions. Finally, conducting an experiment directly through the Alexa system is an innovative approach that builds on past work that has largely relied on naturalness ratings of synthetic voices with no interactive component for the rater themselves (e.g., Marge et al., 2010; Gálvez et al., 2017; Hennig & Chellali, 2012; Schmitz et al., 2007).

This study can serve as a test to the 'Computers are Social Actors' theoretical framework (CASA: Nass et al., 1994; Nass & Moon, 2000) that proposes that humans apply social norms from human-human interaction to *computers* when they detect a cue of humanity in the system. One empirical question for the CASA framework is what cues can trigger computer personification and to what extent this personification graded; that is, do we see cumulative effects of introducing multiple human-like features in a dialog system, or do listeners display a more categorical response to human-likeness? In particular, we ask whether individuals' ratings of social dialog quality vary according to the type and combination of addition for interjections and filler words.

In the following section, we will review the literature for related work on cognitive-emotional expression via interjections and filler words in human-human and human-computer interaction (HCI). Then, we will introduce our overall chatbot dialog system design and our interjection/filler insertion methodology in Section 3, our user study experiment in Section 4, and a perception experiment in Section 5.

2 Related Work

2.1 Limited Prior Work on Interjections and Exclamations in HCI

Despite the prevalence of interjections in human speech patterns, few groups have explored inserting interjections in TTS systems. In human speech, interjections constitute words or phrases that can display emotion (e.g., emotive interjections such as "Yuck!"; cf. Wierzbicka, 1999) or reveal the speaker's "information state" (e.g., "Aha!"). Some interjections are based on existing words (e.g., "Neat!"), while others are based on non-lexical vocal productions (e.g., "Ooh!"; cf. Yang, 2010). Interjections can also signal that the information is newsworthy (e.g., "Really?" in Pammi, 2012). Still, the addition of interjections in TTS voices remains a largely understudied area, while much greater attention has been given to overall prosodic adjustments over the scope of a phrase or utterance (e.g., pitch, duration, etc.) (e.g., Németh et al., 2007) or the introduction of non-linguistic affective bursts in robots (e.g., beeps, buzzes in Read & Belpaeme, 2012). While not introducing interjections per se, but rather modeling new TTS productions based on positive or negative interjections (e.g., "Great!" vs. "Oh dear!"), Syrdal and colleagues (2010) found that speech trained on positive exclamations resulted in higher listener ratings in a 7-utterance simulated dialog; they observed no such effect for TTS adjustments for negative exclamations (e.g., "Oh dear!", "Oops!"). One novel line of research we explore in the present study is whether the presence of an interjection – and the degree of prosodic dynamism in the interjection, such as exaggerating the pitch contour and increasing duration – contributes to a user's perception of the system as being more cognitive-emotionally expressive.

2.2 Mixed Results for Fillers in HCI

Another element signaling cognitive-emotional expression in human conversations is filler words. In certain instances, filler words, or filled pauses (e.g., "um"), can be considered to be a type of disfluency or hesitation in a speaker's production (Clark & Tree, 2002), demonstrating more time for the speaker to "collect" their thoughts (cf. Brennan & Williams, 1995). At the same time, filler words can signal information about the speaker's cognitive state; for example, longer filler words have been shown to signal greater uncertainty or degree of thought on the conversational subject, while the pitch contour on the filler word communicates the speaker's level of understanding (Ward, 2004). In some studies, introduction of filler words in dialog systems has a facilitatory effect on perceived naturalness and expressiveness of the voice (Gallé, et al., 2017; Goble & Edwards, 2018; Marge et al., 2010; Wigdor et al., 2016). For instance, a user's "sensation of engagement" in a conversation with a robot improves with the addition of filler words (Gallé, et al., 2017). Filler words additionally have been shown to impact perceived likeability and engagement with a computer, even for individuals not directly talking to the computer/robot; independent raters gave higher naturalness ratings for "overheard" human-computer conversations when the computer voice included filler words (e.g., using the Talkie dialog system in Marge et al., 2010).

Yet, at the same time, other studies have reported no effect of introducing filler words (e.g., "Hmmm", "uh huh" in Syrdal et al., 2010), or a negative effect for some listeners (e.g., Pfeifer & Bickmore, 2009). This negative response might be expected given their association with as markers of anxiety and unpreparedness for some subjects. However, Christenfeld (1995) additionally observed that listeners' evaluations varied based on their task: when asked to focus on the speech *style*, subjects reported more negative ratings of the filler "um", but subjects had no such negative judgments when they were asked to focus on the content. This raises an important question: how might the experimental task impact the way users perceive these more human-like, but in some cases more "marked", displays of cognitive-emotional expressiveness? Addressing a limitation of prior work having subjects rate stimuli presented in isolation (e.g., Syrdal et al., 2010), our study tests both actual user's responses as well as external raters in assessing the introduction of fillers.

3 Dialog System Design Amazon Alexa Prize Chatbot

For the past two years, Amazon has launched the Alexa Prize Socialbot Challenge to support universities in building conversational bots to advance human-computer interaction. General public users with an Alexa-enabled device or free Alexa application can access the system and talk to the system about various topics (e.g., music, sports, animals, movies, food, weather, etc.) in a conversational manner. When a user engaged the social mode by saying "Let's chat", one of the socialbots in the competition was randomly invoked. After talking to the system, the Alexa Skill system automatically solicited user feedback ("How likely are you to talk to this bot again, on a scale from one to five?"), providing a measure of user engagement.

