How do we speak with ALEXA – Subjective and objective assessments of changes in speaking style between HC and HH conversations

Ingo Siegert*, Julia Krüger**

* *Institute for Information and Communications Engineering, Otto von Guericke University Magdeburg 39106 Magdeburg, Germany (e-mail: ingo.siegert@ovgu.de)*
** *Department of Psychosomatic Medicine and Psychotherapy, Otto von Guericke University Magdeburg, 39120 Magdeburg, Germany (e-mail: julia.krueger@med.ovgu.de)*

Abstract: Nowadays a diverse set of technical solutions is implemented to detect if a system should react to an uttered speech command. Unfortunately, the preferred methods of wake words can result in confusions e.g. when the word has been said but no interaction with the system was intended by the user. Therefore, technical systems should be able to detect their addressing by itself. In order to achieve this goal research concentrates on analyzing speech. Analysing the speaker's self-assessment of his speech characteristics while addressing a system can provide further information, which up to now wasn't considered in the field. Utilizing a new generated voice assistant conversation corpus, this paper presents insights of the participant's addressee behavior and correlates objective and subjective changes in speaking style characteristics between human-human and human-computer conversations. It could be shown that users could recognize changes in some of their speech characteristics. Furthermore, the objective identifiable changes are heavily dependent on the type of interaction. Mostly affected are intonation and stress patterns as well as melody and rhythm patterns. The presence of a confederate speaker does not reveal differences on the addressing behavior.

Keywords: Addressee Detection, Speech Assistant, Multi-Scenario, Multi-User, Speaking-Style

1. INTRODUCTION

Voice assistant systems recently receive increased attention, as the market for commercial voice assistants is rapidly growing: e.g. Microsoft Cortana had 133 million active users in 2016 (cf. Osborne (2016)), the Echo Dot was the best-selling product on all of Amazon in the 2017 holiday season (cf. Dickey (2017)). Furthermore, 72% of people who own a voice-activated speaker say their devices are often used as part of their daily routine (cf. Kleinberg (2018)).
The attractiveness of today's voice assistants is based on their ease of use. Using nothing but speech commands, users can play music, search the web, create to-do and shopping lists, shop online, get instant weather reports, and control popular smart-home products.
Besides making the operation of a technical system as simple as possible, voice assistants should also enable a natural interaction. This type of interaction is characterized by systems that understand natural actions and engage people in a dialog, while allowing them to interact naturally with each other and the environment. Users don't need to wear any device or learn any instruction, as the interaction respects the human perception. Accordingly, the interaction with such systems is easy and seductive for everyone (cf. Valli (2007)). To fulfill these properties, cognitive systems, which are able to perceive their environment and are working on the basis of gathered knowledge and model-based recognition, are needed. In contrast, today's voice assistant's system functionality is still very limited. Although promoted as assistants with at least rudimentary cognitive skills.
Another aspect that still needs improvement is to automatically recognize the addressee of a user's utterance. Nowadays several solutions are implemented to detect if a system should react to an uttered speech command, particularly push-to-talk inputs and wake words [1]. In addition to the unnaturalness of these solutions in the above sense, the currently preferred wake word method is error-prone. It can result in users' confusion, e.g. when the wake word has been said but no interaction with the system was intended by the user. Especially for voice assistant systems that are already able to buy products automatically and in future should be enabled to autonomously make decisions it is crucial to only react when intended by the user.
The following examples illustrate that today's solution of using a wake word is in many ways insufficient. At the end of a San Diego news story the anchorman remarked: "I love the little girl, saying 'ALEXA order me a dollhouse."' Amazon Echo owners who were watching the broadcast found that the remark triggered orders on their own devices (cf. Liptak (2017)). Another recent addressee detection failure high-

[1] The wake word to activate Amazon's ALEXA from its 'inactive' state to be able to make a request is 'Alexa' by default.

lights the privacy issues of these smart devices. According to the KIRO7 news channel, a private conversation of a family was recorded by Amazon's ALEXA and sent to the phone of a random person, who was in the family's contact list. Amazon justified this misconduct as follows: ALEXA woke up due to a word in the background conversation sounding like 'ALEXA', the subsequent conversation was heard as a "send message" request, the customer's contact name and the confirmation to send the message (cf. Horcher (2018)). Besides the given examples there are further examples of malfunctions for Google Now as well, see Tilley (2017).
Thus, additional techniques are needed to detect whether the voice assistant is addressed or not. One possibility is the development of a reliable addressee detection implemented in the system itself.
Regarding systems addressee detection various aspects have been investigated so far, cf. Section 2. Previous research concentrated on analyzing observable users' speech characteristics in the recorded data as well as subsequent analyzes and external ratings. The question whether users themselves recognize differences or even perhaps deliberately change their speaking style when interacting with a technical system (and potential influencing factors for this change) is a matter of basic research which has not been investigated so far. Although it could be shown, that there are differences in the speaking style of users in human-human interaction (HHI) and human-computer interaction (HCI), up to now, a comparison between objectively measurable differences and users' subjectively recognized differences is missing.
The aim of the current study is to identify changes in speaking style by analyzing modifications of features during a multi-party HCI and to investigate whether this change is an explicit or implicit one (degree of awareness). Furthermore, the influence of the type of interaction and of the presence of a second speaker will be investigated, too, because it is assumed that these factors influence the addressee behavior. In this connection, the following research questions will be answered: 1) Do users themselves recognize differences in the interaction with a technical system compared to interacting with another person? 2) How do humans speak with current speech-based assistant systems? 3) Do the differences in the observed and/or reported interaction style differ between a formal and an informal interaction setting? 4) Which differences in the speaking style during the interaction with the technical system can be observed when users interact alone or together with a confederate speaker? The answers to these questions supports the understanding of changes in the speaking style during different addressing task as well as the identification of influencing factors on the addressee behavior.
The remainder of the paper is structured as follows: In Section 2 previous work on addressee detection is discussed. In Section 3 the experimental setup of the utilized dataset and the participant description are presented. In Section 4 the subjective and objective analysis methods are introduced. The results are then presented in Section 5. Finally, Section 6 concludes the paper and presents an outlook.

