DOES USERS' SYSTEM EVALUATION INFLUENCE THEIR SPEECH BEHAVIOR IN HCI? – FIRST INSIGHTS FROM THE ENGINEERING AND PSYCHOLOGICAL PERSPECTIVE

Ingo Siegert[1], *Matthias Busch*[1], *Julia Krüger*[2]

[1]*Mobile Dialog Systems, Institute for Information Technology and Communications, Otto-von-Guericke University Magdeburg*
[2]*Department of Psychosomatic Medicine and Psychotherapy, Otto von Guericke University Magdeburg*
ingo.siegert@ovgu.de, matthias1.busch@st.ovgu.de, julia.krueger@med.ovgu.de

Abstract: In interactions with speech based dialog systems users tend to adapt their speech behavior to their technical counterpart by taking care on the abilities and characteristics they ascribe to the system. Hence, it can be supposed, that different systems may evoke different speech behavior according to the users' evaluation of the system. In order to support this hypothesis we compared a widely similar individualization-focused interaction between users and 1) a WOZ-simulated speech-dialog system (LAST MINUTE Corpus) and 2) a self-developed skill for Amazon's Alexa. We examined measurable acoustic characteristics and subjective system evaluations operationalized by the item-based self-report questionnaire AttrakDiff. The analysis revealed that users' speech behavior showed less acoustic variation in the interaction with Amazon's Alexa. However, system evaluation did not differ significantly between the two experiments. In the discussion of this finding we, inter alia, take into account results from a first unsystematic analysis of interviews regarding users' subjective experiences conducted after the interaction with Amazon's Alexa.

1 Introduction and Related Work

Recently, it seems to become usual for users to interact with technical systems via speech. Researchers are working to further improve the interaction with technical systems. They are led by the vision of future individualized systems being able to adapt their functionality and behavior according to the user's state, including his or her situation, abilities and requirements, perceive their environment and provide a pleasant and enjoyable user experience [1]. Analyses of post-hoc user interviews conducted after an interaction with an individualization-simulating system (LAST MINUTE Corpus) as well as statistical analyses revealed hints that users tend to adapt their prosodic speech behavior to their technical counterpart by taking care on the abilities and characteristics they ascribe to the system during the interaction [3, 4]. Accordingly, in this study we want to further investigate this topic by the following research question: Does the users' evaluation of the system relate to their prosodic speech behavior?

We examine this question by using an HCI conducted for the purpose of system's individualization. If necessary user information cannot be gathered by measurement or observation, asking the user directly is in many cases the only way to get these. Such dialogs include often private and personal questions, e.g. regarding age, family or even emotional situations, and evoke various individual user reactions and experiences. A first study to analyze human-machine interaction including such an individualization-focused dialog was conducted

in 2010/2011 within the LAST MINUTE Corpus (LMC) [2]. We build up our investigation on the LMC and extend it by nearly replicating the individualization-focussed dialog in an interaction with a self-developed Alexa Skill.

2 Material and Methods

2.1 Data Corpora

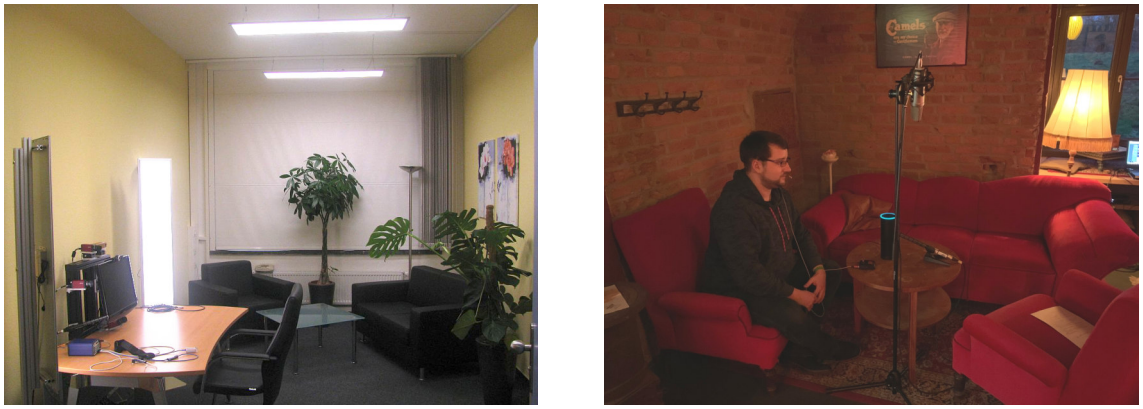

Figure 1 – Snapshots of the two data collection setups. LMC was recorded in a living room environment (left) [2]; the developed Alexa Skill was recorded in a lounge room at sa science convention (right).