Competing in the 2018 Alexa Prize competition, our chatbot, Gunrock (Chen et al., 2018), aims to produce engaging and coherent conversations with real human users. During the competition, our bot achieved an average rating of 3.62 (on a 1-to-5 scale) in over 40,000 conversations; conversations had an average of 18.9 turns, averaging 4.35 minutes in duration. Our bot uses automatic speech recognition and text-to-speech models are provided by Amazon. It has a three-stage natural language understanding pipeline including ASR correction, sentence segmentation, constituency parsing, and dialog act prediction to aid user intent detection. Our system has a hierarchical agenda-based dialog manager that covers different topics, such as movies, music, etc., and a templated-based natural language generation module that allows the system to fill slots with data retrieved from various knowledge sources. Please refer to Chen et al. (2018) for system implementation details.

3.1 Methods of Inserting Interjections (Speechcons)

We designed a framework to introduce 52 distinct interjections pre-recorded by the US English Alexa voice actor. These interjections, known as Speechcons (Amazon, 2018), are "special words and phrases that Alexa pronounces more expressively". For a listening sample, refer to the Speechcon website (Amazon, 2018). We inserted these interjections using Speech Synthesis Markup Language (SSML) tags in the Alexa Skills Kit. These interjections were longer in duration and showed wider pitch variations and exaggerated pitch contours, relative to their unmodified counterparts (see Figure 1).

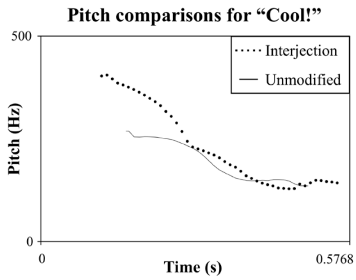


Figure 1: Pitch and duration differences for Speechcon and unmodified production of "Cool!" generated in Praat (Boersma & Weenik, 2018).

Of the 52 interjections (see Table 1 for a breakdown), we inserted 39 phrase-initially using a rule-based system, for the following 5 contextual scenarios, defined by conversational template: when the bot wanted to signal interest about the user's response to encourage the user to elaborate, to resolve an error, to accept a request, to change the topic, and to express agreement of opinion. In each context, we randomly inserted an interjection appropriate for that context (from the subset of pre-categorized interjections) to increase variation and retain user interest. Note that insertion of interjections did not result in any pauses or other incongruencies in the Alexa TTS generation.

Context	Purpose	# Interjections
1	Signal interest	12
2	Error resolution	14
3	Accept request	4
4	Change topic	4
5	Agree	2
Utterance-specific		13
	TOTAL:	**52**

Table 1: Total number of possible interjections added to defined slots in conversational templates.

Interjections were selected for each context by a native English speaker (Author 1) based on the acoustic production of the interjection and its semantic/pragmatic fit in the utterance. First, we selected positive interjections (e.g., "Wow!") that could be used to signal interest (Context 1) and negative interjections (e.g., "Darn!") in error resolution (Context 2); we used the widest variety of interjections for these two contexts as these situations arose most frequently in conversation. We denote the interjection version of words with an exclamation (e.g., "Awesome!").

- **Context 1: To signal interest about the user's response and elicit user's expansion.** We added 12 interjections phrase-initially to show Alexa's interest in the user's answer (after Alexa asks a question and the user provides a response); these interjections included "Awesome!", "Cool!", "Fantastic!", "Super!", "Wow!", "Ooh la la!", "No way!", "Fancy that!", "Interesting!", and more (for a full list, see Appendix A). For example: "[*Wow!...* | *Interesting!...* | *Ooh la la!...*]. Tell me more about it."
- **Context 2: Error resolution.** We also introduced 14 interjections in error resolution templates in order to show Alexa's "feelings" about her misunderstanding. Possible interjections included "Whoops a daisy!", "Darn", "Oh brother". For example: "[*Whoops-a-daisy!...* | *Baa!...* | *Darn!...*] I think you said probably. Can you say that one more time?"

- **Context 3: To accept a request.** We inserted 4 interjections phrase-initially to reflect Alexa's acceptance of the user's request (e.g., such as to change topic), including: "Okey dokey!", "Righto!", "As you wish!" and "You bet!". For example: "[*Okey dokey!... | Righto!... | As you wish!...*] Here's some more info."
- **Context 4: To change the topic.** We used 4 interjections to transition to a new topic, simulating a scenario where Alexa "just remembered" something she wanted to share with the user. We generated 2 interjection versions of "Ooh!" and "Ah!" to use in this context. For example: "[*Ooh!... | Ah!... | All righty!...*] tell me more about you! What else are you interested in? Do you like [music | movies | animals]?"
- **Context 5: To express agreement of opinion.** We inserted 2 interjections phrase-initially to show Alexa's emphatic agreement to the user's opinion: "Yes!" "High Five!". For example: "[*High Five!... | Yes!...*] We share the same thoughts!"

Overall, our rule-based system resulted to the insertion of interjections in 12-18% of turns in each conversation. We implemented these interjections with a following pause (ranging from 150-300ms), using SSML. Note that 13 unique interjections, of the total 52, were added to very specific utterances (e.g., using "Moo!" with cow jokes) without using this rule-based system (see Appendix B for stimuli and descriptions). All the interjections were rated on two axes by a native English speaker (see Appendix A for full word list and classifications; see Table 5 for an example conversation log from in-lab user tests). **Axis 1** is valence: Positive, neutral, or negative. For example, the interjection "Awesome!" was rated as having a positive valence, while "Darn!" was rated as having a more negative valence. **Axis 2** is the interjection emotional orientation: self- or other-oriented (cf. Brave et al., 2005).