2. RELATED WORK

Most authors use either eye-gaze, or language related features (utterance length, keyword, trigram-model), or a combination of both. But, as current voice assistant systems are speech activated only, only related work considering the acoustic channel are reported. Addressee detection studies for speech enabled systems utilize self-recorded databases either with one human and a technical system or groups of humans (mostly two) interacting with each other and a technical system (cf. Shriberg et al. (2012); Vinyals et al. (2012); Tsai et al. (2015); Shriberg et al. (2013); van Turnhout et al. (2005)) or teams of robots and teams of humans (cf. Dowding et al. (2006)). These studies are mostly done using one specific scenario. Just a few researchers analyze how people interact with technical systems in different scenarios (cf. Lee et al. (2013); Baba et al. (2012)). In these studies, the technical system is either a robot (cf. Dowding et al. (2006); Katzenmaier et al. (2004)), a research system (cf. Shriberg et al. (2012); Vinyals et al. (2012)), or a Wizard-of-Oz (WOZ)-system (cf. van Turnhout et al. (2005)). To the best of our knowledge, a current commercial system has not been used so far to examine addressee detection.
Regarding acoustic addressee recognition systems, researchers employ different tasks, as there are no generally accepted benchmark data. In Tsai et al. (2015), 150 multiparty interactions of 2 to 3 people playing a trivia question game with a computer are utilized. The dataset comprises audio, video, beamforming, system state and ASR information. For acoustic analyzes, energy, energy change and temporal shape of speech contour features, in total 47 features, are used to train an adaboost classifier. The authors achieved 13.88% Equal Error Rate (EER).
In Shriberg et al. (2012), data of 38 sessions of two people interacting in a more formal way with a "Conversational Browser" are recorded. Using energy, speaking rate as well as energy contour features to train a Gaussian Mixture Model (GMM) together with linear logistic regression and boosting, the authors achieved an EER of 12.63%. The same data is used in Shriberg et al. (2013). Their best acoustic EER of 12.5% is achieved using a GMM with adaptive boosting of energy contour features, voice quality features, tilt features, and voicing onset/offset delta features.
Baba et al. (2012) used two different experimental settings (standing and sitting) with 10 times two speakers interacting with an animated character. The experimental setup was about two decision-making sessions with formalized commandos. They employed a Support Vector Machine (SVM) and four supra-segmental speech features (F_0, intensity, speech rate and duration) as well as two speech features describing the difference for a speaker from all speakers's average for F_0 and intensity. The reported acoustic accuracy is 75.3% for the participants standing and 80.7% for the participants sitting.
In Batliner et al. (2009) the authors investigate Off-talk vs. On-talk situations, where Off-talk comprises all utterances not directed towards the system. This includes reading, thinking aloud and speaking to other people. As database SmartKom and SmartWeb are utilized. The authors used a highly redundant feature-set comprised from 100 prosodic features (duration, energy, F0, jitter) and 30 POS features. Their best result in recognizing Off-Talk vs. On-Talk is 73.7% Unweighted Average Recall (UAR). Afterwards, they also analyzed the importance of specific features and identified duration for read speech, energy for On-Talk. Furthermore, in Siegert et al. (2018b) it could be shown that an addressee detection system based on acoustic features only achieves an outstanding classification performance

(> 84%), also for inter-speaker groups across age, sex and technical affinity using data from a formal computer interaction (Prylipko et al., 2014) and a subsequently conducted interview representing a HHI (Lange and Frommer, 2011). A very recent work by researchers of AMAZON (cf. Mallidi et al. (2018)) uses long short-term memory neural networks trained on acoustic features, ASR decoder, and 1-best hypotheses of automatic speech recognition output with an EER of 10.9% (acoustic alone) and 5.2% combined for the recognition of device directed utterances. As dataset natural 250 hours (350k utterances) of human interactions with voice controlled far-field devices are used for training. The so far reported research concentrated on analyzing observable users' speech characteristics in the recorded data. Regarding research on how humans identify the addressee during interactions, most studies rely on visual cues (eye-gaze) and lexical cues (markers of addressee), cf. (Jovanovic et al., 2006; Zhang et al., 2016; Beyan et al., 2016). Only few studies analyze acoustic cues.

In Terken et al. (2007) the human classification rate using auditory and visual cues is analyzed. Therefore, Terken et al. (2007) analyzed conversations between a person playing as a clerk of travel agency and two people playing as customers. Furthermore, the authors reported that the tone of voice was useful for human evaluators to identify the addressee in their face-to-face multiparty conversations. Analyzing the judgments of human evaluators in correctly identifying the addressee, the authors stated that the combination of audio and video presentation gave the best performance of 64.2%. But both auditory and visual information alone resulted in a somewhat poorer performance of 53.0% and 55.8%, respectively. Both results are still well above chance level, which was 33.3%.

The authors of Lunsford and Oviatt (2006) investigated how people identified the addressee in human-computer multiparty conversations. To this avail, the authors recorded videos of three to four people sitting around a computer display. The computer system answered questions from the users. Afterwards human annotators were asked to identify the addressee of the human conversation partners by watching the videos. Additionally the annotators should rate the importance of lexical, visual and audio cues for their judgment. The list of cues comprise fluency of speech, politeness terms, conversational/command style, speakers' gaze, peers' gaze, loudness, careful pronunciation, and tone of voice. An overall correct judgment of 63% identifying the human or computer addressee was reported with 86% correctness in identifying the computer as addressee. This emphasizes the difficulty of the addressee detection task. The authors furthermore reported that both audio and visual information are useful for humans to predict the addressee even when both modalities – audio and video – are present. The authors additionally stated that the subjects performed the judgment significantly faster based on the audio information than on the visual information. Regarding the importance of the different cues, the most informative cues are intonation and speakers' gaze (cf. Lunsford and Oviatt (2006)). The human performance of 86% in correctly identifying the system as addressee are in line with the performance of automatic approaches.

