Exploring Humor as a Repair Strategy During Communication Breakdowns with Voice Assistants

Mikkel Clausen
mikkelclausen95@gmail.com
Department of Computer Science, Aalborg University
Aalborg, Denmark

Mikkel Peter Kyhn
mikkel.peter.kyhn@gmail.com
Department of Computer Science, Aalborg University
Aalborg, Denmark

Eleftherios Papachristos
eleftherios.papachristos@ntnu.no
Department of Design, Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway

Timothy Merritt
merritt@cs.aau.dk
Department of Computer Science, Aalborg University
Aalborg, Denmark

ABSTRACT

Voice assistants are becoming increasingly useful and support realistic conversations, yet communication breakdowns occur. We investigate the use of humor as a repair strategy in an experiment where the voice assistant makes a mistake and then utilizes one of four humorous personalities to repair the breakdown in the conversation. We conducted a study with 30 participants, each of whom took the Humor Style Questionnaire (HSQ) to understand their predisposition to humor type, and then engaged in conversation with each of the four humorous personalities and one that was designed to give neutral repair responses (non-humorous). Aggressive personalities were rated as the funniest, yet there was no clear connection between the participant's humor style and their preferred voice assistant personality. While humorous responses were successful in repairing communication breakdowns, participants overall preferred non-humorous responses. This research provides insight into the role of humor in communication breakdown repair with voice assistants.

CCS CONCEPTS

• **Human-centered computing → Personal digital assistants**.

KEYWORDS

conversational agents, funny content, humorous, humor, intelligent personal assistants, digital assistant, voice user interface, intelligent personal assistant

1 INTRODUCTION

Voice assistants have become a familiar part of everyday life in the home with retail products such as the Amazon Echo or the Google Home. Advances in natural language processing and machine learning have enabled voice assistants to communicate verbally, gather information, control IT equipment, and take on more complex tasks [11, 45]. While the current context of voice assistant usage is primarily task-oriented [23], it is expected that as the technology matures, it will adopt more anthropomorphic features [29, 37] allowing for more organic human-like conversations [7]. Even though voice assistants are able to engage in effective conversations, just as with human communication partners, *communication breakdowns* occur, such as missed inputs and misinterpretations [3]. Research on communication breakdowns suggests that these misunderstandings and small failures are inevitable therefore, it is more important to consider how to recover when breakdowns happen [3, 9, 10, 19, 42].

In order to prevent user frustration, when a voice assistants is unable to respond or needs further clarification, it should acknowledge and notify the user that an error has occurred and that new input is needed. The error notification is through a verbal response and it is common to adopt anthropomorphic cues through natural phrasing and a relatable voice [28]. Typically, the responses might explain in a synthetic human voice, "I did not hear you properly", or "I am not sure about that", or "this is what I found on the internet." These neutral responses may be effective, but they often lack the character and tone that we might expect from a familiar friend or acquaintance. Prior research on voice assistants suggests that humorous personalities could be used to help in recovering from communication breakdowns [17]. Humor is commonly used during conversation and throughout many facets of daily life to smooth over difficult situations [12]. Humor has been shown to be useful in conversations serving various purposes including uniting communicators and supporting clarification [27]. Considering the complexity of humor, in this paper we ask, "Which humor type/strategy is most preferred by users when employed by VUIs to recover from communication breakdowns?" In this paper, we present an exploratory study in which 30 participants were asked to interact with four voice assistants that used distinct humor styles to recover from communication breakdowns and one neutral voice assistants that did not use humor.

This paper is organized as follows: first, we review the related research informing the design of voice assistants and the role of humor in conversation, then, we present the details of the user study, thirdly, we present the results, and finally, we discuss the findings and map out important future work.

2 RELATED WORK

2.1 Humor theory

Humor is a cognitive process based on social contextual stimulus, which triggers an emotional response in humans [22]. Many complementary theories about humor have been formulated over the years [6], such as Relief Theory [6, 27], which sees humor as a physiological release of emotional tension or the incongruity theory [4], which focuses on cognition and views humor as a response

to violations of expected patterns. Martin et al. [21] investigated interpersonal differences in humor and identified four humor styles, which describe in what context humor is used, who is the target of the humor, and the state of mind of the speaker. The four humor styles are presented in a 2x2 matrix, formed around the speaker's state of mind, positive or negative, and whom the humor is directed at, self-oriented or other-oriented (see Table1) [21].

- Affiliative: This humor type relates to strengthening bonds between speaker and listener. It is friendly and essentially as non-hostile as humor can be. As seen in the matrix, it is oriented positively toward others and has a positive outlook toward the listener. Wordplay, puns, and witty banter are typical of the affiliative style.
- Aggressive: This humor is known for ridicule, teasing, sarcasm, and other personal attacks masked as humor and is about negative emotions pointed toward the listener. This humor type includes ideas and themes that may not be socially acceptable but can be played off as sarcasm, absurdity, or provocation. The listener can be the joke's target, but groups outside the speaker and listener can also be the target of ridicule.
- Self-enhancing: This humor type is about promoting and being positive about oneself and a positive outlook on life. It is often used to mask potential negativity towards the speaker by subverting the target of ridicule as untouched by the criticism. It is positively self-oriented and may be recognized as a self-defence mechanism, as the speaker tries not to lose face.
- Self-defeating: This humor type is about the speaker being funny at the speaker's own expense. This humor type can involve personal anecdotes in which the speaker does not appear on top, socially uncomfortable, or unacceptable situations. It is about negative humor directed towards oneself and it is often self-deprecating. It invites the listener to show empathy towards the listener.

	Other-oriented	**Self-Oriented**
Positive Styles	Affiliative	Self-Enhancing
Negative Styles	Aggressive	Self-Defeating

Table 1: The four humor styles as represented by [21]

The four humor types provide an overview of how humor is formed and how it can be used in different social contexts. Most people typically use one of those types when making humorous remarks. However, there can be a predisposition towards using one type of humor more frequently, which corresponds to a person's humor style personality. This predisposition has led Martin et al. to create the Humor Style Questionnaire (HSQ), which determines the person's humor style from personality questions. The HSQ is one of the most established and recognized tools for determining an individual's humor style and has been translated into many different languages and used in numerous countries [36].