3.2 Methods of Inserting Fillers

We added 9 fillers used in American English (Barbieri, 2008) in the conversational templates: "um", "hmm", "huh", "ah", "uh", "oh", "ooh", "uh huh", "mhm" (see Table 5 for an example conversation log from in-lab user tests). In all cases, we used SSML to add a pause (ranging from 150-200ms) following the filler word and slow the production of the word "so" (80% of original rate), if it occurred before or after the filler to improve naturalness. We added certain subsets of filler words in three specific contexts: to change topics, when retrieving Alexa's backstory, and as an acknowledgment to the user's utterance. Overall, this resulted in fillers added to a total of 7.8-7.9% of total turns.

- **Context 1: To change topic.** We added two fillers, "um" and "uh", either before or after "so" to introduce a new topic. We additionally reduced the rate of "so" (indicated by underlining in the following examples). For example: "[*Um…sooo*, |*Sooo, um…*| *Uh… sooo* | *Sooo… uh,*] I've been meaning to ask you: do you like to play videogames?
- **Context 2: When retrieving Alexa's backstory**. We added six fillers ("mhmm", "hmm", "um", "uh", "oh", and "ooh") at the beginning of the utterance when the user had asked Alexa a question, simulating that Alexa needed time to consider her own experience and/or opinions. For example: "[*Hmm…,* | *Uh…* | *Oh…* | *Ooh…*| *Mhmm…*] I love all animals, but I think my favorite is probably the elephant".
- **Context 3: As an acknowledgment to the user's answer to Alexa's question.** We added the fillers to act as feedback response tokens. Specifically, we added "ah", "oh", "uh huh", "mhmm", "huh", and "ooh" at the beginning of the utterance to show Alexa's acknowledgment of the content provided by the user (e.g., "*Oh…* legos? Interesting choice!"). Note that while these utterances are often used for backchanneling, where one speaker provides verbal feedback while the other continues to hold the floor (e.g., "uh huh" in Pammi, 2012), we do not classify them as such they did not occur during the user's turn. Given the limitations of the text transcripts of the conversations—in the absence of acoustic-phonetic data—we could not implement a real-time backchanneling mechanism

4 Experiment 1: Chatbot User Study

In the current study, we systematically tested the impact of adding interjections and fillers in the Alexa TTS voice in our chatbot (Chen et al., 2018). We hypothesize that in a social dialog system, adding interjections (e.g., "Awesome!") and filler words (e.g., "um") in appropriate locations, with emotional valence consistencies, will improve overall user ratings. This prediction stems from related work conducted in laboratory settings with other types of interlocutors (e.g., robot in Gallé et al., 2017; Marge et al., 2010), with greater expressiveness of the voice relating to positive ratings by users (e.g. Hennig & Chellali, 2012).

4.1 Experimental Conditions

From November 20, 2018 to December 3, 2018 we conducted an ablation study with four possible conditions, varying according to the presence of interjections and fillers (see Table 2). Condition A was filtered to include interjections (and exclude filler words). Condition B was filtered to include filler words and exclude interjections. Condition C included both interjections and fillers, while Condition D excluded both elements. Condition was randomly invoked for each user. During this timeframe, no other code updates were implemented. A total of 5,527 users participated in the study for a total of 5,582 conversations, with 62,130 conversational turns.

Condition	Interjection	Filler	Users	Conversations
A	✓	-	*n*=1511	*n*=1523
B	-	✓	*n*=1183	*n*=1196
C	✓	✓	*n*=1423	*n*=1443
D	-	-	*n*=1410	*n*=1420
		Total	***n=5527***	***n=5582***

Table 2: Experimental conditions & summary statistics

4.2 Statistical Analysis & Results

We modeled user rating (produced at the end of the interaction on a scale from 1-to-5) with a mixed effects linear regression with the *lme4* R package (Bates et al., 2015), with the fixed effect of Condition (A: Interjection only, B: Filler only, C: Interjection and Filler, or D: Neither) and by-user random intercepts. Effects were contrast coded relative to Condition D (baseline condition).

The linear regression model revealed a main effect of Condition on users' ratings, with significantly higher ratings for the three conditions with manipulations (A: Interjection, B: Filler, and C: Interjection & Filler) relative to baseline (see Table 3 and Figure 2 below). The highest rating improvement was observed for Condition C (Interjection & Filler) with an average increase of 0.749.

User Study: Holistic Ratings

	Coef	*SE*	*df*	*t*	*p*
Intercept	2.88	0.025	8.09E+03	114.37	<0.001 ***
A- Interjection only	0.34	0.019	2.99E+04	17.50	<0.001 ***
B- Filler only	0.34	0.019	2.92E+04	18.02	<0.001 ***
C- Interjection + Filler	0.75	0.019	2.98E+04	38.56	<0.001 ***

Num. observations = 31,065 *REML criterion at convergence: 18039.8*

Table 3: Hierarchical linear regression model output: User ratings based on Condition, relative to the baseline condition ("D").