In summary, the studies so far examined identified acoustic cues as meaningful as visual cues (gaze direction) for human evaluation. But these studies analyzed only a few acoustic characteristics. Furthermore, it must be stated that previous studies are based on the judgments of evaluators, never on the statements of the interacting speakers themselves. Although, the fact that users can be aware of speaking differently with technical systems compared to speaking with humans has already been described in Frommer et al. (2017). Thus, the question whether users themselves recognize differences or even perhaps deliberately change their speaking style when interacting with a technical system has not been evaluated. Thereby, the conducted study relies on both psychometric questionnaires with predefined answers and questionnaires with open questions allowing a subjective relevance setting of the participants.

3. THE VOICE ASSISTANT CONVERSATION CORPUS (VACC)

To analyze the speakers' behavior during a multi-party HCI, the Voice Assistant Conversation Corpus (VACC) was developed, see Siegert et al. (2018a). VACC contains recordings of 27 German speaking participants, all students at the Otto von Guericke University Magdeburg. The aim was to design realistic interactions of comparable human-machine and human-human interactions. Table 1 summarizes the dataset characteristics.

Subjects/Experiments	27
Sex	Male 13 / Female 14
Total Recorded Data	17 h 07 min
Experiment Duration	Mean: 31 min
Age (years)	Mean 24 (Std: 3.32) Min: 20; Max: 32
Language	German

Table 1. Dataset Characteristics

3.1 Experimental Design

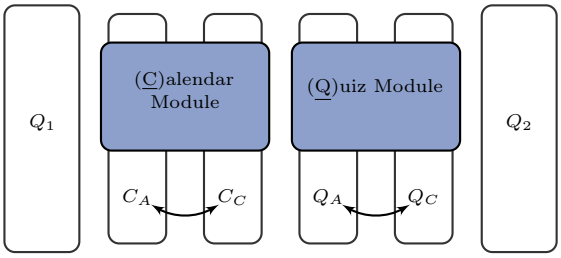

Fig. 1. A sketch of the experimental procedure. Q_1 and Q_2 are the questionnaire rounds. The order of the scenarios (Calendar Module and Quiz Module) is fixed. A and C denote the experimental conditions alone and together with an confederate respectively.

VACC consists of recordings of interaction experiments between the participant and a confederate speaker [2] and Amazon's ALEXA. Additionally, questionnaires presented before and after the experiment are used to get insights about the speakers' addressee behavior, see Fig. 1. The initial instruction of the experiment entailed information about the basic capabilities and the wake word-based addressing of ALEXA. Then, questionnaire round Q_1 was conducted. Two experimental modules followed, arranged according to their complexity level. There were two conditions for each module, which were permuted for

[2] The second speaker was introduced to the participants as "Jannik".

different participants. Thus, each experiment contained four "rounds". A round was finished when the aim was reached or broken up to avoid frustration if hardly any success could be realized. The confederate speaker's role was to only interact with the participant and discuss possible strategies. Afterwards questionnaire round Q_2 was conducted. Although, the proscribed role of the confederate is distinct from that of ALEXA, it was decided for such an attempt to gather natural interactions, as they would occur in daily life when using speech-enabled assistants.

Questionnaire Round 1 In Q_1 a short form of a self-defined questionnaire used in Rösner et al. (2012) was utilized to obtain socio-demographic information as well as information about the participants' experience with technical systems.

Module 1 ("Calendar Module"): In this formal interaction the participant had to make two times three appointments in two pre-defined weeks with the confederate speaker. The participant's calendar was stored online and was only accessible via ALEXA. The participants were instructed that ALEXA could give information about the calendar on request including exemplary commands. In condition C_A ("alone") the participant only got a written information about the confederate's available dates. The participant had to interact with ALEXA alone and could interact in its own manner without a supervision. In condition C_C ("with confederate") the confederate speaker entered the room and could give the information by himself. Thus, the participant had to ask both, ALEXA and the confederate to find available time slots. The confederate speaker was part of the research team and was instructed to interact only with the participant, not with ALEXA.

Module 2 ("Quiz Module"): In this interaction the participant had to answer questions of a quiz (e.g., "How old was Albert Einstein?"). During the explanation of this module, the participants were instructed that ALEXA is not able to give the full answer, but could offer support by solving partial steps or by answering a reformulated question. In condition Q_A the participant had to solve the quiz on its own. In condition Q_C the participant and the confederate speaker built up a team to answer the questions. Thus, these conversations were more informal than the previous calendar task. The confederate (here again only interacting with the participant, not with ALEXA) was instructed to make command proposals to the participant if frustration due to failures was imminent. The quiz in Q_C was more sophisticated than in Q_A to force cooperation between the two speakers and ALEXA.

Questionnaire Round 2 After the experiment, self-defined computer-aided questionnaires were applied (Q_2 in Fig. 1). The first two of them focused on participants' experiences regarding a) the interaction with the voice assistant and the confederate speaker in general, b) possible changes in voice and speaking style while interacting with the voice assistant and the confederate speaker. The second questionnaire asked for recognized differences in the prosodic characteristics monotony, melody, and intonation.
According to the so-called principle of openness in examining subjective experiences (cf. Hoffmann-Riem (1980)), the formulation of questions developed from higher openness and a free, non-restricted answering format in the first questionnaire to lower openness and highly structured answering formats in the second questionnaire. This structure allowed to examine the degree of participants' awareness of changes in the their voice and speaking style: If they already describe changes in some features (e.g. melody or speed) according to the open, initial questions, a higher degree of awareness is indicated than if they report about differences regarding these features only when they are explicitly asked for in the closed questions. The terms used in these questionnaires are chosen in such a way that layman have an idea about them. A third questionnaire focused on previous experiences with voice assistants. Lastly, AttrakDiff, see Hassenzahl et al. (2003), was used to supplement the open questions on self-evaluation of the interaction by a quantifying measurement of the quality of the interaction with the voice assistant (hedonic and pragmatic quality). For results on AttrakDiff, see Siegert et al. (2018a).

3.2 Recording Setup

Fig. 2. A snapshot of the data collection setup. The confederate speaker (left side) and the participant (right side) are sitting around a table, where the voice assistant (Amazon ALEXA Echo Dot) is located.