2.2 Smart speaker interactions

Nijholt et al. [30] present a study showing that combining voice pitch, language cues, and humor benefited the quality of the social interactions between humans and social robots. These findings resulted from an experiment using a voice interface with varying features, such as pitch change and pause, to simulate vocal cues. The implications of this paper show how pitch and vocal cues benefit the quality and perceived user preference.

Beneteau et al. [3] studied communication breakdowns with the Amazon smart speaker Alexa in the context of families. They recorded 59 communication breakdowns from 10 families with children over four weeks to classify the families' repair strategies, support strategies, and voice assistant error signals. This study identified what repair strategies users employed when encountering a communication breakdown with Alexa. From the 59 communication breakdowns, the families' responses were formed into six different repair strategies, with three different communication breakdown signals that indicate a breakdown has occured. Acting on Misunderstanding (AoM) involves a response based on misheard or misunderstood input, such as giving a weather report when the user asked for a song to be played. Neutral Clarification Response (NR) is an explicit response indicating that the interaction or intent is unclear, such as responding with, "Sorry, I don't know that." Specific Clarification Response (SR) occurs when the VA responds by providing specific information and request clarification, such as clarifiying whether "10 o'clock" is in the morning or evening.

In Lopatovska et al. [16], humorous responses by voice assistants were identified and ranked through a week-long online diary. The participants rated what kind of humorous utterances were found the funniest. The results of this study showed that "canned humor" utterances, including wordplay, puns, and pop-cultural jokes were the most frequently occurring humorous responses in voice assistants. Further, some unintentional situations created by communication breakdown were found somewhat humorous to the participants, despite the fact that they were unintentional. This study's findings form implications for what kinds of humorous interactions users find regarding interactions with voice assistants. In another study by Shani et al., [39], unintentional humor in voice assistants is investigated and analysed through the filter of classical humor theories such as superiority, relief, and incongruity [22, 27].

2.3 Personalities for Voice Assistants

Most research on voice assistants and personalities does not exclusively research humor but often compares different types of personalities to each other. The personality trait of being apologetic during communication breakdown is often compared to that of being humorous. In a study by Mahmood et al. [19], test participants joined an online experiment featuring interactive storyboards, in which five voice assistants' personalities were presented, one control (neutral) personality, and four related to sincerity of the apology (serious/casual) and blame assignment (taking blame/shifting blame). This experiment was designed so that errors would occur, and participants were asked to rate the voice assistants on various

dimensions after their interaction. Results showed that the participants ranked the neutral and sincere/accepting personalities higher than those who were casual or shifting blame. However, the neutral personality was ranked lower in its perceived ability to acknowledge mistakes than the sincere/accepting personality.

In the study by Olafsson et al. [31], two voice assistants were developed to motivate health behavior change in 15 participants. The results showed that affiliative humor was most effective in positively motivating behavioral change in the participants. This study shows that conversational agents' humorous personalities can affect users positively. Another study investigated the effect of a humorous conversational agent in an online learning environment [8]. The humor personalities of affiliative, self-defeating, and neutral were used to investigate the effect on learning experience and outcomes. The results showed that the two humorous personalities positively affected the learning experience.

In this section, we presented a subset of research on humor personality of voice assistants and how they relate to humor theories. It was however not always possible to identify which theories formed the basis of some of the previously mentioned research, as for example [10, 19]. It should, however, be noted that the humor style most frequently encountered in literature is affiliative. A notable exception is the study by Ceha et al. [8], which explores affiliative and self-defeating humor styles. Considering that humor is a multidimensional construct, as described by Martin et al. [21], we aim to explore the full range of humor types to investigate potential benefits in communication breakdown scenarios.

3 METHOD

We investigate the effects of the four humor types that were developed by Martin et al.[21] and the repair of breakdowns in conversation with a voice assistant. We follow a similar experimental design as in the study described in Mahmood et al. [19]. We formulated the following hypotheses:

- **H1:** Humor repair strategies are preferred over non-humorous strategies.
- **H2:** Participants ranking high in one of the humor styles will prefer a voice assistant with a corresponding humor personality.
- **H3:** Humor style personalities will affect the participant's perceptions of the voice assistant's intelligence, satisfaction, and willingness to use.

3.1 Experimental design, study design, and conditions

We conducted the experiment in a laboratory using a Wizard of Oz method [14] and an Amazon Alexa through text-to-speech using an SSML skill [43] for pitch change and pause to increase the quality of the conversation with the users as mentioned by Nijholt et al., [30]. Participants interacted with five personalities, four that matched the humor styles by Martin et al. [21] and one with a non-humorous personality as a neutral control condition.

The Amazon Alexa smart speaker was kept to its default preselected voice settings. The task was constructed around using Alexa as an interface for playing music and podcasts, as this is one of the most common uses for Alexa [38]. The overall experiment would be centered around playing music with four specific tasks. Music was played through the Spotify skill, with all of the commands being examples of authentic use of Alexa as an interface for playing music. The four tasks were:

(1) *"Alexa, play 'Limit To Your Love', by James Blake on Spotify."*
(2) *"Alexa, play Rock music."*
(3) *"Alexa, play from playlist, 'Sommer i Tyrol'."*
(4) *"Alexa, play the podcast 'Hello Internet'."*

To simulate a natural communication breakdown, one of the researchers would control when a communication breakdown would occur. Among the four tasks, breakdowns were triggered in two of the four tasks chosen at random. These breakdowns would be announced by a *neutral clarification response* (CR) communication breakdown signal [3]. During the communication breakdowns, the voice assistants would misinterpret the participants and reply with a CR response, calling for a repetition repair strategy [3]. The participants would then repeat the task, and the voice assistants would succeed in hearing the participants correctly, subsequently completing the task. All utterances done by the voice assistant, both errors and successful interactions, were controlled and prompted by one of the researchers, using Alexa skills and routines to simulate a believable conversation between the participants and the voice assistant. Each participant engaged in a total of 20 interactions, consisting of four tasks for each of the five personality types. The presentation order of the assistants was balanced and randomized using a Latin Square design.