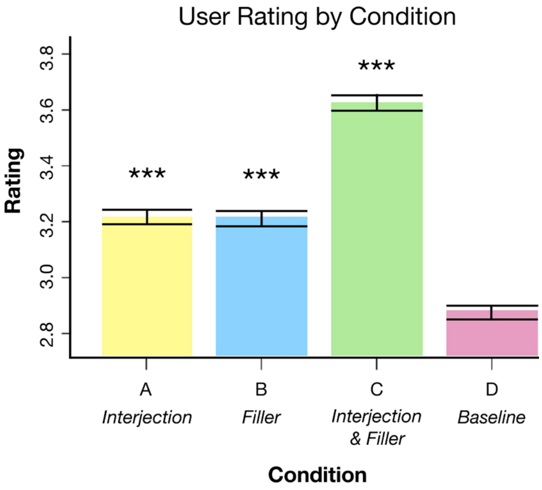


Figure 2: Mean user rating by Condition (error bars represent standard error; asterisks depict significance (p<0.001) relative to the baseline condition, "D")

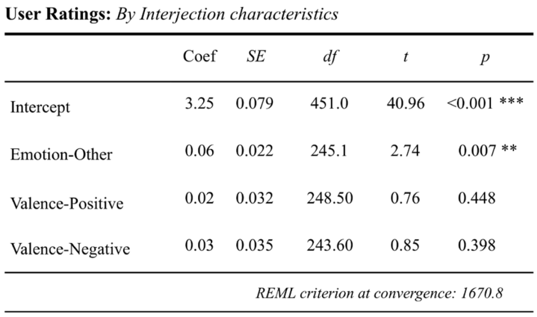
The releveled linear regression model, with Condition C as the reference, tested whether the combined condition (Interjections & Fillers) showed higher ratings relative to the addition of interjections or fillers alone. Results revealed that Condition C indeed showed higher user ratings than Conditions A (Interjections only: β=-0.561, t=-26.16, p<0.001) or B (Filler only: β=-0.326, t=-15.33, p<0.001).

4.3 Interjections Subset Analysis & Results:

We conducted a more fine-grained analysis on the subset of conversations that included the interjections (i.e., Condition A: Interjection, and Condition C: Interjection and filler). In this section, we test whether valence (positive, neutral, negative), emotion orientation (self- versus other), and interjection function (error resolution, change topic, signal interest, etc.) differentially affect user ratings. We predict that more positive interjections, interjections that communicate more other-oriented displays of emotion, and interjections that are used to signal interest (relative to other functions, such as changing topic) will show higher user ratings, in line with prior work (e.g., Bono & Ilies, 2006 Brave et al., 2005; Gibbs & Mueller, 1988).

A mixed effects linear regression model tested the interjection classifications on user's ratings. Fixed effects included Interjection Valence (positive, negative, neutral), Emotion Orientation (self-oriented, other-oriented), and Context (Error resolution, change topic, play, etc). Given the overlap between Emotional Valence and Function (with positive interjections exclusively used to Signal Interest and negative interjections almost always used in Error Resolution, see Appendix A), we tested these two variables in separate models. Random effects included by-user random intercepts.

Model comparisons based on the corrected AIC (Burnham et al., 2011) were conducted with the *MuMIN* R package (Barton, 2017) to test the inclusion of Emotion Valence or Function as main effects, given their colinearity. Model comparisons revealed that the model with the fixed effects of Valence and Emotion Orientation best fit the data (AIC_c=1689.9), relative to the model including Function and Emotion Orientation (AIC_c =1694.78). The retained model output (see Table 4) revealed a main effect of Emotion Orientation, with "other" oriented emotional displays (e.g., "Wow!") associated with higher rating than more self-oriented productions (e.g., "ah"). No differences were observed on the basis of interjection Valence.



User Ratings: *By Interjection characteristics*

	Coef	*SE*	*df*	*t*	*p*
Intercept	3.25	0.079	451.0	40.96	<0.001 ***
Emotion-Other	0.06	0.022	245.1	2.74	0.007 **
Valence-Positive	0.02	0.032	248.50	0.76	0.448
Valence-Negative	0.03	0.035	243.60	0.85	0.398

REML criterion at convergence: 1670.8

Table 4: Mixed effect linear regression output for interjection classification effects on user ratings.

4.4 Qualitative User Study

As part of the Alexa Prize Competition, we additionally recruited users to interact with the system for feedback and bug testing for earlier versions of the dialog system. In September and October 2018, we recorded the interactions of twenty volunteers (12 undergraduates, 8 graduate students). After talking to the socialbot, subjects were asked about their interaction. Several subjects mentioned that they liked the filler words in Alexa's speech as it "sounded like she was actually thinking" or "seemed more realistic". Additionally, we noted that subjects often laughed or smiled when they heard the hyper-expressive interjections while they were part of the conversation (e.g., "Wowza!").

USR:	Hey Alexa, let's chat.
SYS:	Hi, this is an Alexa Prize Socialbot. I'm glad to meet you. How are you doing today?
USER:	I'm okay. How are you?
SYS:	I'm good! So, tell me more about yourself! Do you like movies?
USR:	I do like movies.
SYS:	I'm excited to talk about movies. *Sooo…umm…* **(Filler)**, have you seen any movies lately?
USER:	A Star is Born.
SYS:	*Ah…* **(Filler)** I've seen A Star is Born too! What would you rate this movie on a scale from 1 to 10?
USER:	A ten.
SYS:	*Wow!…* **(Interjection)** A perfect 10! It really is an exceptional movie! Here is something I just learned. Some of the scenes were filmed at the Coachella Festival in 2017…

Table 5: Conversation transcription from an in-lab user test. Interjections and filler words are denoted by *italics* and labeled (original, **annotations**).