The recordings were conducted in a living room-like surrounding, see Fig. 2. The aim of this setting was to enable the participant to get into a natural communication atmosphere (in contrast to the distraction of laboratory surroundings). The participant sat on the sofa (right side of the photo in Fig. 2) and interacted with the voice assistant system, placed on the table in the middle. The confederate speaker (present only in the two-person variants of each scenario) sat on the armchair (left side of the photo in Fig. 2). The positions were identical for all recordings of all participants to ensure comparability.
As voice assistant system, the Amazon ALEXA Echo Dot (2nd generation) was utilized. It was decided to use this system to create a fully free interaction with a currently available commercial system. For this dataset, video records were declined, because current commercial systems – in the focus of this study – do not have video recordings as well. In addition to the requirement of depicting current systems, for research purposes also privacy issues arise. The participants' awareness of video recordings has the danger that the participants behave differently and leading to a possible distortion of a proper speaking style analysis.
Two high-quality neckband microphones (Sennheiser HSP 2-EW-3) were used to capture the voices of the participant and the confederate speaker. Additionally a high-quality shotgun microphone (Sennheiser ME 66) captured the overall acoustics and the output of Amazon's ALEXA.

The recordings were stored in WAV-format with 44.1 kHz sample rate and 16 bit resolution. Afterwards the recordings were manually separated into utterances with additional information about the belonging speaker (participant, confederate speaker, ALEXA). Furthermore a manual labeling was conducted to identify the addressee of each utterance – HHI for statements addressed to the confederate, HCI for statements directed to ALEXA. Additionally, off-talk (OT) for all statements not directed towards a specific speaker and soliloquized parts as well as cross-talk (CT) for parts were the turns of both humans or of the participant and ALEXA overlap were highlighted.

3.3 Participant characterization

All participants were German speaking students. The corpus is nearly balanced regarding sex (13 male, 14 female). The mean age is 24.11 years, ranging from 20 to 32 years. Furthermore, the dataset is not biased towards technophilic students, as different study courses are covered, including computer science, engineering science, humanities and medical sciences.
The participants reported to have at least heard of Amazon's ALEXA before. When asked about experience with ALEXA, only six participants specified that they had used ALEXA prior to this experiment. Five of them used ALEXA rarely for testing, only one participant specified that he uses ALEXA regularly – for playing music. Regarding the experience with other voice assistants, additional ten participants indicated a prior use. As voice assistants, they indicated Apple SIRI, GOOGLE NOW, or Microsoft CORTANA. Seven of them used these voice assistants seldom or just tried it once. Only three reported to use them on a regular basis, e.g. for programming a timer. Thus, in total 16 out of 27 participants reported to have at least basic experience with voice assistants. The nine participants not using any voice assistant before also reported a mistrust in the necessity of voice control and expressed data protection concerns when asked for reasons. Overall this dataset represents a heterogeneous set of participants, which is representative for younger users with an academic background.

4. ANALYSES'S METHODS

4.1 Subjective Evaluation

The collected questionnaires were used to examine subjective reflections on the interaction. Therefore, the first two questionnaires of Q_2 were analyzed. These questionnaires asked for experiences regarding the interaction and the awareness of changes in participants' voice and the participant's speaking style during the interaction.
The analysis of participants' answers concentrated on the questions dealing with subjectively recognized changes in voice and speaking style while interacting with ALEXA and the confederate speaker. The answers were analyzed using qualitative content analysis, see Mayring (2014), in order to summarize the material sticking close to the text. At the beginning, the material was broken down into so-called meaning units. These are text segments, which are understandable by themselves, represent a single idea, argument or information and vary between word groups and text paragraphs in length (cf. Tesch (1990)). These meaning units were paraphrased, generalized, and reduced in accordance with the methods of summarizing qualitative content analysis. Afterwards, they were grouped according to similarities and differences across all participants.
Qualitative research aims at maximizing the variance, in contrast to quantitative approaches aiming at minimizing the variance to gain representative characteristics. Consequently, it was aimed to explore the variance of individual experiences and individual perceptions regarding the users' own voice and speaking style.

4.2 Objective Evaluation

The objective evaluation, on the other hand, is based on statistical comparisons of acoustic characteristics. Acoustic characteristics were automatically extracted using openSMILE (cf. Eyben et al. (2010)). As the related work does not indicate specific feature sets distinctive for addressee detection, a broad set of features extractable with openSMILE was utilized. This set of features has been successfully used in various applications: dialog performance (cf. Ramanarayanan et al. (2017)), acoustic scene classification (cf. Marchi et al. (2016)), user satisfaction (cf. Egorow et al. (2017)), humor prediction (cf. Bertero and Fung (2016)), spontaneous speech (cf. Toyama et al. (2017)), physical pain detection (cf. Oshrat et al. (2016)), emotion recognition (cf. Böck et al. (2017); Eyben et al. (2016)), and addressee detection (cf. Siegert et al. (2018b)).
For feature extraction, it is differentiated between Low-Level-Descriptors (LLDs) and functionals. LLDs comprise the sub-segmental acoustic characteristics extractable for a specific short-time window (usually 25-40ms), while functionals represent super-segmental contours of the LLDs regarding a specific cohesive course (usually an utterance or turn). In Table 2b, the used LLDs and functionals are shortly described. For reproducibility, the same feature identifiers as supplied by openSMILE are used.
To identify changes in the acoustic characteristics, statistical analyzes were conducted utilizing the previously automatically extracted features. To this avail, for each feature the distribution across the samples of the e.g. HCI condition were compared to the distribution across all samples of the e.g. HHI condition by applying a non-parametric U-Test. The significance level was set to $\alpha = 0.01$. This analysis was performed independently for each speaker of the dataset. Afterwards, a majority voting (qualified majority: 3/4) of the analyzed features was applied over all speakers within each dataset. Features with a p-value below α in the majority of the speakers are identified as changed between the compared conditions.
In order to substantiate the informative value of the identified differences, an additional classification task was carried out. A two-class problem has been formulated, utilizing a random-forest (RF)-classifier with a Leave-One-Speaker-Out (LOSO) validation method. This method does reflect reality better than a simple cross-fold validation, by training on the acoustic characteristics of known speakers and testing on acoustic characteristics of unknown speakers. As performance measure the F1-score as the harmonic average of the precision and recall is reported. In the case of the LOSO validation the F1-score is the harmonic mean of the unweighted averaged recall and the unweighted averaged precision over all speakers.