3.2 Communication setup

The voice assistant used phrasing formed explicitly around the personality assigned to that humor style condition. The voice assistant was designed with two communication interactions: success and breakdown. Both communication success and breakdown signals were formed based on the intended humor personality. The communication successful sentences consisted of two parts:

- Personality: The voice assistant would highlight and showcase its personality type by adding phrasing and word usage matching the personality archetype. The phrasing would help the participants to identify what kind of personality they were communicating with. These utterances were not designed to be funny.
- Action response: The voice assistant would follow up on the personality phrases, specifying what action the voice assistant was about to commit (see figure1).

As was the case with successful communication, the breakdowns were specifically phrased to abide by the humor styles described by Martin et al. [21]. These sentences were purposefully constructed to be humorous and fit one of the humor styles.

The communication breakdowns consisted of two parts:

- *Neutral clarification response* (CR) [3]: This part informs the user of a communication breakdown. Participants would be informed that the voice assistant could not hear their commands correctly and provide the call to action, instructing the participants using the repair strategy, 'repetition', to repeat their last voice command.

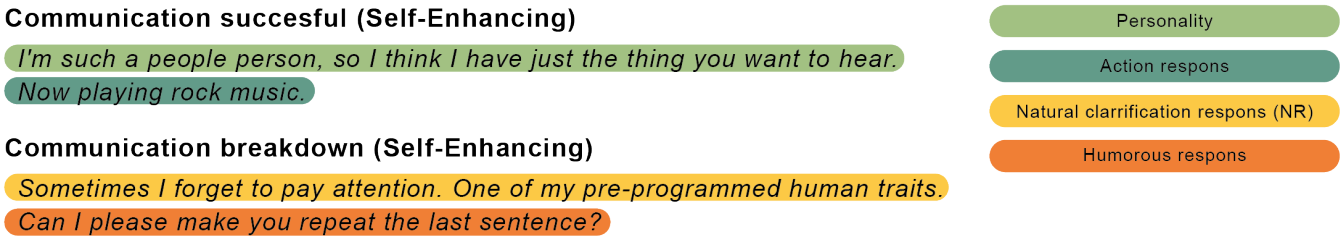

Figure 1: Examples of self-enhancing personality response: Personality, Action response, Natural clarification response, and Humorous response.

- Humorous response: The communication breakdowns were presented with a humorous phrasing, matching the current humor of the corresponding personality type.

The humorous sentences were created in an iterative process in which we systematically generated multiple utterances and evaluated them based on how clearly, they represented the intended personality type. The first step was exploratory, in which two of the authors collected inspiration from the literature about humorous interactions with PAs and popular culture sources such as movies and games involving humorous robots and AI. The next step was generative, in which we created as many humorous sentences as possible that could be probable responses to the four tasks. This was followed by an evaluative phase in which we went through each sentence and explained the humorous intent behind it and how much it corresponded to any of the humour personalities. After undergoing multiple rounds of iteration, we ultimately agreed on which sentences to eliminate. Our evaluation criteria were based on their congruence with the intended personalities and their perceived level of humor. We ended up with five sentences for each personality, and in the final step, we chose the most humorous sentences out of those based on group consensus.

3.3 Measures

Our task design followed a similar design of error responses in previous studies [10, 19]. In the next subsections, we present the metrics we used to measure affinity to humor styles, service recovery satisfaction, perceived intelligence, likeability, and willingness to use the voice assistant.

3.3.1 Subjective measures on humor style and perceptions of voice assistants.

- **Humor Style Questionnaire**. Developed by Martin et al. [21], the Humor Style Questionnaire (HSQ) is among the most prominent self-report scales in the psychology of humor [41], with the questionnaire being translated into multiple languages [36]. The four humor styles presented in the questionnaire are measured with a 32-item self-report Likert scale ranging from 1 (totally disagree) to 7 (totally agree). Each style is assessed with eight items.
- **Service recovery satisfaction**. Following prior research [19], we used two 5-point Likert scale questions ("I am happy with how the error was handled" and "In my opinion, the AI assistant provided a satisfactory response to the error") to measure service recovery satisfaction.
- **Perceived intelligence**. Based on previous research [19, 37], we used Godspeed four items questionnaire[2] to measure the perceived intelligence of the voice assistant on a 5-point semantic rating scale (Cronbach's alpha = .90). We asked the participants to rate their impression of the agent on these dimensions: 1) Incompetent – Competent, 2) Ignorant – Knowledgeable, 3) Irresponsible – Responsible, 4) Unintelligent - Intelligent, and 5) Foolish – Intelligent.
- **Likeability**. Following previous research [19], we used Godspeed three-item questionnaire [2] to measure likeability on a 5-point semantic rating scale. We asked the participants to rate their impressions on the dimensions: 1) Dislike – Like, 2) Unfriendly – Friendly, 3) Unkind - Kind, 4) Unpleasant - Pleasant, and 5) Awful – Nice.

3.4 Participants

A total of 30 participants (17 males, 13 females) were recruited for the study through social media and the authors' personal network. The participants were 19 to 29 (M = 24.1, SD = 2.56) years old. Most participants (n=25) were students at the computer science department of Aalborg University in Denmark and were of Danish nationality. The majority of the participants were recruited through social media and asked to book a time through online appointment software. One of the inclusion criteria was to be proficient in English–it is typical for Danish students we made it clear in the study invitation that the study would be conducted entirely in English. [1] None of the participants reported having any issues understanding the personal assistant or the humorous responses.