2.1.1 Initial Dialog and Users' Reflection in the LAST MINUTE Corpus

The LAST MINUTE Corpus (LMC), recorded in 2010-2011, contains multimodal recordings of interactions between subjects and a speech-based dialog system [5, 6].

System design: The dialog system was simulated by Wizard-of-Oz (WOZ)-technique, represented by a machine-like male voice (MaryTTS) and a graphical interface on a computer screen without any agent. The dialog is designed in such a way that the system continuously initiated each interaction sequence. Furthermore, personal pronouns or active forms indicating a self-reference of the system are avoided. The user interacted alone in a living-room-like surrounding, see Figure 1 (left). The voice of the user is captured via high quality neckband microphone.

Interaction design: After a short self-introduction of the system, the interaction starts with a personalization module, the so-called "Initial Dialog" [7], which this study focuses on[1]. In the Initial Dialog, the system asks the user for personal, even intimate data, as part of an individualization process: After the user is requested to give and spell his or her name, the system openly asks for a self-introduction. The system summarizes all information relevant for the individualization (age, place of residence, profession, place of work, family, body height, clothing size and shoe size) and asks for a revisal and asks for missing data if necessary. Afterwards, users shall report on a recent happy and a recent annoying event, about their hobbies and their use and former experiences with technical devices. In case of very short answers the system asked for further elaboration.

Subjective users' reflection on the interaction: After the Initial Dialog as well as the other experimental modules were conducted, the users answered the AttrakDiff questionnaire, an item-based self-report for the individual evaluation of interactive products, especially usability

[1]The interested reader is referred to [7] to find out more about further experimental modules.

and design [8]. It distinguishes three aspects, the pragmatic quality (PQ), the hedonic quality (HQ), and the attractiveness (ATT). It has been used in previous interaction studies to evaluate the conversational partner, see [9, 10, 11].

Furthermore, 73 of the 130 users additionally took part in a post-hoc interview conducted by the third author [3, 13][2]. The interview focussed on the users' individual experiences during the interaction with the system, users' ascriptions to the system (e.g. aims, attitude towards the user), as well as e.g. former experiences with technical systems. The formulation of the interview questions followed the principle of openness in order to enable the participants to speak freely in an narrative mode.

Sample Description: In total, LMC includes interactions between 130 German-speaking subjects and the simulated system. In this study only subjects from whom high-quality recordings of the interviews are available are selected. This results in 24 subjects with 13 of the young group (18-28) and 11 of the elder group (>60). Sex was nearly balanced with 13 female and 11 male subjects. Regarding the computer usage, a mean value of 28.4 hours (std: 24.7 hours) was reported for the weekly usage. Furthermore, the subjects reported to use a computer since 12 years in average. It has to be noted that four participants did not use a computer at all.

2.1.2 Initial Dialog with the developed Alexa Skill and users' reflections on it

This dataset comprises interactions between subjects and Amazon's Echo Hub using a specifically developed skill. It was recorded during a science convention in Magdeburg in 2019.

System design: The technical system is represented by Amazon's Alexa using the provided male voice (Hans). The dialog is invoked via the developed skill called "Kennenlerngespräch". The participant interacted alone in a separate corner of a cosy lounge, see Figure 1 (right). One other person (third author who also conducted the post-hoc interviews) followed the interaction in a blind spot area to ensure help in case of system failures. The voice of the user is again captured via high quality neckband microphone. For the development of an Alexa skill two components are necessary [14]: an interaction model (A) and a skill back-end (B). For (A) the developer has to determine what actions a user could make by defining an intent. The developer must also specify some training sentences. They will be used by the Alexa Service to perform a Natural Language Understanding (NLU) Task. Furthermore, each developer has to provide (B) to evaluate the information contained in the request to generate a suitable reaction Alexa is providing towards the user. Information not contained in the request are not available for the developer. This includes the transcription of user's utterance and possible alternatives as well as information about the confidence of the NLU and possible alternative intents.

