Enthusiasts, *Pragmatists*, and *Skeptics*: Investigating Users' Attitudes Towards Emotion- and Personality-Aware Voice Assistants across Cultures

Yong Ma
LMU Munich
Munich, Germany
yong.ma@ifi.lmu.de

Yomna Abdelrahman
Bundeswehr University of Munich
Munich, Germany
yomna.abdelrahman@unibw.de

Barbarella Petz
LMU Munich
Munich, Germany
barbarella.petz@gmail.com

Heiko Drewes
LMU Munich
Munich, Germany
heiko.drewes@ifi.lmu.de

Florian Alt
Bundeswehr University of Munich
Germany
florian.alt@unibw.de

Heinrich Hussmann
LMU Munich
Munich, Germany
hussmann@ifi.lmu.de

Andreas Butz
LMU Munich
Munich, Germany
andreas.butz@ifi.lmu.de

ABSTRACT

Voice Assistants (VAs) are becoming a regular part of our daily life. They are embedded in our smartphones or smart home devices. Just as natural language processing has improved the conversation with VAs, ongoing work in speech emotion recognition also suggests that VAs will soon become emotion- and personality-aware. However, the social implications, ethical borders and the users' general attitude towards such VAs remain underexplored. In this paper, we investigate users' attitudes towards and preferences for emotionally aware VAs in three different cultures. We conducted an online questionnaire with $N = 364$ participants in Germany, China, and Egypt to identify differences and similarities in attitudes. Using a cluster analysis, we identified three different basic user types (*Enthusiasts*, *Pragmatists*, and *Skeptics*), which exist in all cultures. We contribute characteristic properties of these user types and highlight how future VAs should support customizable interactions to enhance user experience across cultures.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**;

KEYWORDS

voice interfaces, affective computing, emotion-awareness, personality awareness, culture-dependent attitudes

MuC '22, September 4–7, 2022, Darmstadt, Germany

ACM ISBN 978-1-4503-9690-5/22/09…$15.00
https://doi.org/10.1145/3543758.3543776

ACM Reference Format:
Yong Ma, Yomna Abdelrahman, Barbarella Petz, Heiko Drewes, Florian Alt, Heinrich Hussmann, and Andreas Butz. 2022. *Enthusiasts*, *Pragmatists*, and *Skeptics*: Investigating Users' Attitudes Towards Emotion- and Personality-Aware Voice Assistants across Cultures. In *Mensch und Computer 2022 (MuC '22), September 4–7, 2022, Darmstadt, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3543758.3543776

1 INTRODUCTION

The recent advances in artificial intelligence (AI), speech recognition and synthesis, as well as natural language processing have sparked the widespread use of voice assistants (VAs) in our daily life. VAs exist in both mobile and stationary devices which means they are potentially always in our vicinity. Voice interfaces use an intuitive and natural form of interaction, inspired by human-human communication. In specific contexts, they have advantages over other interfaces such as keyboard and mouse, or touch displays. Voice interfaces keep the users' hands and eyes free for other tasks. They require no space for interaction, are hygienic as there is nothing to touch, and they are claimed to be easy to operate as they work with natural language. While the quality of voice recognition and synthesis is crucial, it is merely one of several relevant aspects of voice-based interaction. Prior work showed that users' perception of a voice plays a significant role [48]. Users tend to intuitively react to the human-like features of a computer voice as if they would interact with a human [36, 48, 57]. This inspired the development of technologies such as Google Duplex [37] where the VA has a natural-sounding human voice instead of a robotic one. If it becomes almost impossible to distinguish a computer voice from a real person, this creates a range of new challenges, such as calibrating trust and expectations.

Existing research asks to consider social and psychological factors, including affect and emotion, trust, credibility, and the relational context [65, 67]. Additionally, emotion- and personality-awareness are substantial for a natural conversation between humans; consequently, there is a vast body of research in the HCI community to

achieve these abilities also for voice interfaces [5, 64, 65]. However, there are still many open questions on the topic, such as the users' acceptance of such a technology, the details on how the voice assistant should react to the users' emotions, and potential ethical borders. Moreover, the answers to some of these questions will vary with other social aspects, such as education level, gender, age group, and culture.

Previous work on VAs reveals a gap in identifying and understanding diverse user groups. As these technologies are now adopted worldwide and used across cultures, identifying specific user groups could reveal design needs and concerns among these specific groups. We set out to identify similarities among users across cultures to identify culture-independent user groups. Therefore, we investigate users' preferences and attitudes towards emotion- and personality-aware VAs across German, Chinese, and Egyptian users. In particular, we address these research questions:

- **RQ1:** What is the general attitude of users towards emotion- and personality-aware VAs?
- **RQ2:** How do different cultures influence the preferences regarding VAs?
- **RQ3:** How does gender influence the preferences regarding VAs?
- **RQ4:** Can we identify culture-independent user groups?

2 RELATED WORK

Our work builds upon previous work on 1) voice assistants and voice AI, and 2) studying diverse user groups. We will therefore address these two fields of related work.

2.1 Voice Assistants and Voice AI

2.1.1 Advanced Voice Assistants. When research on voice-based dialogue systems started more than 40 years ago, the main challenge was understanding spoken words and synthesizing voice output. Four decades later, and with the basic technical challenges solved, voice assistants have become part of the real-life of families [27, 54], and we can move on towards making conversations with VAs more similar to human-human conversations. This includes even the ability to understand dialects, which is achieved by methods based on phonetic posteriorgrams [71] or deep learning algorithms [1, 21]. There is research on emotion recognition in speech [34, 56] and emotion synthesis in voice [4, 55]. AffectAura [43] is one of the first emotion-aware assistants that automatically collects emotional cues over a longer period of time to help users reflect on their emotional well-being. Future voice assistants may also perceive users' personalities and adapt their own, based on personality detection [41, 53] and synthesis [3]. Additionally, voice assistants may eventually be able to track mental health based on speech signal analysis [6, 63] and diagnose mental disorders.

2.1.2 Our relationship with Voice Assistants. Clark et al. [9] have found a clear distinction in the expectation of users between conversing with a purely functional assistant and establishing a relationship with a separate entity, where the functional viewpoint is clearly dominating these days. In our approach, we make a step towards designing the new genre of conversation which is suggested there: Designing voice assistants in a way that they do not try to replace a human counterpart but optimize a functional conversation towards higher trust and less friction. According to Nass et al. [47] users tend to ascribe personality to computers even if the computer does not possess any personality traits. This is also called the ELIZA effect, named after one of the first conversational agents from the sixties [68]. Emotion- and personality-aware VAs have the potential to improve Human-Computer Interaction (HCI), as for example shown by McDuff et al. [22]: Emotion-aware voice assistants can help users to reflect on their feelings and become more self-aware, which can lead to improved well-being and mood [28, 61]. Agents designed according to psychological models of personality have been studied for instance by McRorie et al. [44], and a more recent model has been developed by Völkel et al. [66]. Virtual assistants that show certain personality traits are perceived as more trustworthy [5, 72] and in driving situations they can increase the safety by reducing stress [5, 23]. The personality aspects of current voice assistants are the same for all users, although Braun et al. [5] showed that users benefit more from using voice assistants when their personalities match. Klein et al. [35] presented a survey on user acceptance for such VAs for a German user group, but focused mostly on concerns about privacy issues. In our work, we try to improve their work by investigating the attitude toward VAs' capabilities regarding emotions and personality for users on three continents.

2.1.3 Voice AI. Combining the power of artificial intelligence and conversational agents, we may arrive at hyper-personalized, intelligent agents as proposed by Zhou et al. [72]. Current prototypes of such conversational agents are still based on textual interaction but make use of indicators for personality, and therefore may soon be able to replace a conversation screening for personality in a human resources department. In our work, we do not follow these thoughts much further, but rather focus on the aspect of how users perceive such emotion- and personality-sensitive agents.

2.2 Studying Diverse User Groups

It has been recognized in HCI that often a one-size-fits-all approach is insufficient when trying to optimize the user experience for diverse user populations [10, Chapter 3]. Understanding different user groups with their specific needs and preferences will lead to more specific interface designs and address each group better. Users might differ by basic demographic parameters such as age, gender or education, but also by more complex constructs, such as their cultural background. We will first summarize how culture is (not so much) addressed in existing research, and then briefly discuss general approaches to categorizing users.

2.2.1 Cultural Characteristics. Technologies such as voice assistants are used worldwide in many different markets. However, they are mostly designed in Western countries for Western users [40]. In HCI studies, Western participants are over-represented since 73% of all study findings are based on Western participant samples. Especially African countries, but also other areas of the world are poorly represented. HCI in China has experienced relatively slow progress in the last decades [62] and Ma et al. [38] showed that China is also generally lacking research in the field of HCI. In addition, culture itself as a topic of HCI is far from being mainstream.

While psychology researched cultural constants and variations [15] as early as 1971 (see [45] for a review), the first approaches to intercultural HCI studies are from 1997 [16], followed by a number of theoretical frameworks [11, 24, 59, 60], as well as pairwise comparisons of two cultures as in [17, 19]. A famous framework for inter-cultural comparison was proposed by Hofstede [25] which was used for cross-cultural comparison of consumer behavior [12]. However, according to Kamppuri et al. [32], human factors studies should understand how culture and technology interact instead of simply comparing two cultures. We therefore decided to run our study in Africa, Asia and Europe, and not to focus on the differences between cultures, but to watch out for commonalities and the relation between culture and culture-independent user groups.

2.2.2 User Groups. Users can be grouped by a wide range of characteristics besides culture. Wisniewski et al. [69] categorized users into six profiles depending on their sharing and privacy attitudes on Online Social Networks (OSNs) and offered design implications for each user group. Omidvar-Tehrani et al. [49] developed an interactive framework (IUGA) that explores the user space based on group discovery primitives by creating labeled groups for similar types of users. User groups are created based on the collected user data that consists of basic demographic attributes (such as gender, age, native language) as well as of user interests. To form a user group, multiple users need to exhibit identical values for some of the attributes [49]. The user modeling community has developed a range of approaches for describing such groups of users (see, for example, Fischer [18] for a conceptual and Frias-Martinez et al. [20] for a technical overview).