3.5 Procedure

Before the experiment, participants were asked to complete the Humor style Questionnaire[21]. The experimental procedure of the study consisted of four phases:

(1) *Introduction and consent.* Before the start of the experiment, participants were introduced to the study, and were asked to sign a consent form.
(2) *Experimental task.* Participants were randomly assigned one of the conditions. The order in which they interacted with the five personalities was randomized and counterbalanced using a Latin Square to limit order effects on familiarity and fatigue.

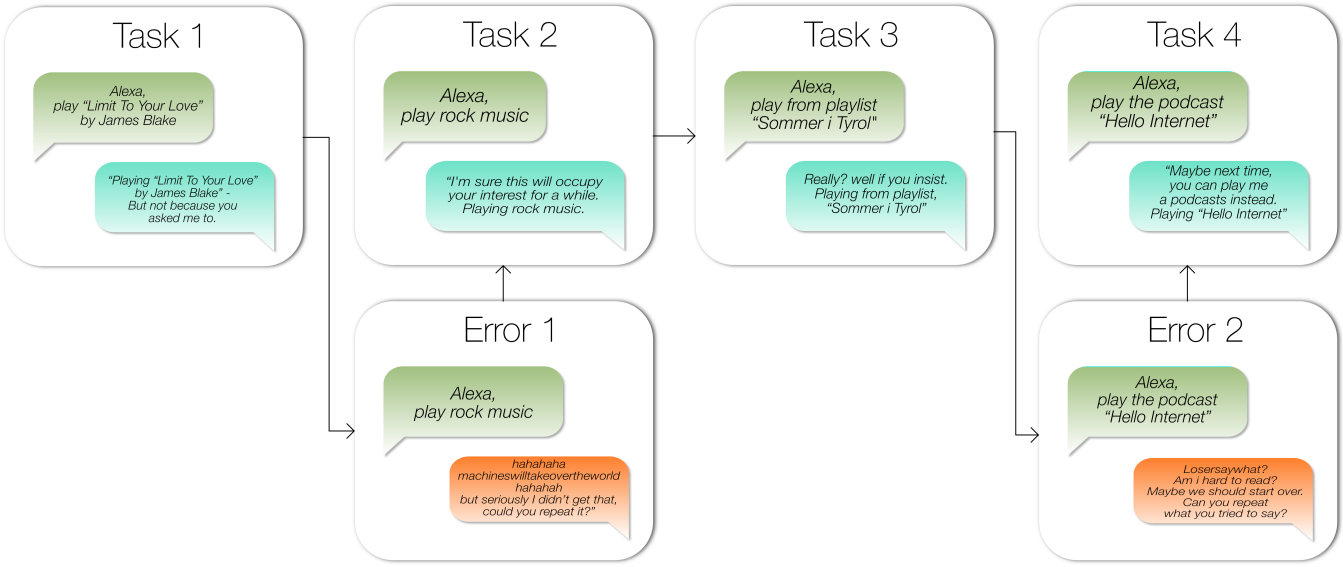

Figure 2: A visual representation of the dialog tree featured in the experiment for the 'aggressive' humor personality.

(3) *Survey.* After each interaction, participants completed a paper questionnaire about their perceptions of the voice assistant. The facilitator also verbally inquired about their perception of the personality they interacted with and wrote down their reply. Afterward, they continued on to the next condition and repeated phases three and four.
(4) *Post study questions.* After completing all the conditions, the facilitator asked two questions on paper: "Which of the five personalities did you find the funniest?" and "Which of the five personalities are you most likely to use in the future?"

4 RESULTS

Our study resulted in 150 evaluations considering that each of the 30 participants interacted with all five personality types. We performed one-way repeated measure ANOVA analysis to investigate the effects of humor personality on the various independent variables we presented in section 3.3. All post hoc pairwise comparisons were conducted using Tukey's HSD test. We followed Cohen's guidelines on reporting effect size [32]. Figure 3 visualizes our main results based on our quantitative data. An affinity diagram was created with qualitative data collected during the post-study interview. This allowed us to identify themes and patterns that emerged from the interview data following a bottom-up approach.

4.1 Service recovery satisfaction

A one-way repeated measures ANOVA was conducted to compare the service recovery satisfaction scores of the agents' five personalities: affiliative, self-enhancing, aggressive, self-defeating, and neutral. The ANOVA revealed a significant main effect of humor personality type on service recovery satisfaction ($F(4,145) = 7.789$, $p = <.001$, $\eta_p^2 = 0.142$). Pairwise comparisons using Tukey's HSD test showed that participants were more satisfied with service recovery by the agent with the neutral personality ($M = 4.317$, $SD = 0.895$) compared to the aggressive humor ($M = 3.317$, $SD = 1.309$) and the self-defeating humor personalities ($M = 2.950$, $SD = 1.124$).

4.2 Perceived intelligence

A one-way repeated measures ANOVA was conducted to compare scores on the perceived intelligence of the five personalities in the agents. The results revealed a significant main effect of humor personality type on perceived intelligence ($F(4,145) = 9.390$, $p = <.001$, $\eta_p^2 = 0.206$). Pairwise comparisons using Tukey's HSD test revealed that the self-defeating humor personality ($M = 2.947$, $SD = 0.995$) was perceived as less intelligent compared to the affiliative ($M = 3.660$, $SD = 0.759$), $p = 0.021$, self-enhancing ($M = 3.533$, $SD = 1.004$), $p = 0.049$, and the neutral humor personality ($M = 4.067$, $SD = 0.583$), $p < 0.001$. Furthermore, the neutral personality style ($M = 4.067$, $SD = 0.583$) was perceived as more intelligent than the aggressive ($M = 2.987$, $SD = 0.830$), $p < 0.001$ and the self-defeating ($M = 2.947$, $SD = 0.995$), $p < 0.001$, humor personalities.