5 Experiment 2: Perception Study

While our user study suggests an improvement on the basis of interjections and fillers, it is possible that other factors played a role in the final ratings (e.g., specific phrasing), as well as the co-occurrence of certain interjections, with particular dialog acts (e.g., Alexa using "Darn!" to resolve errors). To disentangle these factors, we conducted a psycholinguistic experiment using a

Qualtrics survey administered through Amazon's Mechanical Turk[1].

5.1 Participants, Stimuli, and Procedure

A total of 85 Amazon Mechanical Turk workers (i.e., "Turkers") participated in the rating task (note that all Turkers had to have an approval rating of 97% or higher and at least 1000 prior HITs). Stimuli consisted of four 3-utterance dialogs between Alexa and a human male talker (a native English speaker, age 29). The conversation topics were based on those discussed in the main social bot (animals and movies), though were novel utterances. The dialogs systematically varied as to whether the expression of emotion in the interjection (if expressed) was self- or other-oriented and had positive or negative valence.

Using the rules for inserting interjections and fillers (see Sections 3.2 and 3.3) and mirroring the Condition structure from Experiment 1, we systematically generated four conditions for each dialog: A) Interjection addition, B) Filler addition, C) Interjection and Filler addition, and D) Baseline. In each of these conditions, we held the human's response exactly the same, as well as all of the wording (for an example, see Table 6). Using a between-subjects design, we additionally tested whether the conversational context for filler words in the first utterance affects their ratings (e.g., following: "So" versus "Yeah, movies can be really fun….So").

CONDITION 1A: Interjection	CONDITION 1B: Filler
Alexa: So, I've been meaning to ask you. What else are you interested in? Do you like animals? **Human**: I love animals! **Alexa**: *Awesome!* I think my favorite animal is the elephant.	**Alexa**: *Sooo, um…* I've been meaning to ask you. What else are you interested in? Do you like animals? **Human**: I love animals! **Alexa**: Awesome. I think my favorite animal is the elephant.

Table 6: Example dialog (Conditions A and B) excerpt used in the perceptual ratings study. Interjections and fillers are annotated in italics.

In the experiment, subjects heard each utterance (randomly presented) and were asked to rate Alexa on several dimensions using a sliding bar (on a scale of 0-to-100): likeability, naturalness, expressiveness, and engagement (e.g., "How engaged does Alexa sound in the conversation?"). Two listening comprehension questions were included to ensure that Turkers were attending to the stimuli and task at hand (e.g., "What was Alexa's favorite animal?" Correct response: An elephant).

5.2 Analysis and Results

Subjects' ratings for each variable were analyzed with separate linear mixed effects models, with a fixed effect of Condition and by-Subject random intercepts. Results showed a main effect of Condition, where introducing interjections significantly increased ratings of engagement ($\beta=6.1$, $t=3.1$, $p<0.01$), naturalness ($\beta=3.7$, $t=3.5$, $p<0.001$), expressiveness ($\beta=9.0$, $t=7.7$, $p<0.001$), and likeability ($\beta=3.4$, $t=3.1$, $p<0.001$) of Alexa. Furthermore, we observed a significant improvement of introducing both interjections and fillers on perceived expressiveness ($\beta=8.1$, $t=7.0$, $p<0.001$). When introducing fillers only, we observed a negative effect on ratings of likeability ($\beta=-2.8$, $t=-2.5$, $p<0.05$) and engagement ($\beta=-2.4$, $t=-2.1$, $p<0.05$) (see Figure 3).

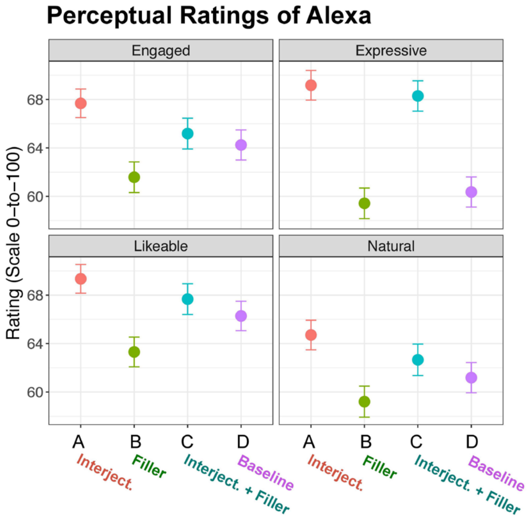


Figure 3: Perceptual Ratings of Alexa for each Condition.

Subset analyses on interjections (Conditions B and C) relative to the baseline were conducted to test for an interaction of Condition*Orientation (self- versus other- oriented emotion) and Condition*Valence (positive, negative, neutral). The models showed significant interactions for both: interjections that were other-oriented ($p<0.001$) and positive in valence ($p<0.001$) showed higher ratings for likeability, engagement,

---

[1] www.MTurk.com

and expressiveness. The subset analysis testing an interaction between the filler condition (relative to baseline) and Conversational Context revealed no effect on ratings.