3 QUESTIONNAIRE DEVELOPMENT

We asked for agreement or disagreement on a 5-point Likert scale for the majority of questions, using statements that were written in an unprejudiced manner. We only utilized a 7-point Likert scale for the general acceptance question in order to obtain more precise clustering of participants (see Figure 22). There were many choices for certain questions' viable responses. For every question, we included a space for free text so that respondents can contribute further information.

3.1 Determining Cultural Identity

Associating a person with a culture is a delicate task [26]. Nationality is not a reliable criterion, as large nations can have several cultures and some cultures spread over different nations, and members of a culture can have nationalities which do not match their culture. Participants may even have parents from two different cultures or may currently live in a third culture. Therefore, we decided to determine participants' culture by asking for their mother tongue since we considered this a strong indicator of their upbringing and hence their cultural socialization.

3.2 Questionnaire Structure

For a systematic approach we analyzed the question space and identified three dimensions: a technology dimension, a social dimension, and a contextual or situational dimension. In several interactive brainstorming sessions, and from current discussions in the media, such as the (default) gender of an assistant's voice[1], we derived 39 questions which filled the structure of the question space. The final selection of questions is explained below.

3.2.1 Demographics. The questionnaire starts with the participants' demographics. Besides the mother tongue, we asked for age, gender, self-assessment of computer knowledge, and voice assistants the participant had already used.

3.2.2 Technology Dimension. We structured the technology dimension along with seven categories: technology in general, voice recognition, voice synthesis, emotion detection, emotion synthesis, personality detection, and personality synthesis. These seven categories were later used to structure the questionnaire on the top level (see Table 2).

3.2.3 Social Dimension. For the social dimension we chose the categories acceptance, interaction details, privacy and security, relation to the voice assistant, and ethics/moral. Acceptance of the technology means whether users are looking forward to it, or whether they are skeptical. The interaction details are questions on how to implement a voice assistant, e.g., what voice to use and whether the device should be configurable or self-adjusting. The privacy and security category includes questions on identification by voice, concerns on who has access to voice data, or competences of a voice assistant (e.g., for money transfer). The relation category deals with questions, such as whether a voice assistant should be a servant or a friend, whether it should respect hierarchies in a family or company, or whether it should be able to speak in the local dialect. The ethics and moral category have questions on gender issues or the usage of voice assistants by children, as Amazon's Alexa offers a child mode[2]. We also asked about ethical limits, such as the usage of voices from people who passed away. Shortly after our survey this topic also received attention in the media[3]. However, we decided not to ask questions on sex (flirting) and religion (praying), which would potentially have made the questionnaire inappropriate in some cultural contexts and might have endangered response numbers.

3.2.4 Context Dimension. The context dimension ranges between public and private, as users' preferences may differ for a private voice assistant in the private home, such as Alexa, or for a public voice interface, for example, in an elevator. However, there are situations which are neither private nor public. A guest present at a private home, or the usage of a voice assistant with a mobile phone while using public transport, makes the private voice assistant only semi-private. Similarly, a public voice assistant becomes semi-public if it is accessible only to a certain group of people.

3.2.5 Limiting Questionnaire Size. With seven categories in the technology dimension, five in the social dimension, and four in the context dimension, we would have arrived at a total of 140 ($7 \times 5 \times 4$) questions if we had intended to ask about all combinations of aspects. For the selection of questions we focused on interaction and

[1] https://www.nytimes.com/2019/05/22/world/siri-alexa-ai-gender-bias.html, last accessed July 6, 2022

[2] https://www.amazon.com/Echo-Dot-4th-Gen-Kids/dp/B084J4QQK1, last accessed July 6, 2022.

[3] https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice, last accessed July 6, 2022.

chose fewer questions on personality, as this topic was researched already by Völkel et al. [66]. Table 1 shows the resulting distribution of questions over the question space categories. The final questionnaire is listed in Table 2. It had 46 questions, seven on demographic data and 39 on user preferences. Participants were able to fill the entire questionnaire within half an hour on average.

3.3 Recruitment of Participants and Data Treatment

The questionnaire was made available online and distributed through our private networks. To obtain a more diverse sample, we complemented recruiting with snowball sampling, where initial participants were asked to spread the link. Participants who completed the questionnaire could choose to either participate in a raffle for one of ten shopping vouchers or to receive credit points for their studies as compensation.

All data was collected anonymously. Participants gave an informed consent to our storing and using their anonymous responses for research purposes. We had a separate contact information form for participants who wished to be compensated with a shopping voucher. This contact information was collected in a separate questionnaire and deleted after compensations were issued. The survey was filled by 576 participants, of which we only considered 364 complete responses for our analysis, as we excluded questionnaires with any empty response fields.

4 RESULTS

4.1 Demographics

We distinguished between three cultures and a category *other*. Among the 149 German participants, 76 self-identified as female and 73 as male, ages ranged from 17 to 64 years ($M = 26.2, SD = 8.30$), 125 participants had previous experience with VAs. From the 102 Chinese participants, 56 self-identified as female and 46 as male. They were aged between 18 and 56 years ($M = 28.8, SD = 5.27$), and 94 of them had previous experience with VAs. Lastly, from the 80 Egyptian participants, 39 self-identified as female and 41 as male. They were aged between 20 and 67 years ($M = 30.6, SD = 7.73$), and 70 of them had previous experience with VAs. 12 participants chose English as their mother tongue and 21 chose other. As shown in Table 3, gender was relatively balanced across all cultures.

Table 1: Distribution of questions over categories. The three dimensions of the question space are projected onto two dimensions (social and technological), merging the context dimension for legibility.

Technology	Acceptance	Relationship	Interaction	Ethics	Privacy	Questions
General	1	5	1	1	1	9
Voice recognition	1	1	1	1	3	7
Voice synthesis	-	-	5	3	-	8
Emotion detection	1	-	1	1	1	4
Emotion synthesis	1	-	2	2	1	6
Perso. detection	1	-	-	-	-	1
Perso. synthesis	-	1	2	-	1	4
Questions	5	7	12	8	7	39

Table 2: List of questions in the English version of the questionnaire. The bold question numbers indicate Likert-scale-based answers. Underlined question numbers indicate categorical choices, and italic indicate free text entry.

	questions on demographic data
A1.	What is your native language?
A2.	How old are you? Please state your age in years.
A3.	Please select your gender.
A4.	What is your professional field or field of study?
A5.	How would you rate your computer skills?
A6.	Which voice assistants have you used before?
A7.	How often do you use voice assistants?
	questions on general technology
B1.	Voice assistants are becoming more and more common. What is your attitude towards voice assistants?
B2.	Do you prefer a standardized voice assistant with the same voice for everyone (like Alexa or Siri) or a voice assistant that has a unique voice exclusively for you?
B3.	Do you want to have conversations with your voice assistant that are not task related e.g. chitchat?
B4.	Voice assistants can be used by anyone who is able to verbally communicate regardless of age. Do you think there should be an age restriction on the usage of voice assistants by minors?
B5.	Do you want to have a child-mode, so children can play and learn with voice assistants
B6.	Do you want to use a voice assistant that supports your mental health?
B7.	How do you want your voice assistant to deal with your mental health?
B8.	Do you think in some situations (e.g. during the COVID-19 pandemic) voice assistants could substitute conversations with human beings?
B9.	Your voice assistant has to know about your preferences to become personalized. How do you want your voice assistant to learn about your preferences?
	questions on voice recognition
C1.	Do you want to be identified by your voice assistant?
C2.	Do you feel comfortable using voice assistants, knowing they record every interaction with you and send the collected data to the server of a company?
C3.	You use your voice assistant together with other members of your household. How should the voice assistant prioritize multiple or contradicting inputs coming from different people at the same time?
C4.	Do you think it is ethically justifiable to allow voice assistants to respond according to hierarchical structures instead of treating everybody equally? This could lead to e.g. needing your confirmation for your child's commands or prioritizing the commands of the voice assistant's owner.
C5.	A personalized voice assistant that fits your preferences and needs must collect at least some data. What type of data may your voice assistant collect and store, so you still feel comfortable?
C6.	When you use your voice assistant and other people are talking in the same room, their unintentional voice input is also recorded and stored by your voice assistant. Do you think these people should be informed that their voice may be recorded?
C7.	You ask your voice assistant to look up vegan recipes. Your voice assistant replies: "You mentioned last month that you do not like carrots, I therefore looked up recipes without carrots for you. Here are my suggestions". The voice assistant adapted its suggestions to information you gave in the past. Do you want your voice assistant to refer to information you gave in the past?
	questions on voice synthesis
D1.	What age do you generally prefer for the voice of a voice assistant?
D2.	What gender do you generally prefer for the voice of a voice assistant?
D3.	Which voice do you prefer for a voice assistant designed for children?
D4.	Do you want your personal voice assistant to be able to speak and understand dialects?

D5.	It is possible to synthesize any voice, even the voice of existing people. Which voice do you choose for your voice assistant?
D6.	If there was the possibility to recreate the voice from a person that has passed away, do you think it would be acceptable to use their voice?
D7.	Most voice assistants come with a female voice by default. A news organization stated: "Women have been made into servants once again. Except this time, they're digital." Do you agree or disagree?
D8.	Do you believe voice assistants should respond differently depending on the users age, e.g. by using child friendly language?
	questions on emotion detection
E1.	A voice assistant that is able to detect emotions responds more considerately and naturally. Do you want to use a voice assistant that is able to detect your emotions?
E2.	What kind of insight do you want on the emotions your voice assistant detects?
E3.	Do you believe a voice assistant could manipulate you (e.g. for commercial purposes) when being able to detect and interpret your emotions?
E4.	Do you want your voice assistant to store the emotions it detects, so it can track your emotional well-being?
	questions on emotion synthesis
F1.	There are linguistic parameters (e.g. words, phrases) and paralinguistic parameters (e.g. pitch, intensity) in a voice that can be interpreted as emotions. How do you want your personal voice assistant to express emotions?
F2.	Imagine you had a bad day. Your voice assistant detects sadness. How do you want your voice assistant to react to how you feel?
F3.	You use your voice assistant to get a news update. Do you want the device to adapt to the content, e.g. in tone or expressed emotions?
F4.	The emotions of your voice assistant could influence your mood. Which emotions should your voice assistant be allowed to express?
F5.	Do you believe a voice interface should be able to show affection to you?
F6.	Do you want your voice assistant to change which emotions it shows when other people (e.g. guests, roommates, partner, etc.) are present?
	questions on personality detection
G1.	Do you want to use a voice assistant that is able to detect and adapt to your personality?
	questions on personality synthesis
H1.	In what kind of relationship do you want your personal assistant to interact with you?
H2.	You share a voice assistant with other members of your household. Do you prefer to have a device with one personality per household or one that shows different personalities depending on the person using it?
H3.	Active voice assistants initiate conversations, make suggestions, and state their opinion. Passive voice assistants do not initiate conversations but rather wait for your commands and they do not make suggestions without being asked. Do you want your personal voice assistant to be more active or passive?
H4.	You use your mobile voice assistant at home in your room. You leave the house to take the bus. In the bus you decide to use your voice assistant again. How do you want the personality of your voice assistant to adapt to the new surrounding?