4.3 Likeability

Similar to perceived intelligence we performed a one-way repeated measures ANOVA and we found a significant main effect of humor personality type to likeability ($F(4,145) = 11.978$, $p = <.001$, $\eta_p^2 = 0.248$). Pairwise comparisons using Tukey's HSD test suggest the aggressive humor personality ($M = 2.760$, $SD = 1.037$) was perceived as less likeable by participants compared to the affiliative ($M = 4.080$, $SD = 0.707$), $p < 0.001$, the self-enhancing ($M = 3.887$, $SD = 0.919$), $p < 0.001$, self-defeating ($M = 3.520$, $SD = 0.725$) $p = 0.005$, and the neutral humor personality ($M = 3.893$, $SD = 0.717$), $p < 0.001$.

4.4 Humor styles and agent preferences

As shown in Figure 4, 19 participants rated the voice assistant with the aggressive humor personality as the 'most humorous'.

Table 2: Means and standard deviations for the four Humor Styles Questionnaire scales for all participants

	Mean	Std. Deviation
affiliative humor	46.07	5.795
self-enhancing humor	37.63	6.965
aggressive humor	28.37	7.365
self-defeating humor	30.80	7.645

Meanwhile, in their preference for 'most likely to use in the future', 13 participants chose the neutral, and 11 chose the affiliative humor personality.

In our analysis of the HSQ (see Table 2), the majority of participants had an affiliative humor style with a total number of 22 participants, six participants had a self-enhancing humor style, and two participants had an equal score in affiliative and self-enhancing. In contrast, no participants had a clear aggressive or self-defeating humor style. Table 2 summarizes the mean and standard deviation for the four HSQ scales for all participants.

5 DISCUSSION

Effective voice assistant error mitigation is critical for retaining user satisfaction, building a positive relationship, and increasing system usage. This study investigated how participants' perceptions of voice assistants and satisfaction with service recovery differed depending on their sense of humor. We now discuss the study results in relation to the hypotheses, the limitations of our work, and the implications for future research.

5.1 Preference towards non-humorous response

Our hypothesis H1 questions that *any humor of the four humor repair strategies is preferred over no strategy*. From our results (see figure 3), it was found that a neutral personality is preferred over an aggressive or self-defeating one regarding service recovery satisfaction and perceived intelligence. There were no significant differences between neutral, affiliative, and self-enhancing. For perceived likeability, the aggressive humor personality was perceived as less likable compared to the other personalities, and there was no significant difference between affiliative, self-defeating, and neutral personalities. This outcome is consistent with earlier studies on affiliative and self-defeating humor personalities [8].

Our results are consistent with previous research about communication mistakes [19] in which participants ranked personalities during communication breakdowns. They found that a neutral personality ranked second highest out of five in service recovery satisfaction, perceived intelligence, and likability. The preferred response was a 'serious and accepting' response. That study focused not exclusively on humor but on apology for the mistakes during communication breakdown, and humor was featured as a casual personality trait. This suggests a preference for neutral voice assistant personalities. It should be noted that experience with voice assistants was not disclosed in the paper by Mahmood et al. [19]. Our data suggest that there may be a correlation between experience with the voice assistant and the desire for less personality during voice assistant interactions. Research on teammates in games suggests that AI is perceived as instrumental and that engagement socially with AI results in differences in judgment emotions [26].

Additionally, participants explained why they preferred the neutral for continued usage, commenting: P13 - *"Just for daily use [...] compared to the second one (aggressive) for example it was very talkative, this one (neutral) was quicker... yeah it's a safe bet"* and P20 - *"I would probably be annoyed if I used it over a longer period"*.

5.2 Unexpected or inappropriate responses

Unexpectedly participants rated the aggressive humor personality as the most humorous and gave comments; P8 - *"I liked it better than the others, it mocks you"*, P10 - *"It is the funniest that have been so far, it makes fun of you"*, and P15 - *"She is quite rude, but... also entertaining"*. This contradicts our hypothesis H2, which questions if *participants ranking high in one of the humor styles will prefer the same corresponding voice assistant humor style personality and rate it as most humorous*. As shown in Figure 4 the aggressive humor personality was rated the most humorous. In contrast, in Table 2 we can see that the majority of participants are highest in the affiliative humor style. A reason for this could be that the study was conducted with Danish participants and Danish humor is known for being sarcastic, aggressive, and without limits [18]. This may have influenced the participants to have a greater tendency toward the aggressive humor style, which may affect the data from the experiment and their verdict on the funniest humor personality. Another reason could be the incongruity theory as explained by Meyers' three humor theories [27, 39]. Participants could be surprised by the aggressive personality's unexpected or inappropriate responses and find it funny, which aligns with prior research on humorous interactions using a voice assistant [16].

5.3 Humorous personality preference

From our hypothesis H3 *humor style personalities presented by the voice assistant, will affect the participant's perceived intelligence, satisfaction, and willingness to use*, as seen in our results on the four humor personalities, the aggressive and self-defeating humor personality is the least preferred by participants in regards to our components. Furthermore, the affiliative and self-enhancing humor personalities, while not having a significant difference on the neutral personality, are still in line with each other regarding our component on likeability. It can be argued that an affiliative or self-enhancing humor personality can be used. This can be further substantiated by the participants' comments on the neutral personality: P10 - *"It's professional, but it does not have much of a personality"*, P15 - *"Simple, it cut all the fat away. More focused on completing the task, not much of a personality though"*, P20 - *"It did what it should, but it sounds quite dead"*, and P26 - *"Didn't feel like there was much of a personality, felt more like I talked to a robot"*. Compared to the paper by Ceha et al. [8], it is mentioned that they constructed two personalities (affiliative and self-defeating) with Martins' four humor personality types. This was due to the humor types being described as either conductive (affiliative, self-defeating) or detrimental (aggressive, self-enhancing) in building