6 Discussion

This paper combines a large-scale user study with a targeted perceptual ratings experiment to test the effect of adding hyper-expressive interjections (e.g., “Awesome!”) and filler words (e.g., “um”, "um”) in a 2018 Amazon Alexa Prize chatbot. Overall, our user study provides evidence that introducing these discrete expressions of cognitive-emotional expression improves users’ experience talking to a social dialog system; this was evidenced by a higher holistic rating that they provided at the end of the interaction on a scale from 1-to-5. Using both a large sample size and in-situ experiment of an Amazon Alexa Skill, such that users directly engaged with their own devices, is a novel methodology for assessing TTS expressiveness that extends prior in-lab studies on users recruited to engage with the system (e.g., Brave et al., 2005; Cowan et al., 2015; Qvarfordt et al., 2005; Yu et al., 2016).

The cumulative effect of adding interjections and fillers (e.g., in Condition C) suggests that individuals might respond better to dialog systems that use greater TTS dynamism, or *variation*, in the ways in which cognitive-emotional expressiveness is conveyed. These findings can inform theoretical frameworks of computer personification (Nass, 1994; Nass & Moon, 2000); while in a conversation with the system, users appear to be reading the minimal and discrete “human” cognitive-emotional cues generated by the TTS voice – and these effects are additive. Additionally, our results support the classification of fillers and interjections as “socio-affective glue” in developing rapport in human-computer interaction (cf., Sasa & Auberge, 2014).

The facilitatory effect of interjections in the user study was additionally replicated in our perceptual ratings study: we found higher ratings of naturalness, expressiveness, and engagement when Alexa used interjections (e.g., <speechcon>"Awesome!”</spcon>“) versus unmodified productions of the same words (e.g., “Awesome.”). At the same time, we find that introducing filler words improves ratings when the user is directly engaging with the socialbot, but independent raters, who are not directly part of the conversation, give lower ratings for filler words. This suggests that the role of the user in the conversation, as well as the conversational context (as being more socially oriented) may be important considerations in evaluating TTS manipulations to improve cognitive-emotional expressiveness.

Finally, this work has practical applications for other dialog system designers, with the Alexa system (e.g., using Speechcons), but also more broadly. That we see an improvement across thousands of users and unique conversations suggests that inserting interjections and fillers plays a key role in perceptions of social dialog quality. We see the potential to use this expressiveness in other types of interactions, including task-oriented dialog (e.g., in tutoring, counselling sessions, etc.).

7 Conclusion

Overall, we present a methodology for inserting interjections and filler words in a socialbot dialog system and empirical validation of their use in a large-scale user study. In comparison to utterance- or phrase- level prosodic manipulations, these word-level “infusions” of cognitive-emotional expression are easier to implement and appear to improve users’ experience. For one, that we see an improvement in ratings across a large-scale pool of users, each with a unique conversation, suggests that introducing these minimal TTS manipulations in other types of dialog systems may be beneficial. Future work testing the implementation of interjections and/or fillers in task versus non-task-oriented systems can further tease apart their generalizability.

Acknowledgments

We would like to acknowledge the help from Amazon in terms of financial and technical support and Dr. Georgia Zellou for feedback on the project.

References

Akasaki, S., & Kaji, N. (2017). Chat Detection in an Intelligent Assistant: Combining Task-oriented and Non-task-oriented Spoken Dialogue Systems. *ArXiv:1705.00746 [Cs]*. Retrieved from http://arxiv.org/abs/1705.00746.

Amazon. (2018). Speechcon Reference (Interjections): English (US) | Custom Skills. Retrieved from https://developer.amazon.com/docs/custom-skills/speechcon-reference-interjections-english-us.html.

Auberge, V., Sasa, Y., Robert, T., Bonnefond, N., & Meillon, B. (2013). Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot. In *WASSS 2013*. HAL ID: hal-00953780.

Baker, F. S. (2015). Emerging realities of text-to-speech software for nonnative-English-speaking community college Students in the freshman year. *Community College Journal of Research and Practice*, 39(5), 423-441. doi.org/10.1080/10668926.2013.835290.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English 1. *Journal of Sociolinguistics*, 12(1), 58-88. doi.org/10.1111/j.1467-9841.2008.00353.x.

Barton, K. (2017). MuMIn : multi-model inference. R package (Version 1.7.2). Retrieved from ci.nii.ac.jp/naid/10030918982/.

Bates, D., Bolker, B., & Walker, S. (2015). *Fitting Linear Mixed-Effects Models Using lme4*. Retrieved from: doi.org/10.18637/jss.v067.i01.

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer (Version 6.0.37). Retrieved from http://www.praat.org/.

Bono, J. E., & Ilies, R. (2006). Charisma, positive emotions and mood contagion. *The Leadership Quarterly*, *17*(4), 317–334. doi.org/10.1016/j.leaqua.2006.04.008.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2), 161–178. dx.doi.org/10.1016/j.ijhcs.2004.11.002.

Brennan, S. E., & Williams, M. (1995). The feeling of another′ s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, 34(3), 383-398. doi.org/10.1006/jmla.1995.1017.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35. doi.org/10.1007/s00265-010-1029-6.

Chen, C.-Y., Yu, D., Wen, W., Yang, Y. M., Zhang, J., Zhou, M., … Iyer, S. (2018). Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data. *2nd Proceedings of Alexa Prize*. m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/pdf/2018/Gunrock.pdf.