Table 3: Age (mean and standard deviation) and gender (number and percentage) distribution by mother tongue for 364 participants.

Mother	Age (Mean, SD)	Male	Female	Total
German	$M = 26.2$, $SD = 8.30$	73 (49.0%)	76 (51.0%)	149
Chinese	$M = 28.8$, $SD = 5.27$	46 (45.1%)	56 (54.9%)	102
Egyptian	$M = 30.6$, $SD = 7.73$	41 (51.3%)	39 (48.8%)	80
other	$M = 26.6$, $SD = 6.97$	18 (54.6%)	15 (45.5%)	33
Overall	$M = 27.9$, $SD = 7.51$	178 (48.9%)	186 (51.1%)	364

4.2 General Attitude

We asked about the general attitude towards voice assistants (cf. Figure 1). The majority (58.5%) of participants have a positive attitude towards voice assistants. However, there is also a group (19.2%) with an averse attitude.

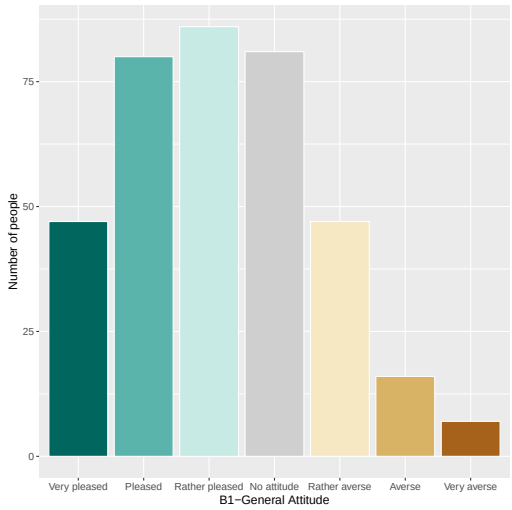

Figure 1: General attitude (B1) towards voice assistants for all participants.

Figure 2 shows an overview of the results of all Likert-scale-based questions. All these questions are on a 5-point scale from 'Yes' to 'No' except for two questions. B2 ranges from 'standardized' to 'personalized 'and H3 ranges from 'active' to 'passive'. It depends on the question whether a bar to the left or the right side means a positive attitude, but for all questions the bars go to both sides. Our results answer **RQ1** by showing that there is no common opinion, and voice assistants often have to deal with users of varying opinions.

Figure 3 shows the correlations between each pair of Likert-scale-based questions. The matrix clearly shows a strong correlation between many pairs of questions. This means that there are groups of questions which are interrelated in the view of the participants. The answers to most questions are spread over a spectrum and are not homogeneous. Altogether this suggests that we may find certain groups of users who answered a certain group of question homogeneously, which means there are different user types. In section 4.6, we will examine this in more detail. Figure 4 and 5 show plots of the users' attitude towards emotion (E1) and personality detection (G1). There is an obvious correlation for this combination as can be seen from the fact that most data is on the diagonal. The largest group are the users who answered both questions with 'rather yes'. The group of users who answered both questions with 'no' are mostly German. This matches the results presented in section 4.6.

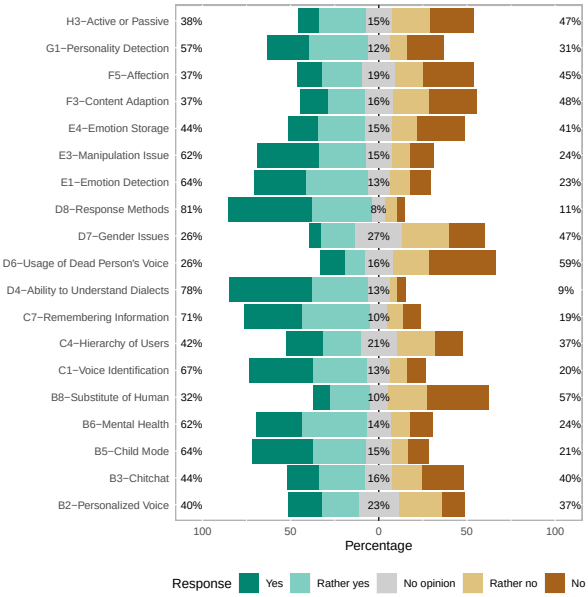

Figure 2: Overview of values for all Likert-based questions. The numbers on the gray field in the middle show the percentage of participants with no opinion. The numbers on the right and left side show the percentage of positive (yes and rather yes) and negative (no and rather no) answers respectively. The overview shows that the questions were answered in both directions.

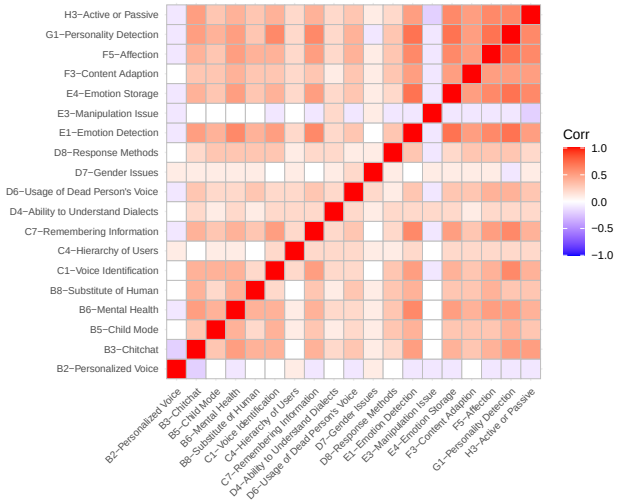

Figure 3: Correlation matrix for all Likert-based questions. The overview (Figure 2) shows that the questions were answered in great variation. The correlated pairs in the matrix tell that a participant who answered one question in a certain way also answers other question in the same way. From this we can conclude that there might be distinct types of users.

4.3 Selected Results

In the following section we discuss the most valuable insights from our results, based on their relevance to our research questions and on how timely they are to the on-going research on VAs.

4.3.1 Dialects. An overwhelming majority of our participants favored voice assistants which can speak dialects (D4). On an emotional level the question on the dialect is a question on the 'feeling of being understood' and 'speaking my language'. There are no differences based on gender, yet based on culture (see Figure 6 and 7). Egyptians had the strongest demand for dialects, potentially because Arab is spoken in many countries with country-specific dialects. Amazon's Alexa already offers five different dialects of English, and three dialects of Spanish. The desire for voice assistants speaking dialects will result in more available dialects in future.

4.3.2 Voice of People who Passed Away. One aim of our survey was to identify ethical limits. As it is technically possible to synthesize existing voices (if there are recordings), and this means also voices of people who passed away, we asked participants what they think about the usage of such voices (D6). These voices could be from relatives, for example, the grandmother, or famous people like actors, entertainers or statesmen. Those people cannot be asked for permission anymore. Another question is whether it is acceptable to revive persons in such a way. However, we simply asked the question without mentioning those aspects or giving examples. The results are depicted in Figure 8 and 9. The majority of the participants did not like the idea. The results by culture show that the concerns are biggest among Egyptians, followed closely by Germans.

4.3.3 Taking Care of Mental Health. Mental health is a sensitive topic. Health problems, especially psychological problems, are confidential as they may affect the relation to others. It was interesting for us to find out, whether users could build up the same bond of trust with a voice assistant as with a doctor. Figure 10 and 11 show the results for question B6. The majority of the people would like a voice assistant that takes care of their mental health. This also means that the trust in the diagnosis done by an artificial intelligence is quite high.

4.3.4 Private and Public Use. We asked whether a voice assistant running on a smartphone should change its personality when not being in a private environment but in public transport (H4). Figure 12 shows the results. The largest group wants to have a neutral personality in public. However, the second largest group is the group of people who do not care about the missing privacy and want no change in personality. The question what your personal communication device tells others about you does not only arise for personalized voice assistants. We know this question already from the personal ringtone of a mobile device.

4.3.5 Obeying Hierarchies. The answers for the question whether the voice assistant should obey hierarchies are evenly distributed over the Likert-scale from 'yes' to 'no'. It seems that there is no clear opinion on this topic. Although there is a small tendency for the

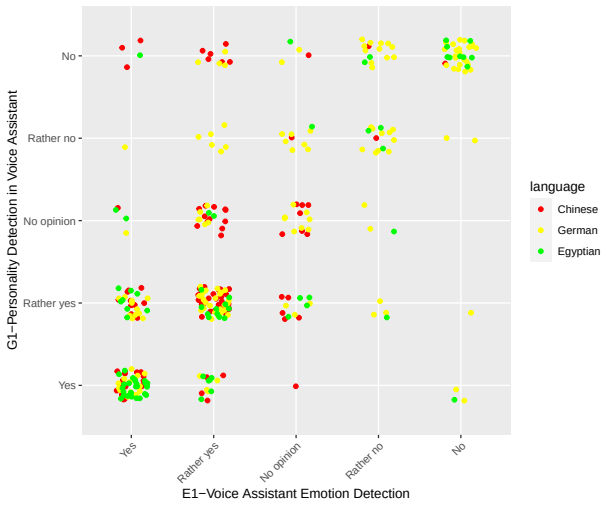

Figure 4: Attitude towards emotion and personality detection across cultures. The dots are distributed randomly around their integer coordinates to not cover each other. Every dot represents a participant. The data is correlated as most dots lie on the diagonal. The plot suggests some dependencies on culture.