Interaction design: The Alexa Skill is developed on the basis of the rules for LMC "Initial Dialog" defined in LMC's Operator's Manual [7]. Thereby, it is secured that the content and course of the skill's dialog is kept similar to the LMC. The training of intents was challenging due to the large amount of open questions in LMC Initial Dialog. Therefore, necessary changes of the conversation defined in [7] had to been made.

- Instead of asking the user to provide several information at once, rephrasing the relevant attributes the system understood and asking the user for any missing information, a question for each attribute is used.
- Some interactions entail follow up questions in case of a very short answers. To solve this, the number of uttered words are counted. All utterances smaller than four words invoke a question to provide more information.

[2]The interested reader is referred to [3, 12] in order to study the results of the qualitative interview analysis.

- In case of no input within eight seconds, Alexa quits the session immediately. This period can be extended by defining a reprompt in the back-end. Thus, a reprompt to each response was added repeating the same response.
- Furthermore, due to privacy issues the Alexa skill does not ask the users for their name. After the user started the skill with its invocation phrase "Alexa starte Kennlerngespräch", he will be greeted and informed about the aim and the purpose of the following interaction by a short system's introduction similarly to the introduction in LMC.

Subjective users' reflection on the interaction: The reflection phase was arranged similar to that in LMC. After the interaction with Alexa the users answered the AttrakDiff questionnaire. Afterwards, a short version of the post-hoc user interviews used on the LMC was conducted, too, in order to get insights on the users' subjective experience and their evaluation of the interaction [13].

Sample: In total, this dataset comprises 20 interactions between German speaking subjects and Amazon's Alexa. The age ranges from 18 to 66 (mean 36 years). The sex was nearly balanced with 9 female and 11 male participants. Furthermore, a similar questionnaire for the technical experience as for the LMC is used. In order to take recent technological development into account, this questionnaire was supplemented by questions on the use of smartphones and voice assistants. Regarding the computer usage the participants reported a mean usage of 46.3 hours (std: 15.8 hours) and reported to use computers since 21 years on average. For the smartphone usage the subjects reported a mean usage of 27.2 hours (std: 16.3 hours) and a usage since 9 years (std: 3.1 years) on average. One subject reported to not have used a smartphone at all. Furthermore, 9 participants stated to have used a voice assistant before, while all participants stated to have at least heard about them.

2.2 Analysis Methods

2.2.1 Analysis of prosodic speech behavior

The speech behavior was operationalized by automatically measurable acoustic characteristics. Therefore, a statistical feature comparison of various acoustic characteristics automatically extracted by openSMILE [15] and the *emobase* set was conducted. This set comprises 988 supra-segmental features as functionals from sub-segmental speech descriptors. For the identification of changed acoustic characteristics, the feature distribution of the samples of the interaction with a technical system were compared to the distribution across all samples of the interview condition for the same dataset by applying a non-parametric U-Test. The significance level was set to $\alpha = 0.001$, as previous investigations on LMC suggest a huge variation between the human-machine interaction part and the interview part [4]. The analysis was performed independently for each subject. Afterwards, a majority voting (qualified majority: 3/4) was applied over all speakers within each dataset. The same approach has been used previously to analyze the difference in the addressee behavior of participants in various experiments, see [4, 9].

2.2.2 Analysis of users' system evaluation

In order to examine relations between the users' evaluation of both systems, we analyzed the AttrakDiff questionnaire for the LMC and the Alexa Skill. In order to compare the systems' evaluation, we used descriptive statistics, in particular, means of all items for both samples as well as sum scores of the sub-scales.

3 Results

3.1 Comparison of speech behavior in LMC-Initial Dialog and Alexa Skill-Initial Dialog

For both experiments, a significant difference for certain supra-segmental features between the human-computer interaction (HCI) and the interview part can be observed. For the LMC's Initial Dialog with 546 (55%) much more significant different features are found than for the Alexa Skill with just 133 (13%) significant different features.