Christenfeld, N. (1995). Does it hurt to say um?. *Journal of Nonverbal Behavior*, *19*(3), 171-186. doi.org/10.1007/BF02175503.

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. doi.org/10.1016/S0010-0277(02)00017-3.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., & Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human−computer dialog. *International Journal of Human-Computer Studies*, *83*, 27-42. doi.org/10.1016/j.ijhcs.2015.05.008.

Gallé, M., Kynev, E., Monet, N., & Legras, C. (2017). Context-aware selection of multi-modal conversational fillers in human-robot dialogs. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 317–322). doi.org/10.1109/ROMAN.2017.8172320.

Gálvez, R. H., Benuš, Š., Gravano, A., & Trnka, M. (2017). Prosodic Facilitation and Interference while Judging on the Veracity of Synthesized Statements. *Proc. Interspeech 2017*, 2331–2335. doi.org/10.21437/Interspeech.2017-453.

Gibbs Jr, R. W., & Mueller, R. A. (1988). Conversational sequences and preference for indirect speech acts. *Discourse Processes*, *11*(1), 101–116. doi.org/10.1080/01638538809544693.

Goble, H., & Edwards, C. (2018). A robot that communicates with vocal fillers has… Uhhh… greater social presence. *Communication Research Reports*, *35*(3), 256-260. doi.org/10.1080/08824096.2018.1447454.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. doi.org/10.1038/466029a.

Hennig, S., & Chellali, R. (2012). Expressive synthetic voices: Considerations for human robot interaction. In *RO-MAN, 2012 IEEE* (pp. 589–595). IEEE. doi.org/10.1109/ROMAN.2012.6343815.

Liu, H., Lin, T., Sun, H., Lin, W., Chang, C.-W., Zhong, T., & Rudnicky, A. (2017). RubyStar: A Non-Task-Oriented Mixture Model Dialog System. *ArXiv:1711.02781 [Cs]*. Retrieved from http://arxiv.org/abs/1711.02781.

Lopatovska, I., & Williams, H. (2018, March). Personification of the Amazon Alexa: BFF or a mindless companion. *In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 265-268). ACM. doi.org/10.1145/3176349.3176868.

Marge, M., Miranda, J., Black, A. W., & Rudnicky, A. I. (2010). Towards improving the naturalness of social conversations with dialog systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialog* (pp. 91–94). Association for Computational Linguistics. http://aclweb.org/anthology/W10-4318.

Montero, J. M., Gutierrez-Arriola, J. M., Palazuelos, S., Enriquez, E., Aguilera, S., & Pardo, J. M.

(1998). Emotional speech synthesis: From speech database to TTS. In *Fifth International Conference on Spoken Language Processing*. https://www.isca-speech.org/archive/archive_papers/icslp_1998/i98_1037.pdf.

Mustafa, M. B., Ainon, R. N., Zainuddin, R., Don, Z. M., Knowles, G., & Mokhtar, S. (2010). Prosodic Analysis And Modelling For Malay Emotional Speech Synthesis. *Malaysian Journal of Computer Science*, *23*(2), 102-110. https://ejournal.um.edu.my/index.php/MJCS/article/view/6399.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3), 171. doi.org/10.1037/1076-898X.7.3.171.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. doi.org/10.1111/0022-4537.00153.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 72–78). ACM. doi.org/10.1145/191666.191703.

Németh, G., Fék, M., & Csapó, T. G. (2007). Increasing prosodic variability of text-to-speech synthesizers. In *Eighth Annual Conference of the International Speech Communication Association*. https://www.isca-speech.org/archive/interspeech_2007/i07_0474.html.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, *17*(11), 776. doi.org/10.1037/h0043424.

Pammi, S. C. (2012). Synthesis of listener vocalizations: towards interactive speech synthesis. PhD thesis, Naturwissenschaftlich-Technische Fakultät I, Universität des Saarlandes, Saarbrücken, Germany. doi.org/10.22028/D291-26277.

Pfeifer, L. M., & Bickmore, T. (2009, September). Should agents speak like, um, humans? The use of conversational fillers by virtual agents. In *International Workshop on Intelligent Virtual Agents* (pp. 460-466). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-04380-2_50.

Qvarfordt, P., Beymer, D., & Zhai, S. (2005, September). Realtourist–a study of augmenting human-human and human-computer dialog with eye-gaze overlay. In *IFIP Conference on Human-Computer Interaction* (pp. 767-780). Springer, Berlin, Heidelberg. doi.org/10.1007/11555261_61.

Read, R., & Belpaeme, T. (2012, March). How to use non-linguistic utterances to convey emotion in child-robot interaction. In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (pp. 219-220). ACM. doi.org/10.1145/2157689.2157764.

Sasa, Y., & Auberge, V. (2014). Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the" socio-affective glue". In *SpeechProsody 2014*. hal.inria.fr/hal-00953723.

Sasa, Y., & Aubergé, V. (2017, October). SASI: perspectives for a socio-affectively intelligent HRI dialog system. In *1st Workshop on "Behavior, Emotion and Representation: Building Blocks of Interaction*". hal.inria.fr/hal-01615470/.

Schmitz, M., Krüger, A., & Schmidt, S. (2007). Modelling personality in voices of talking products through prosodic parameters. In *Proceedings of the 12th international conference on Intelligent user interfaces* (pp. 313–316). ACM. doi.org/10.1145/1216295.1216355.