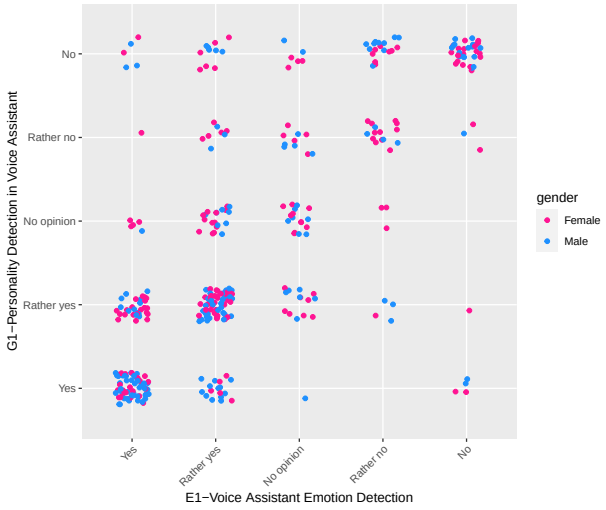

Figure 5: Attitude towards emotion and personality detection across gender. The distribution for gender is quite even.

male participants to obey hierarchies, there are no clear differences by gender (see Figure 13). Looking at the answers by culture (see Figure 14), the distribution is not uniform. The highest peak for Egyptians is on 'yes', for Chinese it is on 'no opinion', and for the Germans on 'rather no'.

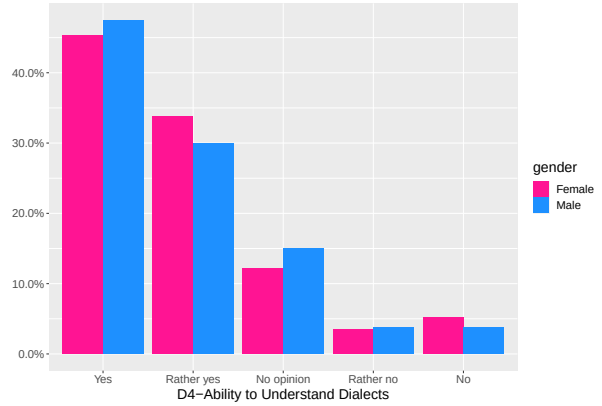

Figure 6: Results by gender for question D4 on whether a voice assistant should be able to speak dialects.

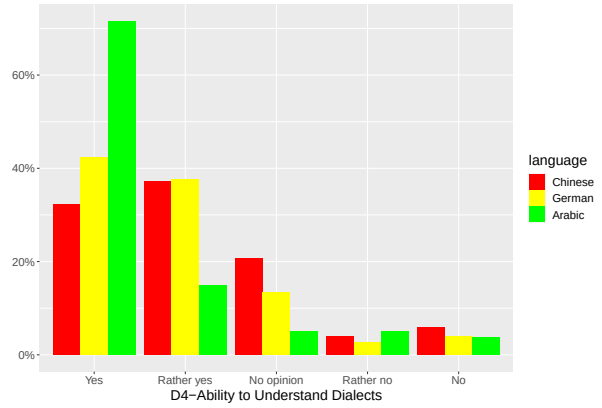

Figure 7: Results by culture for question D4 on whether a voice assistant should be able to speak dialects.

The social hierarchy of voice assistant users also has implications. Imagine a smart home environment with voice assistant control and the child telling the voice assistant to turn up the heating. Then the father says that the heating should be turned down. Existing voice assistants always execute the last order and this give the child the possibility to override the father's command and start a childish game. At this moment the father may wish hierarchy-awareness of the voice assistant. Now the mother comes in and the question arises whether she is on the same, a lower or higher hierarchy level than the father. The same question arises for the grandparents. Typically, there are (potentially complex) hierarchy levels in a family, but they may be unspoken or hidden. A hierarchy-aware VA would need to make these unspoken hierarchies explicit to some extent.

4.4 Gender Influence

We analyzed all Likert-based questions with a Wilcoxon-signed-rank test regarding gender. The results are summarized in Table 4. Out of 21 questions there are only five questions where the p-value is below the 5%-significance level. However, we should keep in mind that when looking for a 5%-significance level one of twenty

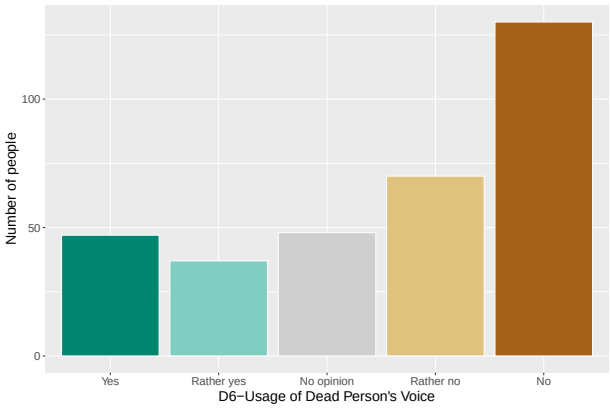

Figure 8: Results for question D6 on the usage of voices from people who passed away.

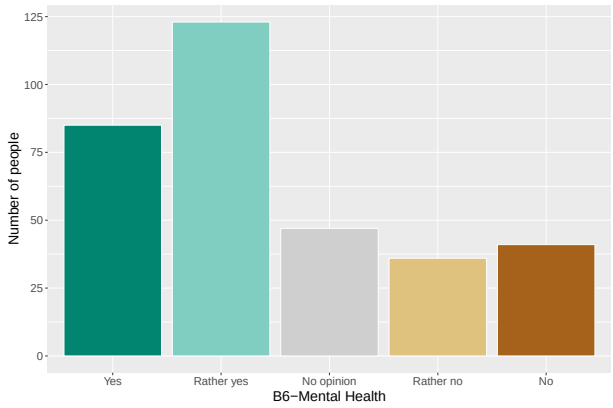

Figure 10: Results for question B6 on whether a voice assistant should take care of the user's mental health.

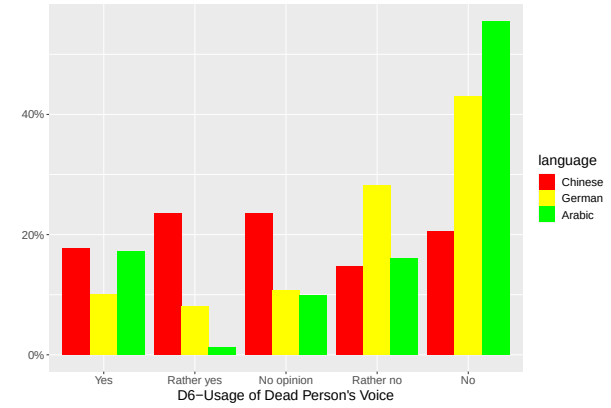

Figure 9: Results by culture for question D6 on the usage of voices from people who passed away.

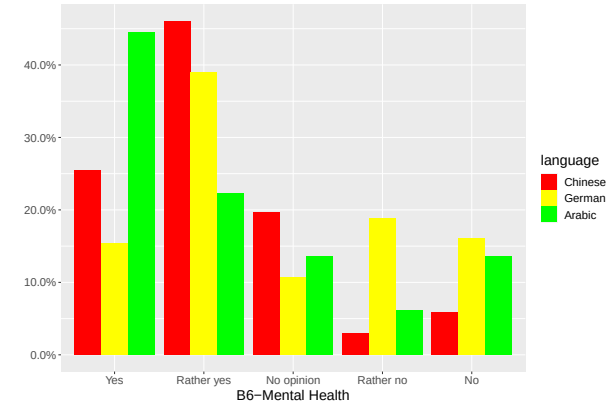

Figure 11: Results by culture for question B6 on whether a voice assistant should take care of the user's mental health.

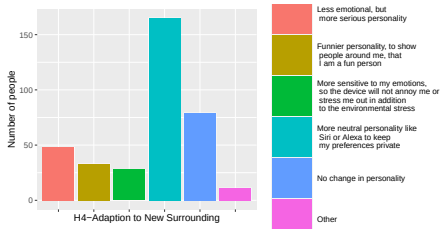

Figure 12: Results for question H4. Most people want to have a neutral voice assistant personality in public.

tests may report significance by chance. Therefore, we applied a Bonferroni correction and tested on a significance of 5%/21 = 0.0024 for the 21 questions in the gender comparison. There are only two questions (D7, D8), for which the test indicated a significant difference. Therefore, we argue that there are only small differences between the genders.

One of the significant differences between the genders is the question about the gender issue (D7), which asked about the attitude on female voice as default for voice assistants. Women see this more skeptical than men. As this is a gender-specific question, it is not surprising that the answers differ by gender. Figure 15 shows the distribution of the answers by gender. The other question with a significant difference between the genders is whether a voice assistant should respond differently to a child (D8). The distribution of the answers by gender is given on the right side of Figure 16. Women think children should not be treated like adults. However, this result should be taken with care, as there was a similar question (B5) asking whether there should be a child mode, which does not show a significant difference for gender.

4.5 Differences by Culture

As mentioned before, our intention to conduct the study in different cultures was rather to better understand diverse user populations, than to show that cultures are significantly different. Even if there are significant differences in the answers by cultures, this does not necessarily mean that the reason for the difference lies in the culture. The reason for differences could also lie in different wealth or in the number of voice assistants used currently.