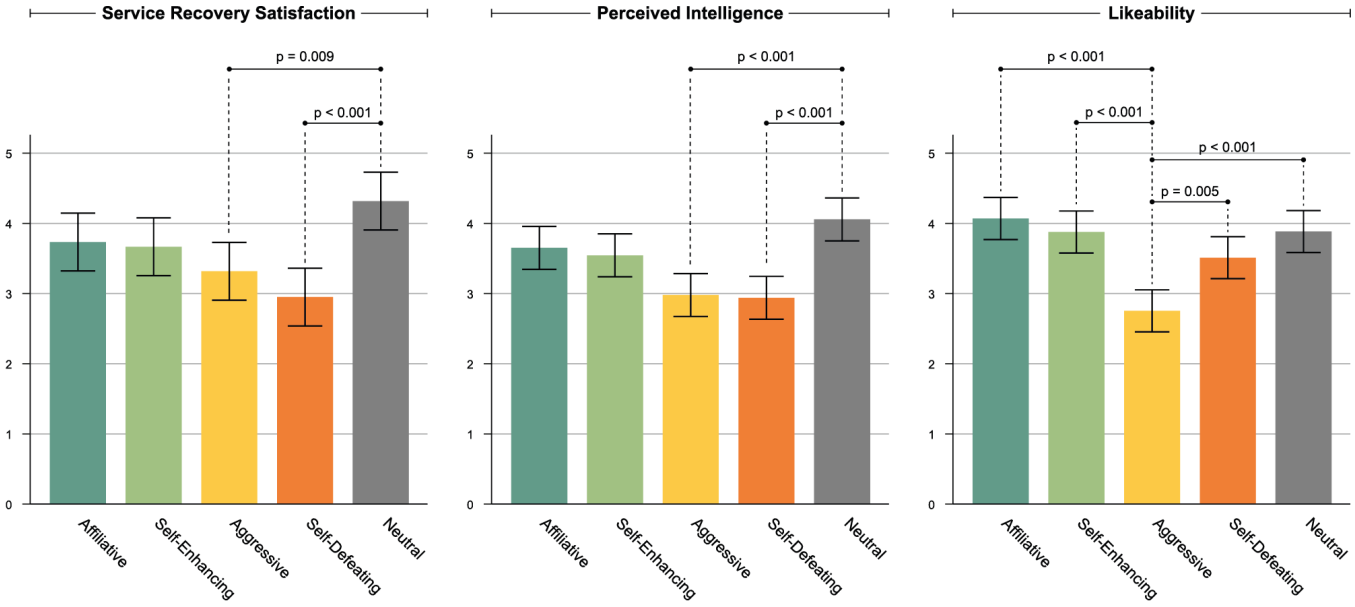

Figure 3: Results of the components service recovery satisfaction, perceived intelligence, and likeability. One-way repeated measures ANOVA was conducted to discover the effects of the five agent personalities. All pairwise comparisons were conducted using Tukey's HSD test. The standard deviation is represented by the error bars and only the significant comparisons ($p < 0.05$) are highlighted.

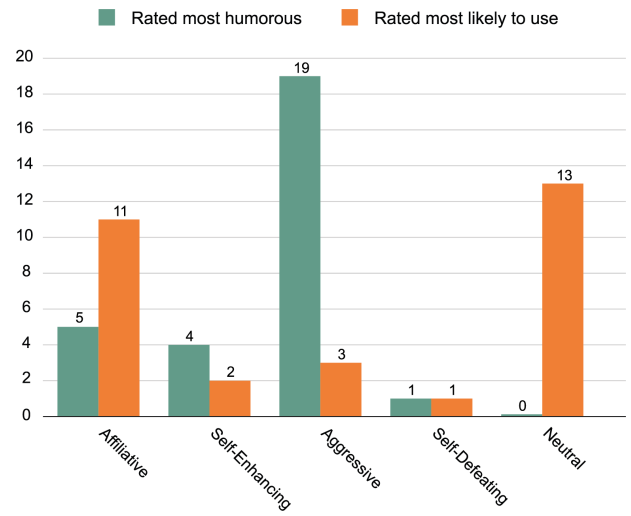

Figure 4: Participant ratings for the "most humorous" and "most likely to use." The count shown represents the number of participants who selected that personality.

relationships with others [22], and in their paper, conductive humor personality types were chosen. Compared to our data, we see no apparent connection between the conductive or detrimental elements. Additional work should examine the perception of the VA more deeply. In research focused on perception and responses to AI teammates in games, findings suggest that beyond the actual behaviors and actions, the identity of the agent led to very different perceptions [25], e.g. actions by a human were perceived as more enjoyable than the same actions by an AI [24].

5.4 Humor frequency and length

During the experiment, several participants stated that the length of responses presented by the four humor personalities was far too long. This was both regarding the communication success and the communication breakdowns. These participants also preferred the neutral personality for its minimal but precise responses, saying: P12 - *"Good, really good... she comes with quick responses compared to the last one... I liked that"* and P22 - *"Maybe its responses are a bit long [...] some of the others have very long responses"*, and P18 - *"She should just do what she is told"*. Furthermore, participants also expressed their dislike toward the frequency of humor used in the four personalities, commenting: P6 - *"I don't feel like that they have to be funny every time, maybe just once in a while can they come with a funny remark"* and P20 - *"It would be nice if I could turn it off [referring to the humor], but also existing to know what it has to say"*. This outcome is also in line with previous studies where participants noted the overuse of humor and referred to the frequency of jokes [8]. This implies that there may be an ideal balance between a short precise answer and the use of humor. Future research should take up these challenges with studies outside the lab and examine how humorous voice assistants can fit within the diversity of situations in daily life.

5.5 Voice assistant personalization and adaptation

Recent research has looked for meaningful ways to personalize and adapt voice assistants to suit the personality of the user [17]. The participants in our study also mentioned this during the experiment: P21 - "It could be cool if you could like... like choose different personalities" and P28 - *"I'm more a fan of this personality, but it could be nice if I could choose between 10 variations"*, and P29 - *"[...] I would probably like a combo between the two (neutral and affiliative)"*. This study was focused on repair strategies that used specific clarification responses (SR) to breakdowns [3]. Next steps toward personalization should explore other clarification responses including *acting on a misunderstanding* (AoM) and *neutral clarification response* (NR) [3], which may fit well into specific contexts and activities. There are various strands of research that could enable more sophisticated personalization, for example, GPT-4 has demonstrated some early signs of having a theory of mind [5], which may in the future open possibilities for context-aware sarcasm and other more advanced and personalized humor.