Regarding a detailed feature analysis, it is apparent that *loudness* and *energy* related descriptors are only identified for LMC's Initial Dialog and not for the Alexa Skill. Another speech descriptor showing differences between both experiments is the *zero crossing rate*, which is heavily affected in the LMC's Initial Dialog and not occurring in the analysis of the Alexa Skill. This characteristic is a key feature to classify percussive sounds. But to draw conclusions for the current object of investigation a future in-depth analysis has to be performed. Regarding the spectral domain, the Alexa Skill only shows differences for higher order Mel-Frequency Cepstral Coefficients (MFCCs). These descriptors describe the spectral micro structure of the speech signal. In contrast, for LMC's Initial Dialog all MFCCs are affected suggesting a quite different speech behavior between the analyzed parts – the HCI and the interview. For the Line Spectral Pairs (LSPs) no differences are apparent within both experiments. Also the fundamental frequency is affected to a similar extend in both experiments. Analyzing the functionals, it can be seen that for the Alexa Skill mostly functionals related to the first and second order moment are affected (mean, median, range). For LMC's Initial Dialog also higher order moments (skewness, curtosis) and regression approximations are affected. It can be concluded that the subjects in the LMC's Initial Dialog show a much more divergent speech behavior between the interaction with the technical system than during the interaction with the human interview partner. Within the Alexa Skill the subjects' speech behavior is much more similar between the interaction with a technical system and the interaction with a human interview partner.

3.2 Comparison of the AttrakDiff Evaluation

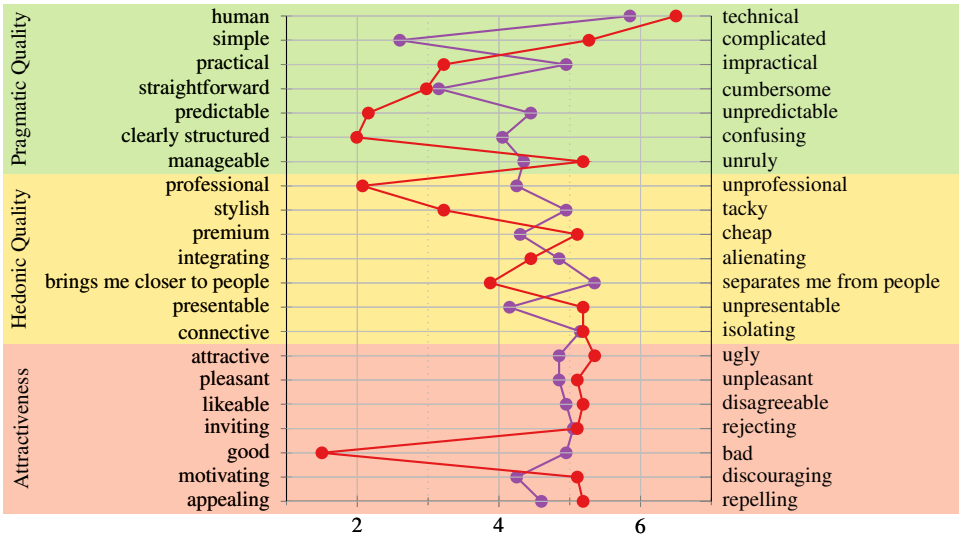

Figure 2 – Evaluation of the AttrakDiff questionnaire for the two analysed experiments: LMC (—●—) and the Alexa Skill (—●—).

The overall assessment for the dimension PQ barely differs between the two systems (LMC: 4.27,

Alexa Skill: 4.33). For the dimensions HQ (4.36 to 4.01) and ATT (3.66 to 3.97) the difference is slightly larger, but still not significant.

In some aspects, however, the assessments of the two systems differ significantly. For PQ, the LMC system is judged to be more practical, more predictable, and clearly structured. In contrast, the Alexa Skill is assessed as simpler. Furthermore, this system is assessed as less technical and less unruly. Regarding HQ, the LMC system is considered as more professional, more stylish, and less separating from people, while the Alexa Skill is evaluated as rather presentable and rather premium. Concerning ATT, the LMC system is rated as good, while the Alexa Skill is rated as rather bad. The Alexa Skill is assessed as less discouraging and less repelling than the LMC System. All other items are evaluated quite similar.

4 Discussion

In this study we examined users' system evaluation and speech behavior in comparison of two different data corpora, which entail initial individualization-focussed interactions. Our results show, that speech behavior differed significantly in a way that the participants interacting with the Alexa Skill showed less variation in speech behavior than those interacting with the system simulated in the LMC. However, the overall evaluation of the system (AttrakDiff) did not differ significantly between the two experiments, although there were differences in some of the items of AttrakDiff (especially simplicity, predictability, structuredness, professionality, and goodness).