Name	Description
alphaRatio	Ratio between energy in low frequency region and high frequency region
F0	Fundamental frequency
F0_env	Envelope of the F0-contour
F0semitone	Logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz
FXamplitude	Formant X amplitude in relation to the F0 amplitude
FXfrequency	Centre frequency of 1st, 2nd, and 3rd formant
FXbandwidth	Bandwidth of 1st, 2nd, and 3rd formant
lspFreq[0-7]	Line spectral pair frequencies
mfcc_[1-12]	Mel-Frequency cepstral coefficients 1-12
pcm_intensity	Mean of the squared windowed input values
pcm_loudness	Normalized intensity
pcm_zcr	Zero-crossing rate of time signal (frame-based)
slope0-500	Linear regression slope of the logarithmic power spectrum within 0-500Hz
slope500-1500	Linear regression slope of the logarithmic power spectrum within 500-1500Hz
jitterLocal	Deviations in consecutive F0 period lengths
shimmerLocal	Difference of the peak amplitudes of consecutive F0 periods

(a) Short description of utilized Low-Level-Descriptors.

Name	Description
amean	Arithmetic mean
stddev	standard deviation
max	maximum value
maxPos	absolute position of max (in frames)
min	minimum value
minPos	absolute position of min (in frames)
range	max-min
quartile1	first quartile (25% percentile)
quartile2	second quartile (50% percentile)
quartile3	third quartile (75% percentile)
percentile50.0	50% percentile
percentile80.0	50% percentile
iqrY-X	Inter-quartile range: quartileX-quartileY
pctlrange0-2	inter-percentile range: 20%-80%
skewness	skewness (3rd order moment)
kurtosis	Kurtosis (4th order moment)
linregc1	Slope (m) of a linear approximation
linregc2	Offset (t) of a linear approximation
linregerrA	Linear error computed as the difference of the linear approximation
linregerrQ	Quadratic error computed as the difference of the linear approximation
meanFallingSlope	Mean of the slope of falling signal parts
meanRisingSlope	Mean of the slope of rising signal parts

(b) Short description of utilized functionals.

Table 2. Overview of investigated Low-Level-Descriptors (LLDs) and functionals.

5. RESULTS

5.1 Analysis of subjective assessment of speaking style changes

The analyzes of the subjective reports on changes in voice and speaking style are based on the first two questionnaires given after the interaction. According to the formulation of the questions given, the answers vary from more open, extensive ones to more closed ones. The participants used headwords or sentences to describe their experiences and evaluations. These texts made up a total number of 5 725 words. At first, participants' answers on the more open questions regarding the subjective reflections upon their voice and speaking style while interacting with ALEXA in comparison to the interaction with the confederate speaker were analyzed. In general, all 27 participants report, that they recognized differences in their speaking style.

Subjective experiences of the interaction with the confederate speaker The interaction with the confederate speaker is described as "free and reckless" (B [3]) and "intuitive" (X). Participants stated that they "spoke like [they] always do" (G) and "did not worry about" the interaction style (M). The participants explain this behavior by saying that the interaction with humans is simply natural. However, some of them reported particularities when speaking with the confederate speaker, e.g. one participant stated: "I spoke much clearer with Jannik, too. I also addressed him by saying 'Jannik"' (C). This showed that there are participants who adapt themselves to the speaking style during the interacting with ALEXA (see following paragraph). Another participant reported that the information can be reduced when speaking with the confederate speaker: "I only need one or two words to communicate with him and speak about the next step" (H). Altogether, interacting with the confederate speaker is described as "more personal" (E) and "friendly" (E) than interacting with ALEXA.

Subjective experiences of the interaction with the ALEXA Speaking with ALEXA is described as more extensively. Only a few participants experienced it as "intuitive" (AB) and spoke without worrying about their speaking style: "I did not worry about the intonation, because ALEXA understood me very well" (Y). Another one did think about how to speak with ALEXA only when ALEXA did not understand him (B). Besides these few exceptions, all of the other participants report about differences in their voice and speaking style when interacting with ALEXA. The interaction is described as "more difficult" (P), "not that free" (B), "different to interacting with someone in the real world" (M); there is "no real conversation" (I), "no dialog" (J) and "speaking with Jannik was much more lively" (AB).

Subjective experiences of changes in the speaking style characteristics Differences are reported in relation to the prosodic characteristics loudness, intonation, and rhythm (monotony, melody). In the case of *loudness*, participants reported to "strive much more to speak louder" (J) with ALEXA, e.g. because "I wanted that she replied directly on my first interaction" (M). In combination with reflections upon *intonation* one participant said: "I tried to speak particularly clearly and a little bit more louder, too. Like I wanted to explain something to a child or asked it for something." (W). Furthermore, many participants stated that they stressed single words, e.g. "important keywords"

[3] Participants were anonymized by using letters in alphabetic order.

(V), and speak "as clearly and accurately as possible" (G), e.g. "to avoid misunderstandings" (F). However, a few participants explained that they did not worry about intonation (Q, Y) or only worried about it, if ALEXA did not understand them (B, O). Regarding *melody* and *monotony*, participants emphasized to speak in a staccato-like style because of the slowness and aspired clearness of speaking, the repetition of words, and the worrying about how to further formulate the sentences.

Closed evaluation of the different speaking style characteristics The second questionnaire used a more closed answering format. Participants assessed variations of different speaking style characteristics between the interaction with the confederate speaker and ALEXA. Thereby it was explicitly asked for separate assessments of the Calendar and Quiz module. Table 3 shows the response frequencies.

characteristic	R	N	K	I
Intonation	16/17	7/5	4/4	0/1
Monotony	19/19	6/6	2/2	0/0
Melody	10/11	8/7	7/7	2/2

Table 3. Response frequencies for the self-assessment of different speaking style characteristics for the Calendar module (first number) and the Quiz module (second number). Given answers are: **R**eported difference, **N**o difference, I don't **K**now, **I**nvalid answer

It could be seen that all participants indicate to deliberately have changed speaking style characteristics. Only in the Quiz module two participants denied changes in all speaking style characteristics or indicated that they do not know if they changed the characteristic asked for (K, AB). In the Calendar module all participants answered at least one time with "yes" when asking for changes in speaking style characteristics. Furthermore, in the Quiz module more differences were individually recognized by the participants than in the Calendar module.