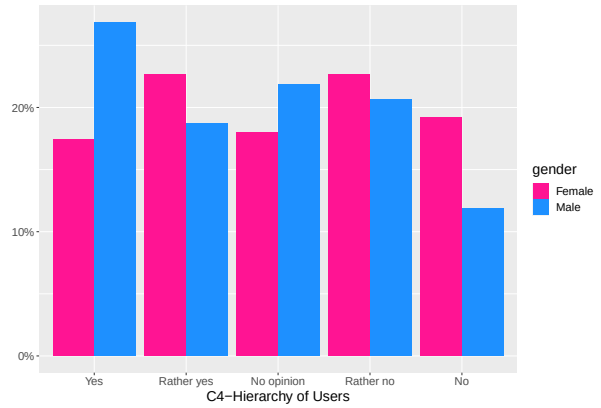

Figure 13: Results by gender for question C4 on whether a voice assistant should obey hierarchies.

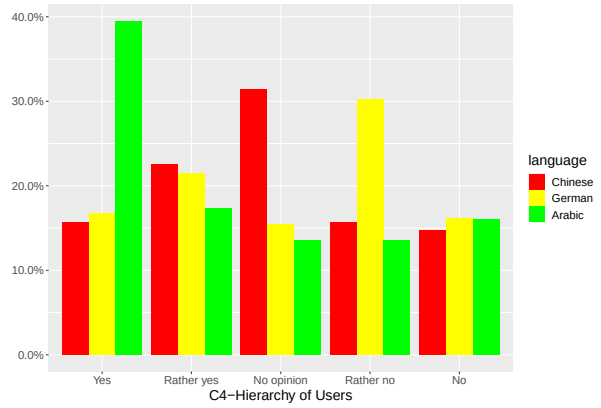

Figure 14: Results by culture for question C4 on whether a voice assistant should obey hierarchies.

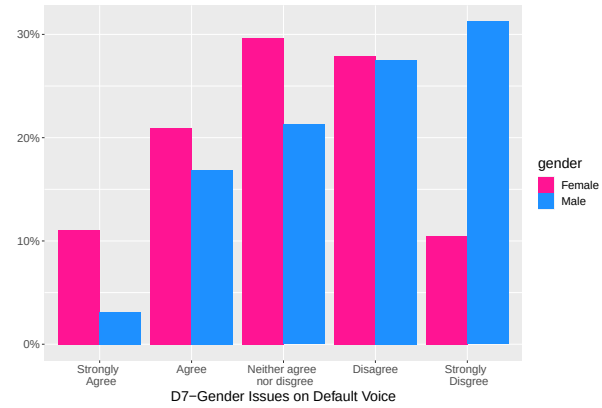

Figure 15: Opinion on default female voice (D7) by gender. This is one of two questions where we found significant differences by gender.

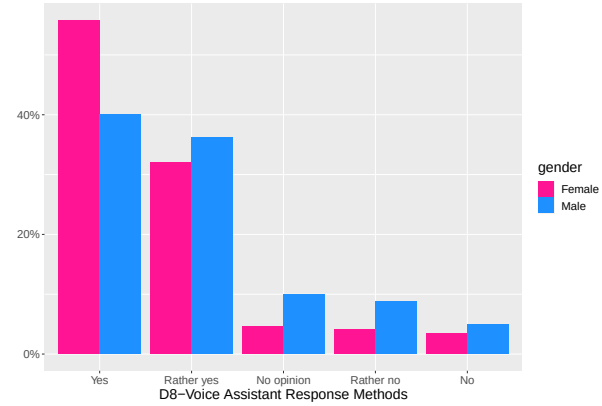

Figure 16: Opinion on whether children should get a different response (D8) by gender. This is one of two questions where we found significant differences by gender.

For all cultural comparisons we dropped all results from participants which stated English or other as their mother tongue, as their number is relatively small and will not give valid statistical results. Table 4 shows the results of a Wilcoxon-signed-rank test for comparisons of the mean values for the Likert-scale-based questions. As we did three comparisons for 21 questions, we used a Bonferrroni correction and set the significance level to 5%/63 (=0,00126%). It shows differences for the general attitude (B1) across all cultures. There are differences in the general attitude towards voice assistants across the cultures. The bar chart for this question across cultures is given in Figure 17. Most Chinese welcome the technology, Egyptians have a balanced opinion, and only some Germans are very averse towards emotion- and personality-aware voice assistants.

Figure 18 shows the answers to the question whether a female voice for a voice assistant degrades women to servants. The majority does not agree to the statement in all cultures, however among the cultures the Chinese do agree much more than the Egyptians, and the Germans are between both positions.

4.6 *Enthusiasts*, *Pragmatists*, and *Skeptics*

As shown in section 4.2 and Figure 3, participants' answers correlate. Together with the fact that there are participants on both sides of the scale in Figure 2, we took this as an indication that there may exist different types of users, and decided to run a cluster detection on our results. For this, we used all questions answered on a Likert-scale to detect clusters. In particular, these are B1, B3, B5, B6, B8, C1, C4, C6, C7, D4, D6, D7, D8, E1, E3, E4, F3, F5, G1. This means that we had to search for clusters in a 19-dimensional space. All these Likert-scale questions provide numeric values in the same range from 1 to 5 except one question (B1) with a range from 1 to 7.

For grouping the data points [29], K-means and agglomerative hierarchical clustering [8, 39, 46] are the two main methods which are considered adequate for a clustering analysis of Likert-scale data. The same methods have also been proposed for detecting user groups before [20]. Typically, K-means clustering is done after a principal component analysis (PCA) [13] to reduce the dimensionality before starting the cluster detection algorithm. Jin and Han [30] proposed K-medoids clustering as a variant of K-means that

Table 4: p-values of Wilcoxon signed-rank test for Likert-scale-based questions comparing gender and the three cultures. Values below 0.001 are given as <.001. Values at or below 0.05 are shown in blue. Values below 0.0024 (0.05/21) for gender and 0.0009 (0.05/63) for culture show significant differences after a Bonferroni correction and are colored in green.

Likert Scale Questions	Gender	Chinese-German	Chinese-Egyptian	Egyptian-German
B1. General Attitude	.157	**<.001**	**.005**	**.039**
B2. Personalized Voice	.145	.366	.472	.997
B3. Chitchat	.529	**<.001**	.216	**.006**
B5. Child mode	.296	.056	**.007**	**<.001**
B6. Mental health	.674	**<.001**	.442	**<.001**
B8. Substitute of human	.102	**<.001**	**.002**	.169
C1. Voice identification	.160	.089	**<.001**	**<.001**
C4. Hierarchy of users	.081	.303	**.016**	**.002**
C6. Informing others	.500	**<.001**	**<.001**	.233
C7. Remembering information	**.018**	.107	**.008**	**<.001**
D4. Ability to understand dialects	.699	.049	**<.001**	**<.001**
D6. Usage of dead person's voice	.367	**<.001**	**<.001**	.342
D7. Gender issues	**<.001**	**.044**	**.005**	.181
D8. Response to children	**<.001**	**<.001**	.928	**<.001**
E1. Emotion detection	.106	**<.001**	.750	**<.001**
E3. Manipulation Issue	.679	**<.001**	**<.001**	.980
E4. Emotion Storage	**.033**	**<.001**	.361	**<.001**
F3. Content adaption	.281	**<.001**	.062	**<.001**
F5. Affection	.201	**<.001**	.641	**<.001**
G1. Personality detection	**.003**	**<.001**	.073	**<.001**
H3. Active or passive	.091	**<.001**	.236	**<.001**

is more robust to noise and outliers, which was also confirmed by Shamsuddin and Mahat [58]. We therefore used K-medoids after a PCA with the PAM (Partitioning Around Medoids) algorithm [33]. One key parameter for both K-means and K-medoids is the selection of a value for K, which can be done using the elbow method, the silhouette coefficient algorithm or the gap statistics algorithm [70]. Both the elbow method (see Figure 19) and the gap statistics (see Figure 20) suggested a value of $k = 3$, so we used this in our K-medoids cluster detection. The result is depicted on Figure 21.

The next step was to find a meaning and a name for the clusters. After closer inspection, we felt that the three clusters somehow represented users who are positive, neutral or skeptical towards emotion and personality-aware voice assistants. To test this assumption, we plotted the general attitude (B1) for all three clusters. Figure 22 shows the result and seems to confirm our assumption. We then decided to call these user groups the *Enthusiasts*, *Pragmatists*, and *Skeptics*, respectively. Finally, we evaluated the membership in clusters with respect to cultures and gender. The result is shown in Figure 23 and 24. The figures show that the user types are equally distributed by gender, but that the distribution differs across cultures. The Chinese seem to be mostly *Enthusiasts* or *Pragmatists*, while the percentage of *Skeptics* is relatively small. Germans, on the other side, have the highest portion of *Skeptics*. Half of the Egyptians are *Enthusiasts*, while the rest are either *Pragmatists* or *Skeptics*.

To check whether the automatically detected clusters give a better partitioning than language or gender, we calculated their Dunn Index. The Dunn Index [7, 51] is the ratio of the smallest distance between data not in the same cluster to the largest distance within a cluster. The higher the distance between clusters and the smaller the cluster diameters, the higher the Dunn Index. For clusters by gender, the Dunn Index D_{gender} is 0.116. For clusters by culture, the Dunn Index $D_{culture}$ is 0.091. The Dunn Index for the user types $D_{usertype}$ is 0.170 and thus substantially higher than the other two. This implies that a separation by detected user type is more descriptive than by gender or culture alone (**RQ2-4**).

5 DISCUSSION

Our results confirmed some things we had expected before starting the study, but also revealed a number of surprises. For one thing, there do not seem to be any notable differences between genders, except for the two clearly gender-related items. This means that

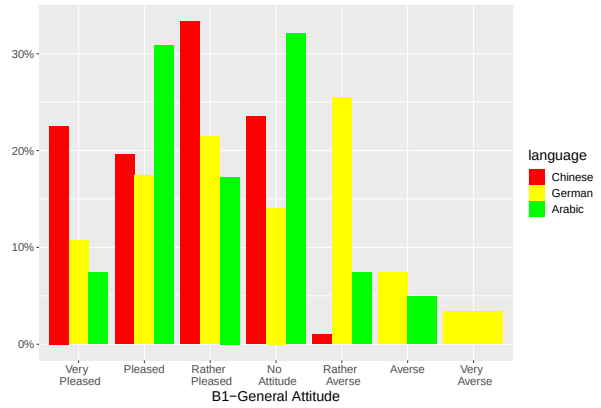

Figure 17: General attitude towards voice assistants across cultures. The Chinese have a rather positive attitude, the Egyptians' attitude is balanced, and Germans are more skeptical.