The missing difference in system evaluation may be discussed under the light of technological development and advances in dialog design in the last years accompanied by an increasing use of voice assistants. LMC was already recorded in 2010/2011, nearly a decade ago. In this time, speech interaction possibilities with technical systems were rare and systems showed basic problems in speech recognition and processing. Speech assistants like Alexa or Siri (in the current version) were visions. The dialog design in the LMC entailed a sole system initiative throughout the Initial Dialog. The user could only react to the offers and requests of the system. The Initial Dialog for the Alexa Skill was implemented as similar as possible to the LMC. Hence, this dialog design may have met the expectations of an interaction with a voice assistant in 2010/2011, but it may not be seen as convenient nowadays, since systems like Alexa or Siri allow a more natural and reciprocal interaction and shape users' expectations towards a speech based HCI.

In line with this argumentation, differences in the evaluation of single items in AttrakDiff seem to be explainable: Participants in 2010/2011 evaluated the LMC system as much better, more professional and more stylish than those interacting with the Alexa Skill, enabling nearly the same dialog design. But LMC users evaluated it as being more complicated, although more practical, predictable and clearly structured, maybe because they were not familiar with such speech based interaction.

However, speech behavior differed significantly in a way that participants interacting with the Alexa Skill showed less variation in speech behavior than those interacting with the system simulated in the LMC. Since our results do not underline the hypothesis that the system evaluation (operationalized by AttrakDiff) may explain these differences, other reasons have to be discussed. One reason, of course, may lay in the increasing use of voice assistants. Maybe users have "learned" that technical systems nowadays are able to "understand" them although they do not speak in an emphazised way. Additionally, first unsystematic analysis of the interviews conducted after the interaction with the Alexa Skill provide insights into the subjective experiences of the interaction with it. These indicate that users may be less motivated to engage in enabling a smooth interaction with the Skill because of different reasons. For example,

they reported that they experienced the interaction as "cold heartlessly structured according to system a" (V01[3]) and "very very very impersonal" (V01), which is explained by the content of the system's questions and the missing reference to the user's answers or the premature termination of user's answer by the system ("I didn't feel noticed", V10). Furthermore, the system is clearly experienced as a machine, which is "not a bit human" (V02). Many participants reported on the fear of possible data misuse ("data gobbling [..] that's the basic problem I have with these devices today [..] it's quite scary to be overheard", V08). This leads to the reported tendency of trying to transport as few emotions as possible through speech ("I didn't want to let any emotions flow into it [..because] I don't know to what extent the machine can use this", V02) and limitations of self-disclosure ("superficial answers", V02). Feelings underlying these experiences, amongst them annoyance and scepticism, may have contributed to the observed variation in speech behaviour ("at the beginning I still liked it to talk to the system and at the end I didn't want it anymore (...) my formulations have become shorter, more concise, more precise, but also reserved", V16).

Although we designed the Alexa Skill experiment to be as similar as possible to the LMC Initial Dialog, there a some limitations, which may have biased the answers in AttrakDiff. Participants interacting with the Alexa Skill were recruited during a convention about objectivity of future AI. It can be assumed that this resulted in a bias of the sample compared to the LMC sample: Convention visitors may of course be interested in technology, but they may furthermore be rather critical, especially regarding topics of data storage and reuse. Furthermore, in LMC a WOZ-system was used, therefore it was possible to intervene directly and fewer interaction disturbances due to system errors occurred than in the Alexa Skill Initial Dialog. Moreover, in LMC the AttrakDiff-rating was conducted after all experimental modules were completed. Hence, the modules following the Initial Dialog influenced the answers.

A stronger incorporation of the interview contents (subjective experience, what emerges besides the information in AttrakDiff, which motives behind speech behavior) will be part of our future work. Also a comparison to an Initial Dialog considering the specific capabilities of modern voice assistants is under discussion.

References

[1] VALLI, A.: *Notes on natural interaction*. Tech. Rep., University of Florence, Italy, 2007.

[2] FROMMER, J., D. RÖSNER, R. ANDRICH, F. RAFAEL, S. GÜNTHER, M. HAASE, and J. KRÜGER: *LAST MINUTE: An Empirical Experiment in User-Companion Interaction and its evaluation*, pp. 253–276. Springer International Publishing, Cham, 2017.

[3] KRÜGER, J.: *Subjektives Nutzererleben in der Mensch-Computer-Interaktion: Beziehungsrelevante Zuschreibungen gegenüber Companion-Systemen am Beispiel eines Individualisierungsdialogs*. Qualitative Fall- und Prozessanalysen. Biographie – Interaktion – soziale Welten. Verlag Barbara Budrich, 2018. URL `https://books.google.de/books?id=v6x1DwAAQBAJ`.