5.2 Objective Analyses

Regarding the objective analyzes, features distinctive for HHI and HCI were identified using the previous described statistical analyzes. It is assumed that the same features can be identified as distinctive for both modules having a different interaction style. Additionally, it is analyzed whether the absence of the confederate speaker influences the speaking style with the technical system.

Distinctive Features between HCI and HHI in the Calendar Module In the statistical analysis of the features between speakers' HHI and HCI utterances, there were only significant differences for a few feature descriptors in the Calendar module, cf. Table 4. Primarily, characteristics from the group of energy related descriptors (*pcm_intensity*, *pcm_loudness*) were significantly larger when the speakers are talking to ALEXA. Regarding the functionals, this applies to the absolute value (mean) as well as the range-related functionals (*stddev*, *range*, *iqr's*, and *max*). This shows that the participants were in general speaking significantly louder towards ALEXA than to the confederate speaker. The analysis of the data revealed that the participants start uttering their commands very loud but the loudness drops to the end of the command. As further distinctive descriptors only spectral characteristic *lspFreq[1]* and *lspFreq[2]* were identified, having a significantly smaller first quartile.
Regarding the recognition results, it can be seen that even with a very simple recognition system, a performance of 81.97% was achieved in distinguishing HCI and HHI utterances, see Table 4. In comparison to the classification results of more sophisticated classifiers reported in the related work chapter (Section 2) being around 87% this is already a satisfactory result.

Distinctive Features between HCI and HHI in the Quiz Module In contrast to the Calendar module, several features in the quiz module showed a significant difference between HCI and HHI utterances of the participants. This comprises energy related descriptors (*pcm_intensity*, *pcm_loudness*, *alphaRatio*) partly identified in the Calendar module as well as spectral characteristics (*lspFreq[0-6]*, *mfcc[2,4]*, *F0semitone*, *F2amplitude*, *F3amplitude*) and the *pcm_zcr* as a measure for the "percussiveness". The energy-based features behave in the same way as in the Calendar module: the participants generally speak louder. For the group of spectral descriptors the distribution over almost all examined functionals is changed, i.e. here the articulation is strongly different in the addressing of ALEXA (HCI) and the confederate (HHI). The larger number of significantly distinctive features found in the Quiz module is also reflected in the improved recognition performance. The F1-score to distinguish human-directed utterances from system-directed utterances increased up to 88.24%. The complete overview can be found in Table 4.

The impact of the second speaker onto the addressing behavior Finally it is analyzed whether the presence or absence of the confederate speaker is reflected in the participants addressee behavior towards ALEXA. The feature distribution between the HCI utterances of the conditions "alone" for each module (C_A and Q_A) and the condition "with confederate" (C_C and Q_C) are compared. It is assumed that the presence of the confederate speaker does not change the addressing behavior of the participants. Consequently, no acoustic descriptor should show a significant difference between the two conditions "alone" and "with confederate".
This assumption is confirmed by the statistical analysis. No significant differences in the characteristic distributions were found for both modules, even if the α-level is increased to 0.05. Additionally, a classifier was developed to analyze the discriminative power of the acoustic characteristics in recognizing if the participant is interacting with ALEXA alone or with the presence of the confederate speaker. For this case the classifier is just slightly above chance level for both modules, with 59.63% and 66.87% respectively. The complete overview can be found in Table 4.

6. CONCLUSION AND OUTLOOK

The presented study analyzes subjective and objective changes in speaking style characteristics when addressing humans or technical systems. Therefore, the VACC is utilized providing real-life multi-party HCI of one participant interacting with ALEXA alone and with another confederate speaker in two different task settings. Besides

	Calendar		Quiz	
	HHI vs. HCI	HCI_{CA} vs. HCI_{CC}	HHI vs. HCI	HCI_{QA} vs. HCI_{QC}
identified distinctive LLDs	*pcm_intensity*, *pcm_loudness*	–	*lspFreq[0-6]*, *mfcc[2,4]*, *pcm_intensity*, *pcm_loudness*, *pcm_zcr*, *alphaRatio*, *F0semitone*, *F2amplitude*, *F3amplitude*	–
F1-score	0.8197 (0.09224)	0.5963 (0.11991)	0.8824 (0.05439)	0.6687 (0.06154)

Table 4. Overview of identified distinctive LLDs ($p<0.05$) and achieved classification rates (mean and std of F-score) for different compared situations within VACC.

audio recordings of the interaction this dataset additionally provides self-assessments of the participants to reveal insights of their own experiences in the interaction with ALEXA and the confederate speaker.

The analyzes of the participants' reports on recognized changes in speaking style showed that all of them were aware of the differences between their interaction with ALEXA and with the confederate speaker. This answers the **first research question (Do users themselves recognize differences in the interaction with a technical system compared to interacting with another person?)**. Considering the answers of the participants, it is surprising that differences are described in detail already in the open answering format indicating a high level of awareness. All speech and voice characteristics explicitly asked for in the second questionnaire (intonation, monotony, and melody) were brought up by them, and are even extended by reports on differences in loudness of speaking (which was not considered in the second questionnaire). However, it has to be emphasized that in the open answering format none of the participants described differences in all of these characteristics. When asked for differences more precisely during the second questionnaire differences regarding a variety of speech and voice characteristics come to mind and could be described.

The objective analyzes of the speech characteristics revealed that in general the participants are talking much louder in the interaction with ALEXA than in the interaction with the confederate speaker. This answers the **second research question (How do humans speak with current speech-based assistant systems?)**. Furthermore, the type of interaction heavily influences the speaking style, as it can be seen in the comparison of the identified distinctive features for the Calendar and Quiz Module. The changes in speaking style mostly affect the interaction towards the confederate speaker. In a more formal interaction – especially in the simplified setting used in VACC – the interaction for ALEXA and the confederate speaker is quite similar with adjacency pairs of participant's request and ALEXA's/confederate speaker's answer. Whereas in the Quiz module the interaction with the confederate speaker is a more living discussion about solution strategies while the interaction with ALEXA remains a simple request/answer interaction. But nevertheless the acoustic differences are distinctive enough to archive adequate recognition results of over 81%.