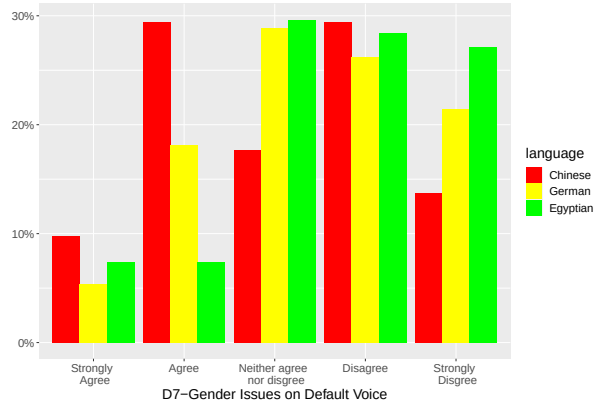

Figure 18: Answers for the gender issue across cultures. Egyptians see less problems in a female default voice.

it will not make any sense to develop gender-specific Voice Assistants. In contrast, some intercultural differences appear to exist, but they are relatively small, given many uncertainties in these judgments. We were unable to find clear regularities in these differences that would allow clear design guidelines for culture-specific Voice Assistants. Finally, there clearly exist three different types of attitudes towards emotional voice assistants, and these exist in in a cross-cultural way. This means that there is a chance for a universal design of emotional Voice Assistants which is helpful across cultures. Coincidentally, Pell et al. [52] also found that basic emotions manifest themselves in similar acoustic features across different languages. It seems that both the expression and recognition of emotion and the attitude towards VAs using this capability vary independently of language or culture.

Our cluster analysis revealed that the differences in attitudes are better characterized by splitting users into *Enthusiasts*, *Pragmatists*, and *Skeptics* instead of differentiating by culture, language or gender, but how can we use this in designing better Voice Assistants? The three clusters essentially differ in their level of agreeing with

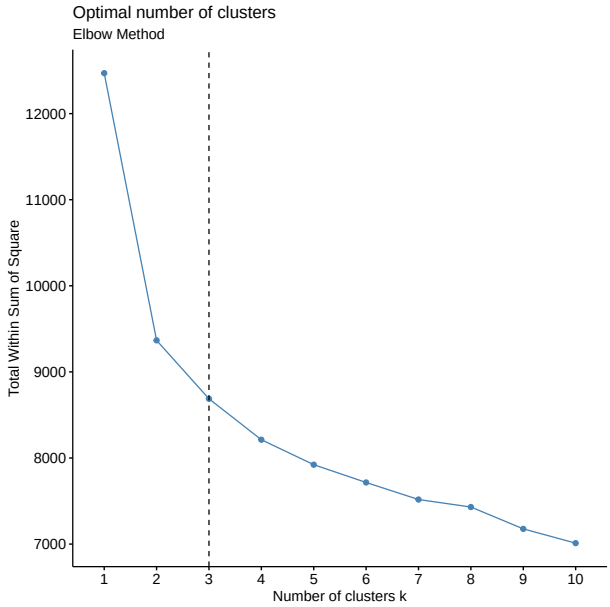

Figure 19: k-value estimation with the elbow method. The methods suggests 3 for the number of clusters.

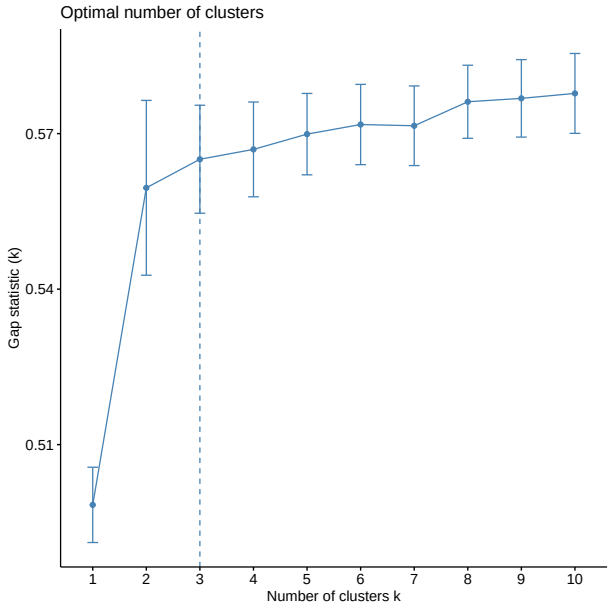

Figure 20: k-value estimation with the gap statistics. Also this method suggests 3 for the number of clusters.

their Voice Assistants having emotional features. Therefore, we probably will have to design voice assistants in a way that their emotional behavior and their emotion detection is adjustable. Ideally, we would have a scale on which we could set the level of emotionality for the assistant and one could imagine commands such as "Alexa, less emotions please".

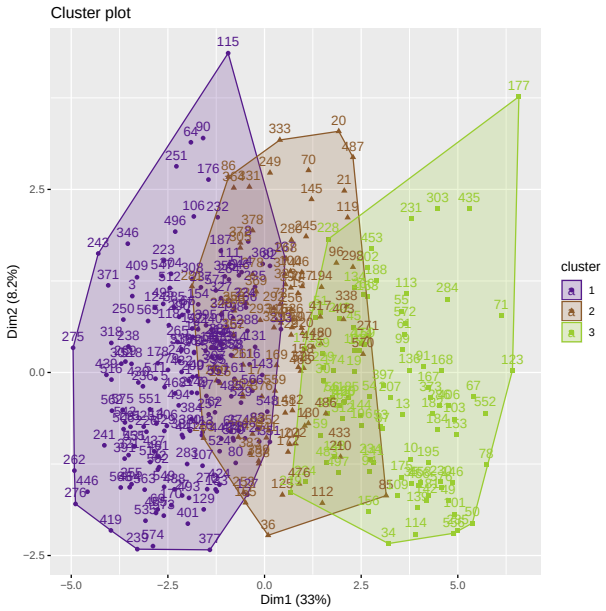

Figure 21: Visualization of the three clusters detected by K-medoids cluster detection after PCA.

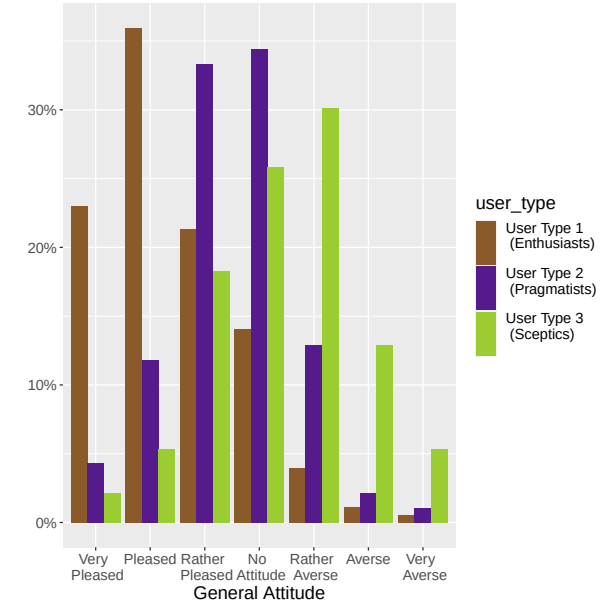

Figure 22: Distribution of general attitude (B1) by membership in a cluster. This bar chart allows to associate cluster member to attitude and gives the clusters a meaning. The three clusters represent positive, neutral and skeptical users.

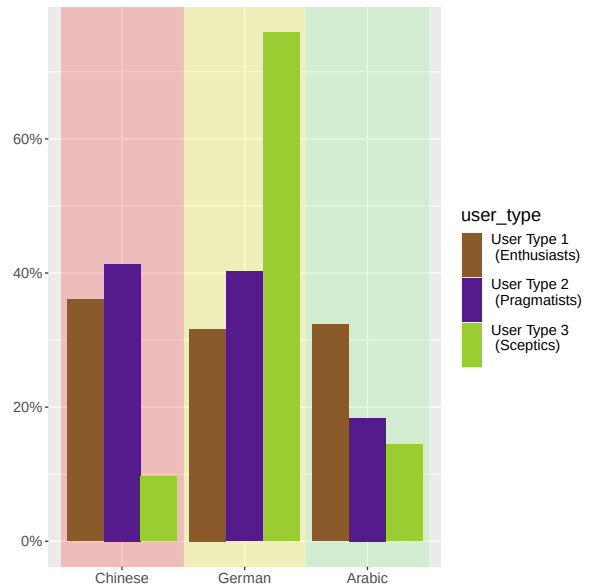

Figure 23: Cluster membership by culture. Every culture has all types of users, however the distribution varies.

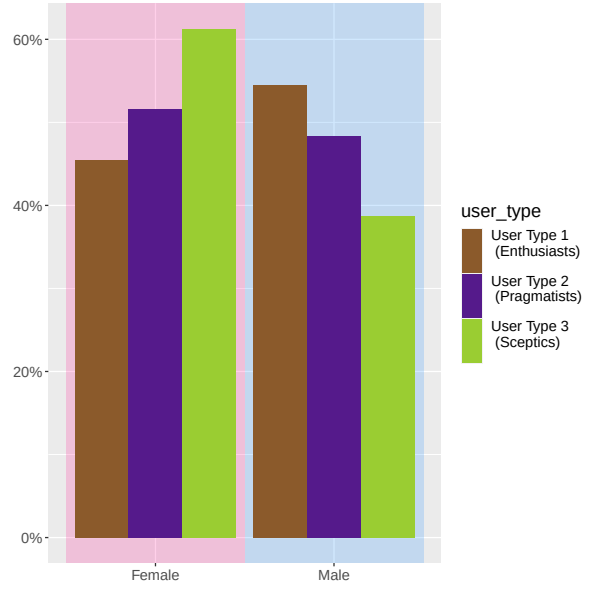

Figure 24: Cluster membership by gender. For gender, the user types are almost equally distributed.