5.6 Implications for design

The importance of humor in the creation and maintenance of human interpersonal relationships has consistently been studied over the years [15, 40]. While there have been various attempts to use anthropomorphisation in voice assistants, chatbots, and robots that leverage humor to improve user perceptions, the results are mixed. This is particularly relevant given that numerous studies on voice assistant applications have reported high levels of user satisfaction, but low levels of intention to continue using the technology [34, 35]. There is mounting evidence that utilizing humor in voice assistants can be risky due to its subjective nature and potential for inappropriate content. Our study found that neutral or affiliative humor was better received after communication breakdowns with a voice assistant, regardless of perceptions of funniness or personality matching. Our results suggest that in the context of communication breakdowns, humor should be used carefully as it may not have the intended effect if it is too aggressive. However, as is the case in other domains (e.g., website aesthetics [33]) involving subjective perceptions from the application domain is important. While neutral or affiliative humor may be preferable in communication breakdowns, more aggressive humor may be appropriate in other situations. Further research in different contexts and domains would be necessary to help designers determine the appropriateness of different types of humor in varying circumstances.

5.7 Limitations and Future Work

There are various limitations of our work. In regards to research design, we utilized the Wizard of Oz (WoZ) technique [14], and conducted the study in a lab setting with preconfigured responses, which does not accurately reflect how users would interact with voice assistant or smart speaker in their home. While our results provide initial findings about humor as a repair strategy, in future studies it would be useful to study more complex and prolonged interactions.

Another limitation of this study is the unidimensional representation of nationality in our participant sample. Research provides strong evidence that disparities in humor perceptions exist across diverse cultural heritages and nationalities (e.g., [13, 20, 44]). Therefore it would be fair to assume that a more diverse sample regarding cultural background may have yielded different results. Our study provides insights into the humor style preferences of a cross-section of people from Denmark. Future research should also engage participants from other cultural backgrounds to examine how people from other backgrounds and culture experience humor as a repair strategy.

We have used the HSQ, an established and widely used questionnaire for determining humor styles that is available in many different languages. Schermer et al. [36] conducted a large-scale study comparing humor styles of 28 countries using the HSQ. Unfortunately, Denmark or any other Scandinavian country was not part of this study which would allow us to compare our results. Lundquist [18], showed that Danish humor preferences are aligned with aggressive humor. Our results provide support for this and if we would project our HSQ findings into the ranking by Schermer et al. [36] our participants to be approximately ranked sixth highest in aggressive humor compared to other countries. However, due to the limited sample size and sampling method, we cannot make reliable assumptions about the overall Danish population. This highlights the importance of conducting larger studies within the Danish context and including a more diverse range of cultural backgrounds.

Longitudinal studies could be valuable in researching humorous voice assistant personalities to understand how users appropriate and respond to humor over time. We hypothesize that there would be increased satisfaction with the appropriately selected humor personality, as also proposed by Lopatovska et al. [17]. Furthermore, a more sophisticated voice assistant could be developed that adapts to the user over time and involves contemporary issues and contextually relevant humorous responses.

6 CONCLUSION

Maintaining fluid interactions and conversations with a voice assistant necessitates appropriate error mitigation and repair strategies. The results from our user study suggest that participants were more satisfied and perceived the neutral voice assistant personality as more intelligent than a voice assistant with an aggressive or self-defeating humor personality. Furthermore, a voice assistant with an aggressive humor personality was found to be less likable by the participants than other voice assistant personalities. Our findings have implications for developing humor-based error-mitigation strategies for voice assistants.

REFERENCES

[1] 2022. EF EPI 2022 – EF English Proficiency Index. https://www.ef.nl/epi/

[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[3] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300473

[4] Arthur Asa Berger. 2017. *An anatomy of humor.* Routledge.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al.

2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[6] Moniek Buijzen and Patti M Valkenburg. 2004. Developing a typology of humor in audiovisual media. *Media psychology* 6, 2 (2004), 147–167. https://doi.org/10.1207/s1532785xmep0602_2

[7] Cuicui Cao, Yingying Hu, and Haoxuan Xu. 2022. A Mind in Intelligent Personal Assistants: An Empirical Study of Mind-Based Anthropomorphism, Fulfilled Motivations, and Exploratory Usage of Intelligent Personal Assistants. *Frontiers in Psychology* 13 (2022). https://doi.org/10.3389/fpsyg.2022.856283

[8] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. *Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?* Association for Computing Machinery, New York, NY, USA, 14. https://doi.org/10.1145/3411764.3445068

[9] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 2018. Why Doesn't It Work? Voice-Driven Interfaces and Young Children's Communication Repair Strategies. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. Association for Computing Machinery, New York, NY, USA, 337–348. https://doi.org/10.1145/3202185.3202749

[10] Xiang Ge, Dan Li, Daisong Guan, Shihui Xu, Yanyan Sun, and Moli Zhou. 2019. Do Smart Speakers Respond to Their Errors Properly? A Study on Human-Computer Dialogue Strategy. In *Design, User Experience, and Usability. User Experience in Advanced Technological Environments*, Aaron Marcus and Wentao Wang (Eds.). Springer International Publishing, Cham, 440–455.

[11] Jose Maria C Ibardaloza, Juan Antonio G Mapua, and Wilson M Tan. 2021. Modifying Voice-User Interfaces for Resiliency and Offline Management of IoT Devices. , 131–137 pages.

[12] Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in psychology* 10 (2019), 123.

[13] Tonglin Jiang, Hao Li, and Yubo Hou. 2019. Cultural differences in humor perception, usage, and implications. *Frontiers in psychology* 10 (2019), 123.

[14] J. F. Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '83)*. Association for Computing Machinery, New York, NY, USA, 193–196. https://doi.org/10.1145/800045.801609

[15] Herbert M Lefcourt. 2001. *Humor: The psychology of living buoyantly*. Springer Science & Business Media.