In comparison between the subjective and objective analyzes and to answer the **third research question (Do the differences in the observed and/or reported interaction style differ between a formal and an informal interaction setting?)**, it can be stated that humans are aware of their different addressing behavior. To compare the self-assessments regarding different speaking styles and the automatic extractable acoustic characteristics, the description of Ramanarayanan et al. (2017) is used. In the present study only acoustic/prosodic evaluations were analyzed. The assessments regarding pronunciation (diction, sentence length and speaking rate) are focus of subsequent analyzes. Intonation and stress are related to the basic functionals (mean, minimum, maximum, range, standard deviation) of the fundamental frequency and energy related descriptors. Melody and monotony as categories of the speech rhythm are related to changes in functionals describing the mean distance, the mean deviation and, the range and quartile-ranges of fundamental frequency's semitones, formant frequencies, formant bandwidths descriptors. Changes in the range of spectral descriptors describe the tendency of a monotonic voice.

According to this comparison of acoustic descriptors, the subjective self-assessments are supported by the objective statistical feature-distribution comparison in general. Amongst the prosodic evaluations the majority of participants indicated to change intonation and rhythm. But the objective analyzes have revealed that these perceived changes are not reflected equally for every type of interaction. Within the formal Calendar module differences are nearly only identifiable within energy related descriptors (intonation, loudness) and much less within rhythm related descriptors. Whereas within the Quiz module several prosodic characteristics changed between speaking with ALEXA and speaking with the confederate speaker (intonation, loudness, rhythm). Additionally, it has to be noted that neither in the Calendar module nor in the Quiz module distinctive changes of the fundamental frequency could be observed.

Regarding the influence of the confederate speaker onto the addressee behavior, the analyzes reveal that this does not have any influence. Neither the statistical analyzes nor the recognition experiments suggest the presumption that the participants address ALEXA differently when the confederate speaker is present. No distinctive features could be identified and the performance of the developed recognizers are just slightly above chance level for both Calendar and Quiz module. This answers the **fourth research question (Which differences in the speaking style during the interaction with the technical system can be observed when users interact alone or together with a confederate speaker?)**.

Additionally, the comparison of the more positive participant remarks and the more negative participant remarks regarding the interaction with ALEXA and the confederate

speaker with the participant characterization reveals no particularities. It moreover emphasizes the assumption that the addressing behavior is independently of user characteristics. Some participants with previous experiences with voice assistants as well as some participants with less experience describe the interaction with ALEXA as intuitive. The same could be observed for technophilic and non-technophilic participants.

Before discussing future work, some remarks have to be made about limitations of the present study. A main limitation of this work can be seen in the relatively small number of participants, preventing sub-group analyzes, e.g. regarding regular usage of voice assistants. Furthermore, the interaction initiation with ALEXA using a wake word impairs the naturalness of the interaction, which may be an additional factor for the differences in the addressing behavior. Moreover, this study did not consider questions 1 and 2 of the first questionnaire. These questions ask for the overall evaluation of ALEXA and of the confederate speaker within the interaction. Also the objective evaluations and subjective assessments did not include the analysis of the pronunciation (diction, sentence length and speaking rate) and the speech fluency.

Future work will deal with the analysis of diction based subjective and objective evaluation and the identification of a general set of characteristics that distinguish human addressed from system addressed utterances. Hereby the influence of different factors of the technical system (voice, wake word, artificial presence) and of the participants (technical affinity, age, prior experience) will be analyzed. Also in-depth analyzes of reported individual changes in comparison to their objectively measurable characteristics have to be conducted to further get insights on user specific addressee behavior. Thereby, a special focus will be laid on the subjectively reported motives for changing speaking style including the users' individual asciptions towards the technical system. In this regard, Krüger (2018) reports correlations between users' individual ascription of cognitive abilities towards the technical system (e.g. quality of speech recognition) and the subjectively reported speaking style in the interaction with the system. The results of these investigations will light up the users' inner processes revealing specialties in their speaking style in HCI. Thereby, voice assistant systems are enabled to perceive their environment and react properly. This ability is one component in the further development from limited assistance systems towards cognitive assistants. A robust addressee-detection allows voice assistant systems to offer a real conversation mode, which is not only based on the simple continuation of listening after certain dialog steps (asking for the weather, setting up shopping lists, etc.) and reacting to a stop word as it is implemented actually in Google Now Konzelmann (2018). Furthermore, a proper addressee detection for device directed utterances also allows voice assistants to take part and support trustworthy multi-user cooperative tasks with future cognitive systems (Dylla et al., 2013; Buchholz et al., 2017).

REFERENCES

Baba, N., Huang, H.H., and Nakano, Y.I. (2012). Addressee identification for human-human-agent multiparty conversations in different proxemics. In *Proc. of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 6:1–6:6.

Batliner, A., Hacker, C., and Nöth, E. (2009). To talk or not to talk with a computer. *JMUI*, 2(3). doi:10.1007/s12193-009-0016-6.

Bertero, D. and Fung, P. (2016). Deep learning of audio and language features for humor prediction. In *Proc of the 10th LREC*. Portorož, Slovenia.

Beyan, C., Carissimi, N., Capozzi, F., Vascon, S., Bustreo, M., Pierro, A., Becchio, C., and Murino, V. (2016). Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proc. of the 18th ACM ICMI*, 317–324.

Buchholz, V., Kulms, P., and Kopp, S. (2017). Konzept zur Überwachung und Assistenz von Mensch-Maschine Systemen am Beispiel der kooperativen Durchführung eines Praktikumsversuches zur Regelung eines Drei-Tank-Systems. *Kognitive Systeme*, 1, s.p.

Böck, R., Egorow, O., Siegert, I., and Wendemuth, A. (2017). Comparative study on normalisation in emotion recognition from speech. In P. Horain, C. Achard, and M. Mallem (eds.), *Proc of the 9th IHCI 2017*, 189–201. Springer International Publishing, Cham.

Dickey, M.R. (2017). The echo dot was the best-selling product on all of amazon this holiday season. TechCrunch. [Online; posted 26-Dec-2017].