However, we have no indication for a default level of emotionality derived from a simple user profile (e.g., language, cultural background or gender). Therefore, an improved emotion-aware VA might, for example, start with a low emotionality setting, just above neutral, and ask the user from time to time whether they like or dislike this feature. Alternatively, a direct configuration via preferences or in a startup dialog is plausible, but there is a considerable risk that this approach would not be used much, since it does not provide any immediate benefit to the user. The ultimate solution we imagine is that the VA could recognize implicitly how well its emotionality is received by its user and self-adjust to the optimal level. This will, however, require substantial future work.

Our results also leave a number of questions unanswered, and even raise new ones: How, for example, should VAs react to the emotions they detect? Should they mirror them or rather try to counteract, and if so, at which level of intensity? Recent work by Völkel et al. [64] has started to explore the desired personality of VAs, but an intercultural perspective will have to be added for scalability across different markets. In addition, it also remains open how to detect the matching personality implicitly from dialog data. Other recent work by Ma et al. [42] and Juslin et al. [31] has started to investigate the difference between real and fake emotions. It remains open which ones VAs should react to.

Finally, a Voice Assistant that competently uses emotions also has the potential to use them in a bad way and cause considerable emotional damage. This potential problem is amplified by the tailored and personal nature of a potential emotional adaption. Similar to the type of misinformation that we can observe by so-called filter bubbles in social networks, this could lead to emotional bubbles with your personal and personalized Voice Assistant. This raises the question whether potentially, some of Asimov's robot laws [2] would have to be transferred to this situation, stating, for example, that "A Voice Assistant may not emotionally harm a human being or, through inaction, allow a human being to come to harm."

6 LIMITATIONS

The reason for running our survey in different cultures was to get a better understanding of diverse user groups. However, our participants are still far from representative, obviously because we sampled just 3 cultures, but also because we recruited the participants from the authors' networks, where academics are over-represented. Another common drawback of questionnaires is the self report bias [14].

We also had a considerable portion of participants who started the questionnaire, but did not fill it completely. It is not clear why these participants quit the survey and we assume that the size of the questionnaire is to blame, but it is also entirely possible that some participants felt provoked by some of our questions. In this case we might have missed some averse opinions and in consequence there might be a general bias towards positive attitudes.

Finally, some of the questions we asked in our survey are probably deeper than they look. Participants had only a few minutes to decide about their opinion, but we found that among the authors, we could discuss for hours about most questions. It remains likely that participants were not aware of all possible consequences and implications. They might actually change their minds after having had more time to think about these implications. However, there is no simple way around this in a study that attempts to involve a larger number of participants in a manageable time frame, and the answers provided certainly have their value as a snapshot of current opinions and attitudes.

When comparing cultures, existing research also suggests that there are inherent risks in the unavoidable between-groups comparison [50]. It was stated that the differences observed in such studies may disappear with a more widespread access to the technologies studied. In addition, public opinions may change over time, which means that the same questionnaire might produce different results if conducted in several years, however, since the field of voice assistants and emotion recognition is currently developing very fast, we can only claim to provide a temporary appraisal of user attitudes towards emotion-aware voice assistants. On the positive side, we hope that our findings will help to improve future voice assistants and thereby contribute to the very development that might invalidate them.

7 CONCLUSION

Our results show that the majority of participants welcomes emotion- and personality-aware voice assistants in general (*Enthusiasts*). However, there is also a considerable fraction of users who are rather skeptical (*Skeptics*). For one thing, this indicates that attitudes are split across society and that VAs will have to meet a spectrum of attitudes and preconceptions. However, there were no clear and simple correlations between these attitudes and simple parameters, such as gender or culture alone. Of course, there was also a large group of participants with a mostly neutral opinion (*Pragmatists*).

In fact we found only very few statistically significant differences for gender. The differences we did find were mostly across cultures. However, this does not mean that people from different cultures are generally different. Actually, our cluster analysis revealed that there are different basic types of users which are present in all cultures (see Figure 23). The difference between cultures then can be attributed to the distribution of user types in each.

Another conclusion from our results is that there is no need to develop gender-specific voice assistants, as gender-specific differences were small in all our analyses. We can also conclude that it would be an inefficient approach to develop culture-specific voice assistants beyond the language aspect as, despite all differences, each culture has all kinds of users. Instead, VAs should rather be tailored to the respective basic user type. The consequence of different user types is that there will be no voice assistant which fits all. Therefore, voice assistants should be configurable to personal preferences. The range should start from voice assistants without any emotion- and personality-awareness and a slightly robotic voice up to full emotion- and personality-awareness with a very natural voice.

ACKNOWLEDGMENTS

We thank all study participants for their time and effort, as well as our anonymous reviewers for their valuable feedback. One of the authors has been funded by the German Research Foundation (DFG) project no. 425869382, and by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr [Voice of Wisdom]. Another author was funded by the China Scholarship Council (CSC).

REFERENCES

[1] Takanori Akiyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2018. Prosody-aware subword embedding considering Japanese intonation systems and its application to DNN-based multi-dialect speech synthesis. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, IEEE, New York,NY,USA, 659–664. https://doi.org/10.23919/APSIPA.2018.8659465

[2] Isaac Asimov. 1941. Three laws of robotics.

[3] Matthew P Aylett, Alessandro Vinciarelli, and Mirjam Wester. 2017. Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing* 11, 2 (2017), 361–372. https://doi.org/10.1109/TAFFC.2017.2763134

[4] Alice Baird, Shahin Amiriparian, and Björn Schuller. 2019. Can deep generative audio be emotional? Towards an approach for personalised emotional audio generation. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, IEEE, New York, NY, USA, 1–5. https://doi.org/10.1109/MMSP.2019.8901785

[5] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. *At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300270

[6] SG Brederoo, FG Nadema, FG Goedhart, AE Voppel, JN De Boer, J Wouts, S Koops, and IEC Sommer. 2021. Implementation of automatic speech analysis for early detection of psychiatric symptoms: What do patients want? *Journal of psychiatric research* 142 (2021), 299–301. https://doi.org/10.1016/j.jpsychires.2021.08.019

[7] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software* 61, 1 (2014), 1–36. https://doi.org/10.18637/jss.v061.i06

[8] Lim Kok Cheng, Ali Selamat, Mohd Hazli Mohamed Zabil, Md Hafiz Selamat, Rose Alinda Alias, Fatimah Puteh, Farhan Mohamed, and Ondrej Krejcar. 2019. Comparing the Accuracy of Hierarchical Agglomerative and K-Means Clustering on Mobile Augmented Reality Usability Metrics. In *2019 IEEE Conference on Big Data and Analytics (ICBDA)*. IEEE, IEEE, New York, NY, USA, 34–40. https://doi.org/10.1109/ICBDA47563.2019.8987044

[9] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. *What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[10] Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About face: the essentials of interaction design*. John Wiley & Sons, Hoboken, New Jersey, U.S.

[11] Donald L Day. 1998. Shared values and shared interfaces: The role of culture in the globalisation of human-computer systems. https://doi.org/10.1016/S0953-5438(97)00025-8

[12] Marieke De Mooij and Geert Hofstede. 2011. Cross-cultural consumer behavior: A review of research findings. *Journal of International Consumer Marketing* 23, 3-4 (2011), 181–192. https://doi.org/10.1080/08961530.2011.578057

[13] Chris Ding and Xiaofeng He. 2004. K-Means Clustering via Principal Component Analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, Alberta, Canada) *(ICML '04)*. Association for Computing Machinery, New York, NY, USA, 29. https://doi.org/10.1145/1015330.1015408

[14] David Dunning, Chip Heath, and Jerry M Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest* 5, 3 (2004), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

[15] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124. https://doi.org/10.1037/h0030377

[16] Vanessa Evers and Donald Day. 1997. The role of culture in interface acceptance. In *Human-Computer Interaction INTERACT'97*. Springer, Springer, Boston, MA, 260–267. https://doi.org/10.1007/978-0-387-35175-9_44

[17] Sybil Eysenck and Ahmed Abdel-Khalek. 1989. A Cross-Cultural Study of Personality: Egyptian and English Children. *International journal of psychology : Journal international de psychologie* 24 (02 1989), 1–11. https://doi.org/10.1080/00207594.1989.10600028

[18] Gerhard Fischer. 2001. User modeling in human–computer interaction. *User modeling and user-adapted interaction* 11, 1 (2001), 65–86. https://doi.org/10.1023/A:1011145532042

[19] Olaf Frandsen-Thorlacius, Kasper Hornbæk, Morten Hertzum, and Torkil Clemmensen. 2009. Non-universal Usability?: A Survey of How Usability is Understood by Chinese and Danish Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 41–50. https://doi.org/10.1145/1518701.1518708

[20] Enrique Frias-Martinez, Sherry Y Chen, and Xiaohui Liu. 2006. Survey of data mining approaches to user modeling for adaptive hypermedia. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 36, 6 (2006), 734–749. https://doi.org/10.1109/TSMCC.2006.879391

[21] Zhenye Gan, Rui Wang, Yue Yu, and Xin Zhao. 2020. Voice Conversion from Tibetan Amdo Dialect to Tibetan U-tsang Dialect Based on StarGAN-VC2. In *2020 International Conference on Big Data Economy and Information Management (BDEIM)*. IEEE, IEEE, New York, NY, USA, 184–187. https://doi.org/10.1109/BDEIM52318.2020.00049

[22] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Towards Understanding Emotional Intelligence for Behavior Change Chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, New York, NY, USA, 8–14. https://doi.org/10.1109/ACII.2019.8925433

[23] Javier Hernandez, Daniel McDuff, Xavier Benavides, Judith Amores, Pattie Maes, and Rosalind Picard. 2014. AutoEmotive: Bringing Empathy to the Driving Experience to Manage Stress. In *Proceedings of the 2014 Companion Publication on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS Companion '14)*. Association for Computing Machinery, New York, NY, USA, 53–56. https://doi.org/10.1145/2598784.2602780