[16] Irene Lopatovska. 2020. Classification of humorous interactions with intelligent personal assistants. *Journal of Librarianship and Information Science* 52, 3 (2020), 931–942.

[17] Irene Lopatovska, Alice Griffin, Kelsey Gallagher, Caitlin Ballingall, Clair Rock, and Mildred Velazquez. 2019. User recommendations for intelligent personal assistants. *Journal of Librarianship and Information Science* 52 (04 2019), 15. https://doi.org/10.1177/0961000619841107

[18] Lita Lundquist. 2014. Danish humor in cross-cultural professional settings: linguistic and social aspects.

[19] Amama Mahmood, Jeanie W Fung, Isabel Won, and Chien-Ming Huang. 2022. Owning Mistakes Sincerely: Strategies for Mitigating AI Errors. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 578, 11 pages. https://doi.org/10.1145/3491102.3517565

[20] G Neil Martin and Erin Sullivan. 2013. Sense of Humor Across Cultures: A Comparison of British, Australian and American Respondents. *North American Journal of Psychology* 15, 2 (2013).

[21] Rod Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality* 37 (02 2003), 48–75. https://doi.org/10.1016/S0092-6566(02)00534-2

[22] Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*. Academic press.

[23] Michael McTear, Zoraida Callejas, David Griol, Michael McTear, Zoraida Callejas, and David Griol. 2016. Conversational interfaces: past and present. *The Conversational Interface: Talking to Smart Devices* (2016), 51–72. https://doi.org/10.1007/978-3-319-32967-3_4

[24] Tim Merritt, Kevin McGee, Teong Leong Chuah, and Christopher Ong. 2011. Choosing human team-mates: perceived identity as a moderator of player preference and enjoyment. In *Proceedings of the 6th International Conference on Foundations of Digital Games*. 196–203.

[25] Tim Merritt, Christopher Ong, Teong Leong Chuah, and Kevin McGee. 2011. Did you notice? artificial team-mates take risks for players. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer, 338–349.

[26] Tim R Merritt, Kian Boon Tan, Christopher Ong, Aswin Thomas, Teong Leong Chuah, and Kevin McGee. 2011. Are artificial team-mates scapegoats in computer games. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 685–688.

[27] John C Meyer. 2000. Humor as a double-edged sword: Four functions of humor in communication. *Communication theory* 10, 3 (2000), 310–331.

[28] Yohan Moon, Ki Joon Kim, and Dong-Hee Shin. 2016. Voices of the internet of things: An exploration of multiple voice effects in smart homes. , 270–278 pages.

[29] Billie Akwa Moore and Jacqueline Urakami. 2022. The impact of the physical and social embodiment of Voice User Interfaces on User Distraction. *International Journal of Human-Computer Studies* (2022), 102784. https://doi.org/10.1016/j.ijhcs.2022.102784

[30] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. 2013. Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics* 5, 2 (2013), 171–191.

[31] Stefan Olafsson, Teresa K O'Leary, and Timothy W Bickmore. 2020. Motivating health behavior change with humorous virtual agents. , 8 pages.

[32] Julie Pallant. 2007. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows Version 15* (3rd ed.). Open University Press, USA.

[33] Eleftherios Papachristos and Nikolaos Avouris. 2013. The influence of website category on aesthetic preferences. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part I 14*. Springer, 445–452.

[34] Eleftherios Papachristos, Dorte P Meldgaard, Iben R Thomsen, and Mikael B Skov. 2021. ReflectPal: Exploring Self-Reflection on Collaborative Activities Using Voice Assistants. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part IV 18*. Springer, 187–208.

[35] Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–27.

[36] Julie Aitken Schermer, Radosław Rogoza, Maria Magdalena Kwiatkowska, Christopher Marcin Kowalski, Sibele Aquino, Rahkman Ardi, Henrietta Bolló, Marija Branković, Razieh Chegeni, Jan Crusius, et al. 2019. Humor styles across 28 countries. *Current Psychology* (2019), 16.

[37] Eike Schneiders, Eleftherios Papachristos, and Niels van Berkel. 2021. The effect of embodied anthropomorphism of personal assistants on user perceptions. In *Proceedings of the 33rd Australian Conference on Human-Computer Interaction*. 231–241. https://doi.org/10.1145/3520495.3520503

[38] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[39] Chen Shani, Alexander Libov, Sofia Tolmach, Liane Lewin-Eytan, Yoelle Maarek, and Dafna Shahaf. 2022. "Alexa, Do You Want to Build a Snowman?" Characterizing Playful Requests to Conversational Agents. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 423, 7 pages. https://doi.org/10.1145/3491101.3519870

[40] Michelle N Shiota, Belinda Campos, Dacher Keltner, and Matthew J Hertenstein. 2004. Positive emotion and the regulation of interpersonal relationships. *The regulation of emotion* 68 (2004).

[41] Paul J Silvia and Rebekah M Rodriguez. 2020. Time to renovate the Humor Styles Questionnaire? An item response theory analysis of the HSQ. *Behavioral Sciences* 10, 11 (2020), 173.

[42] Julian Striegl, David Gollasch, Claudia Loitsch, and Gerhard Weber. 2021. Designing VUIs for Social Assistance Robots for People with Dementia. , 145–155 pages.

[43] Topvoiceapps. 2022. *SSML WYSIWYG EDITOR AND TESTER*. Topvoiceapps. Retrieved june, 6, 2022 from https://topvoiceapps.com/ssml

[44] Xiaodong Yue, Feng Jiang, Su Lu, and Neelam Hiranandani. 2016. To be or not to be humorous? Cross cultural perspectives on humor. *Frontiers in psychology* 7 (2016), 1495.

[45] Linghan Zhang, Sheng Tan, Zi Wang, Yili Ren, Zhi Wang, and Jie Yang. 2020. Viblive: A continuous liveness detection for secure voice user interface in iot environment. , 884–896 pages.