Schröder, M. (2007, September). Interpolating expressions in unit selection. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 718-720). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-540-74889-2_66.

Skantze, G., House, D., & Edlund, J. (2006). User responses to prosodic variation in fragmentary grounding utterances in dialog. In *Ninth International Conference on Spoken Language Processing*. isca-speech.org/archive/interspeech_2006/i06_1229.html.

Syrdal, A. K., Conkie, A., Kim, Y. J., & Beutnagel, M. C. (2010). Speech acts and dialog TTS. In *Seventh ISCA Workshop on Speech Synthesis*.

Tao, J., & Tan, T. (2005). Affective computing: A review. In *International Conference on Affective computing and intelligent interaction* (pp. 981–995). Springer. doi.org/10.1007/11573548_125.

Tokuhisa, R., & Terashima, R. (2009, July). Relationship between utterances and enthusiasm in non-task-oriented conversational dialogue. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (pp. 161-167). https://dl.acm.org/citation.cfm?id=1654628.

Ward, N. (2004). Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody 2004, International Conference.*

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8), 1177-1207. doi.org/10.1016/S0378-2166(99)00109-5.

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals.*

Cambridge University Press. doi.org/10.1017/CBO9780511521256.

Wigdor, N., de Greeff, J., Looije, R., & Neerincx, M. A. (2016, August). How to improve human-robot interaction with Conversational Fillers. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN) (pp. 219-224). IEEE. doi.org/10.1109/ROMAN.2016.7745134.

Yang, L. C. (2010). Meaning and Context: Prosodic Variation of Interjections. In *Conversational Speech. In Speech Prosody 2010-Fifth International Conference.* https://www.isca-speech.org/archive/sp2010/papers/sp10_380.pdf

Yu, Z., Nicolich-Henkin, L., Black, A. W., & Rudnicky, A. (2016). A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialog* (pp. 55-63). http://www.aclweb.org/anthology/W16-3608.

Appendix A. Interjection (Speechcon) Classifications

Function	**Valence**			**Emotional Orientation**
	Positive	***Neutral***	***Negative***	
Signal Interest	Great! Awesome! Fantastic! Super! Ooh la la! Wowza!	Wow! Cool! Interesting! Fancy that! No way!		***Other***
	Aha!			***Self***
Resolve error		Jiminy cricket! Whoops a daisy!	Darn! Shoot! Yikes! Oh boy! Oh dear! Oh brother! Ouch! Tsk tsk!	***Other***
		Ruh roh!	Baa! Oof! Uh oh!	***Self***
Accept request		Okey dokey! Righto! As you wish! You bet!		***Other***
Change topic	Spoiler alert* *(only with disclosure)*	Ahem! All righty!		***Other***
		Ooh! Ah!		***Self***
Express agreement of opinion	High five! Yes!			***Other***
Joke *(phrase-finally)*	Just kidding!*		Wah wah* Neener neener!*	***Other***
			D'oh!*	***Self***
Joke *(specific context)*		Woof!^ Moo!^ Meow!^ Kerplop!^ Honk!^		***Other***
Other context and module-specific interjection	Yum!^ Aww!^			
Response to user after telling a joke	Tee hee!^			

Table A1. Interjections that are only used in very constrained contexts are annotated with an asterisk (*); those that are only used in one, specifically specified sentence are annotated with a carat (^).

Appendix B. Methods of Inserting Sentiment-Specific Interjections

We additionally added 10 interjections in sentiment-specific utterances. These were not interchangeable (unlike Contexts 1-4 described in Section 3.3). We used the interjection, “Spoiler Alert!” to change the topic by leading in to a disclosure by Alexa (see example A below). We introduced 2 interjections as a response to humor, that occurred after a response to a joke. “Tee hee!” and “Woohoo!” (see examples B and C). We implemented “Yum!” specifically in the food module, in response to the user’s favorite food (see example D). Similarly, we added the interjection, “Aww!” as a response to the user disclosing information about their pet in the animal module and “Woof!” and “Meow!” to respond if they indicated they liked dogs or cats, respectively (see examples E-G).

a) *Spoiler alert!...* Did you know? I am definitely more of a dog person than a cat person. How about you? Do you like animals?
b) *Woohoo!...* I’m glad you get my awesome humor.
c) *Tee hee!...* I LOL’d at that as well | If I could giggle I would.
d) *Yum!...* That sounds really delicious.
e) *Woof!...* I love dogs.
f) *Meow!...* I love cats.
g) *Aww!...* That’s so cute.

Table B1: Examples of sentiment-specific interjections (denoted in italics).

We added several interjections (e.g., “Moo!”, “Honk!”, “Woof!”, “Just Kidding!”) at the end of utterances to complement jokes and express playfulness (see examples G-K in Table B-2).

h) What do you call a cow during an earthquake? … A milkshake. ... *Moo!*
i) What do you call blueberries playing the guitar?... A jam session. ... *Wah wah!*
j) What did the traffic light say to the car? …Don't look! I'm about to change... *Honk!*
k) Why wouldn't the shrimp share his treasure?... Because he was a little shellfish... *Neener neener!*
l) Yeah, wouldn't it be (interesting|weird) if I could poop? ... *Kerplop!*

Table B2: Examples of sentiment-specific interjections (denoted in italics) added phrase-finally

Additionally, we added “Kerplop!” in our response if a user asked Alexa if she “poops” (a frequent question in the user studies) (see Table B2 above).