[24] Morten Hertzum, Torkil Clemmensen, Kasper Hornbæk, Jyoti Kumar, Qingxin Shi, and Pradeep Yammiyavar. 2007. Usability constructs: a cross-cultural study of how users and developers experience their use of information systems. In *International Conference on Usability and Internationalization*. Springer, Springer, Berlin, Heidelberg, 317–326. https://doi.org/10.1007/978-3-540-73287-7_39

[25] Geert Hofstede. 2010. The GLOBE debate: Back to relevance. *Journal of International Business Studies* 41, 8 (2010), 1339–1346. https://doi.org/10.1057/jibs.2010.31

[26] Adrian Holliday. 2010. Complexity in cultural identity. *Language and Intercultural Communication* 10, 2 (2010), 165–177. https://doi.org/10.1080/14708470903267384

[27] Matthew B Hoy. 2018. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly* 37, 1 (2018), 81–88. https://doi.org/10.1080/02763869.2018.1404391

[28] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. 2013. *Echoes from the Past: How Technology Mediated Reflection Improves Well-Being*. Association for Computing Machinery, New York, NY, USA, 1071–1080. https://doi.org/10.1145/2470654.2466137

[29] Anil K Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters* 31, 8 (2010), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

[30] Xin Jin and Jiawei Han. 2010. *K-Medoids Clustering*. Springer US, Boston, MA, 564–565. https://doi.org/10.1007/978-0-387-30164-8_426

[31] P. Juslin, P. Laukka, and T. Bänziger. 2018. The Mirror to Our Soul? Comparisons of Spontaneous and Posed Vocal Expression of Emotion. *Journal of Nonverbal Behavior* 42 (2018), 1 – 40. https://doi.org/10.1007/s10919-017-0268-x

[32] Minna Kamppuri, Roman Bednarik, and Markku Tukiainen. 2006. The Expanding Focus of HCI: Case Culture. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles* (Oslo, Norway) *(NordiCHI '06)*. Association for Computing Machinery, New York, NY, USA, 405–408. https://doi.org/10.1145/1182475.1182523

[33] Leonard Kaufman and Peter J. Rousseeuw. 2008. *Partitioning Around Medoids (Program PAM)*. John Wiley & Sons, Inc., Hoboken, New Jersey, U.S., 68–125. https://doi.org/10.1002/9780470316801.ch2

[34] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* 7 (2019), 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

[35] Andreas M Klein, Andreas Hinderks, Maria Rauschenberger, and Jörg Thomaschewski. 2020. Exploring Voice Assistant Risks and Potential with Technology-based Users.. In *WEBIST*. ACM, New York, NY, USA, 147–154. https://doi.org/10.5220/0010150101470154

[36] Jong-Eun Roselyn Lee and Clifford I Nass. 2010. Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives*. IGI Global, Hershey, Pennsylvania, USA, 1–15. https://doi.org/10.4018/978-1-61520-901-9.ch001

[37] Yaniv Leviathan and Yossi Matias. 2018. Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

[38] Jingyi Li, Yong Ma, and Changkun Ou. 2019. Cultivation and Incentivization of HCI Research and Community in China: Taxonomy and Social Endorsements. In *CHI '19 Workshop: HCI in China*. ACM, New York, NY, USA.

[39] Kok Cheng Lim, Ali Selamat, Mohd Hazli Mohamed Zabil, Yunus Yusoff, Md Hafiz Selamat, Rose Alinda Alias, Fatimah Puteh, Farhan Mohamed, and Ondrej Krejcar. 2019. A Comparative Usability Study Using Hierarchical Agglomerative and K-Means Clustering on Mobile Augmented Reality Interaction Data. In *Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques*. IOS Press, Amsterdam, Netherlands, 258–271. https://doi.org/10.1109/ICBDA47563.2019.8987044

[40] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. https://doi.org/10.1145/3411764.3445488

[41] Zhen-Tao Liu, Abdul Rehman, Min Wu, Weihua Cao, and Man Hao. 2020. Speech personality recognition based on annotation classification using log-likelihood distance and extraction of essential audio features. *IEEE Transactions on Multimedia* 23 (2020), 3414–3426. https://doi.org/10.1109/TMM.2020.3025108

[42] Yong Ma, Heiko Drewes, and Andreas Butz. 2021. Fake Moods: Can Users Trick an Emotion-Aware VoiceBot?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–4. https://doi.org/10.1145/3411763.3451744

[43] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. *AffectAura: An Intelligent System for Emotional Memory*. Association for Computing Machinery, New York, NY, USA, 849–858. https://doi.org/10.1145/2207676.2208525

[44] Margaret McRorie, Ian Sneddon, Gary McKeown, Elisabetta Bevacqua, Etienne de Sevin, and Catherine Pelachaud. 2012. Evaluation of Four Designed Virtual Agent Personalities. *IEEE Transactions on Affective Computing* 3, 3 (2012), 311–322. https://doi.org/10.1109/T-AFFC.2011.38

[45] Batja Mesquita and Nico H Frijda. 1992. Cultural variations in emotions: a review. *Psychological bulletin* 112, 2 (1992), 179. https://doi.org/10.1037/0033-2909.112.2.179

[46] Catherine Michalopoulou and Maria Symeonaki. 2017. Improving Likert scale raw scores interpretability with K-means clustering. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 135, 1 (2017), 101–109. https://doi.org/10.1177/0759106317710863

[47] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56 (03 2000), 81–103. https://doi.org/10.1111/0022-4537.00153

[48] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[49] Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, and Alexandre Termier. 2015. Interactive User Group Analysis. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (Melbourne, Australia) *(CIKM '15)*. Association for Computing Machinery, New York, NY, USA, 403–412. https://doi.org/10.1145/2806416.2806519

[50] Daphna Oyserman. 2017. Culture three ways: Culture and subcultures within countries. *Annual Review of Psychology* 68, 1 (01 2017), 435–463. https://doi.org/10.1146/annurev-psych-122414-033617

[51] Malay K Pakhira, Sanghamitra Bandyopadhyay, and Ujjwal Maulik. 2004. Validity index for crisp and fuzzy clusters. *Pattern recognition* 37, 3 (2004), 487–501. https://doi.org/10.1016/j.patcog.2003.06.005

[52] M. Pell, S. Paulmann, Chinar Dara, Areej Alasseri, and S. Kotz. 2009. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *J. Phonetics* 37 (2009), 417–435. https://doi.org/10.1016/j.wocn.2009.07.005

[53] Tim Polzehl, Sebastian Möller, and Florian Metze. 2010. Automatically assessing personality from speech. In *2010 IEEE Fourth International Conference on Semantic Computing*. IEEE, IEEE, New York, NY, USA, 134–140. https://doi.org/10.1109/ICSC.2010.41

[54] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. *Voice Interfaces in Everyday Life*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[55] Georgios Rizos, Alice Baird, Max Elliott, and Björn Schuller. 2020. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, New York, NY, USA, 3502–3506. https://doi.org/10.1109/ICASSP40776.2020.9054579

[56] Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 5 (2018), 90–99. https://doi.org/10.1145/3129340

[57] Katie Seaborn and Jacqueline Urakami. 2021. Measuring Voice UX Quantitatively: A Rapid Review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3411763.3451712

[58] Norin Shamsuddin and Nor Mahat. 2019. *Comparison Between k-Means and k-Medoids for Mixed Variables Clustering*. Springer, Singapore, 303–308. https://doi.org/10.1007/978-981-13-7279-7_37

[59] Andy Smith and Fahri Yetim. 2004. Global human–computer systems: cultural determinants of usability. https://doi.org/10.1016/j.intcom.2003.11.001

[60] Huatong Sun. 2002. Exploring cultural usability. In *Proceedings. IEEE International Professional Communication Conference*. IEEE, IEEE, New York, NY, USA, 319–330. https://doi.org/10.1109/IPCC.2002.1049114

[61] Lee Taber and Steve Whittaker. 2018. *Personality Depends on The Medium: Differences in Self-Perception on Snapchat, Facebook and Offline*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174181

[62] Feng Tian, Xiangshi Ren, Xiangmin Fan, Wei Li, Haipeng Mi, Tun Lu, Chun Yu, and Dakuo Wang. 2019. HCI in China: Research Agenda, Education Curriculum, Industry Partnership, and Communities Building. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299005

[63] Peter Tonn, Yoav Degani, Shani Hershko, Amit Klein, Lea Seule, and Nina Schulze. 2020. Development of a Digital Content-Free Speech Analysis Tool for the Measurement of Mental Health and Follow-Up for Mental Disorders: Protocol for a Case-Control Study. *JMIR research protocols* 9, 5 (2020), e13852. https://doi.org/10.2196/13852

[64] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benajmin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445536

[65] Sarah Theres Völkel, Penelope Kempf, and Heinrich Hussmann. 2020. Personalised Chats with Voice Assistants: The User Perspective. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 53, 4 pages. https://doi.org/10.1145/3405755.3406156

[66] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376210

[67] Sarah Theres Völkel, Ramona Schödel, and Heinrich Hussmann. 2018. Designing for Personality in Autonomous Vehicles: Considering Individual's Trust Attitude and Interaction Behavior. In *Workshop "Interacting with Autonomous Vehicles: Learning from other Domains" at CHI 2018*. ACM, New York, NY, USA.

[68] Joseph Weizenbaum. 1966. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168

[69] Pamela J. Wisniewski, Bart P. Knijnenburg, and Heather Richter Lipford. 2017. Making privacy personal: Profiling social network users to inform privacy education and nudging. *International Journal of Human-Computer Studies* 98 (2017), 95–108. https://doi.org/10.1016/j.ijhcs.2016.09.006

[70] Chunhui Yuan and Haitao Yang. 2019. Research on K-value selection method of K-means clustering algorithm. *J* 2, 2 (2019), 226–235. https://doi.org/10.3390/j2020016

[71] Guanlong Zhao, Shaojin Ding, and Ricardo Gutierrez-Osuna. 2019. Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In *INTERSPEECH*. Elsevier, Amsterdam, Netherlands, 2843–2847. https://doi.org/10.21437/Interspeech.2019-1778

[72] Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Trans. Interact. Intell. Syst.* 9, 2–3, Article 10 (March 2019), 36 pages. https://doi.org/10.1145/3232077