[4] SIEGERT, I., T. SHURAN, and A. F. LOTZ: *Acoustic addressee-detection – analysing the impact of age, gender and technical knowledge*. In W. M. ANDRE BERTON, UDO HAIBER (ed.), *Elektronische Sprachsignalverarbeitung 2018. Tagungsband der 29. Konferenz*, vol. 90 of *Studientexte zur Sprachkommunikation*, pp. 113–120. TUDpress, Ulm, Germany, 2018.

[3] We used individual codes in order to mark utterances of the participants.

[5] RÖSNER, D., J. FROMMER, R. FRIESEN, M. HAASE, J. LANGE, and M. OTTO: *LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions*. In *Proc. of the 8th LREC*, pp. 96–103. Istanbul, Turkey, 2012.

[6] PRYLIPKO, D., D. RÖSNER, I. SIEGERT, S. GÜNTHER, R. FRIESEN, M. HAASE, B. VLASENKO, and A. WENDEMUTH: *Analysis of significant dialog events in realistic human-computer interaction*. *Journal on Multimodal User Interfaces*, 8, pp. 75–86, 2014.

[7] FROMMER, J., D. RÖSNER, M. HAASE, J. LANGE, R. FRIESEN, and M. OTTO: *Detection and Avoidance of Failures in Dialogues – Wizard of Oz Experiment Operator's Manual*. Pabst Science Publishers, Lengerich, 2012.

[8] HASSENZAHL, M., M. BURMESTER, and F. KOLLER: *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität*. In G. SZWILLUS and J. ZIEGLER (eds.), *Mensch & Computer 2003*, vol. 57 of *Berichte des German Chapter of the ACM*, pp. 187–196. Vieweg+Teubner, Wiesbaden, Germany, 2003.

[9] SIEGERT, I. and J. KRÜGER: *How do we speak with alexa - subjective and objective assessments of changes in speaking style between hc and hh conversations*. *Kognitive Systeme*, 1, 2019.

[10] SIEGERT, I., J. NIETZOLD, R. HEINEMANN, and A. WENDEMUTH: *The restaurant booking corpus - content-identical comparative human-human and human-computer simulated telephone conversations*. In W. M. ANDRE BERTON, UDO HAIBER (ed.), *Elektronische Sprachsignalverarbeitung 2019. Tagungsband der 30. Konferenz*, vol. 90 of *Studientexte zur Sprachkommunikation*, pp. 126–133. TUDpress, Dresden, Germany, 2019.

[11] SIEGERT, I. and J. KRÜGER: *Speech melody and speech content didn't fit together" – differences in speech behavior for device directed and human directed interactions*. In *Advances in Data Science: Methodologies and Applications*. Springer International Publishing, 2020. In print.

[12] KRÜGER, J., M. WAHL, and J. FROMMER: *"es is komisch es is keen mensch" – zuschreibungen gegenüber individualisierten technischen assistenzsystemen. eine interviewstudie zum nutzer/innenerleben in der mensch-computer-interaktion*. *ZQF – Zeitschrift für Qualitative Forschung*, pp. 233–251, 2018.

[13] LANGE, J. and J. FROMMER: *Subjektives Erleben und intentionale Einstellung in Interviews zur Nutzer-Companion-Interaktion*. In *Proceedings der 41. GI-Jahrestagung*, vol. 192 of *Lecture Notes in Computer Science*, pp. 240–254. Bonner Köllen Verlag, Berlin, Germany, 2011.

[14] KUMAR, A., A. GUPTA, J. CHAN, S. TUCKER, B. HOFFMEISTER, M. DREYER, S. PESHTERLIEV, A. GANDHE, D. FILIMINOV, A. RASTROW ET AL.: *Just ask: Building an architecture for extensible self-service spoken language understanding*. *arXiv preprint arXiv:1711.00549*, 2017.

[15] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor*. In *Proc. of the ACM MM-2010*. 2010.

[16] KRÜGER, J., M. WAHL, and J. FROMMER: *Making the system a relational partner: Users' ascrip-tions in individualization-focused interactions with companion-systems*. In *Proc. of the 8th CENTRIC 2015*, pp. 48–54. Barcelona, Spain, 2015.