Dowding, J., Clancey, W.J., and Graham, J. (2006). Are you talking to me? dialogue systems supporting mixed teams of humans and robots. In *AIAA Fall Symposium Annually Informed Performance: Integrating Machine Listing and Auditory Presentation in Robotic Systems*. Washington, DC; USA.

Dylla, E., Rehder, T., Helker, S., Fu, X., and Söffker, D. (2013). Konzept zur Überwachung und Assistenz von Mensch-Maschine Systemen am Beispiel der kooperativen Durchführung eines Praktikumsversuches zur Regelung eines Drei-Tank-Systems. *Kognitive Systeme*, 1, s.p.

Egorow, O., Siegert, I., and Wendemuth, A. (2017). Prediction of user satisfaction in naturalistic human-computer interaction. *Kognitive Systeme*, 1.

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., and Truong, K.P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of the ACM MM-2010*.

Frommer, J., Rösner, D., Andrich, R., Friesen, R., Günther, S., Haase, M., and Krüger, J. (2017). Last minute: An empirical experiment in user-companion interaction and its evaluation. In S. Biundo and A. Wendemuth (eds.), *Companion Technology: A Paradigm Shift in Human-Technology Interaction*, 253–275. Springer International Publishing, Cham.

Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwillus and J. Ziegler (eds.), *Mensch & Computer 2003*, volume 57 of *Berichte des German Chapter of the ACM*, 187–196. Vieweg+Teubner, Wiesbaden, Germany.

Hoffmann-Riem, C. (1980). Die Sozialforschung einer interpretativen Soziologie – Der Datengewinn. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 32, 339–372.

Horcher, G. (2018). Woman says her amazon device recorded private conversation, sent it out to random contact. KIRO7. [Online; updated 25-May-2018].

Jovanovic, N., op den Akker, R., and Nijholt, A. (2006). Human perception of intended addressee during computer-assisted meetings. In *Proc. of the 11th EACL*, 169–176.

Katzenmaier, M., Stiefelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proc. of the 6th ACM ICMI*, 144–151.

Kleinberg, S. (2018). 5 ways voice assistance is shaping consumer behavior. think with Google. [Online; posted Jan-2018].

Konzelmann, J. (2018). Chatting up your google assistant just got easier. The Keyword, blog.google. [Online; posted Jun-21-2018].

Krüger, J. (2018). *Subjektives Nutzererleben in der Mensch-Computer-Interaktion: Beziehungsrelevante Zuschreibungen gegenüber Companion-Systemen am Beispiel eines Individualisierungsdialogs*. Verlag Barbara Budrich.

Lange, J. and Frommer, J. (2011). Subjektives Erleben und intentionale Einstellung in Interviews zur Nutzer-Companion-Interaktion. In *Proceedings der 41. GI-Jahrestagung*, volume 192 of *Lecture Notes in Computer Science*, 240–254. Bonner Köllen Verlag, Berlin, Germany.

Lee, H., Stolcke, A., and Shriberg, E. (2013). Using out-of-domain data for lexical addressee detection in human-human-computer dialog. In *Proc. NAACL*, 221–229. Atlanta, USA.

Liptak, A. (2017). Amazon's alexa started ordering people dollhouses after hearing its name on tv. The Verge. [Online; posted 07-Jan-2017].

Lunsford, R. and Oviatt, S. (2006). Human perception of intended addressee during computer-assisted meetings. In *Proc. of the 8th ACM ICMO*, 20–27. Banff, Alberta, Canada.

Mallidi, S.H., Maas, R., Goehner, K., Rastrow, A., Matsoukas, S., and Hoffmeister, B. (2018). Device-directed utterance detection. In *Proc. of the INTERSPEECH'18*, 1225–1228.

Marchi, E., Tonelli, D., Xu, X., Ringeval, F., Deng, J., Squartini, S., and Schuller, B. (2016). Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification. In *Proc of the DCASE2016 Workshop*, 543–547.

Mayring, P. (2014). *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. SSOAR, Klagenfurt.

Osborne, J. (2016). Why 100 million monthly cortana users on windows 10 is a big deal. TechRadar. [Online; posted 20-July-2016].

Oshrat, Y., Bloch, A., Lerner, A., Cohen, A., Avigal, M., and Zeilig, G. (2016). Speech prosody as a biosignal for physical pain detection. In *Proc. of Speech Prosody*, 420–424.

Prylipko, D., Rösner, D., Siegert, I., Günther, S., Friesen, R., Haase, M., Vlasenko, B., and Wendemuth, A. (2014). Analysis of significant dialog events in realistic human-computer interaction. *Journal on Multimodal User Interfaces*, 8, 75–86.

Ramanarayanan, V., Lange, P., Evanini, K., Molloy, H., Tsuprun, E., Qian, Y., and Suendermann-Oeft, D. (2017). Using vision and speech features for automated prediction of performance metrics in multimodal dialogs. *ETS Research Report Series*, 1.

Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., and Otto, M. (2012). LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions. In *Proc. of the 8th LREC*, 96–103. Istanbul, Turkey.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Heck, L. (2012). Learning when to listen: Detecting system-addressed speech in human-human-computer dialog. In *Proc. of the INTERSPEECH'12*, 334–337. Portland, USA.

Shriberg, E., Stolcke, A., and Ravuri, S. (2013). Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. In *Proc. of the INTERSPEECH'13*, 2559–2563. Lyon, France.

Siegert, I., Krüger, J., Egorow, O., Nietzold, J., Heinemann, R., and Lotz, A. (2018a). Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon's ALEXA. In *Proc. of the 11th LREC*. Paris, France.

Siegert, I., Shuran, T., and Lotz, A.F. (2018b). Acoustic addressee-detection – analysing the impact of age, gender and technical knowledge. In *Elektronische Sprachsignalverarbeitung 2017. Tagungsband der 28. Konferenz*, volume 90, 113–120. Ulm, Germany.

Terken, J., Joris, I., and De Valk, L. (2007). Multimodalcues for addressee-hood in triadic communication with a human information retrieval agent. In *Proc. of the 9th ACM ICMI*, 94–101. Nagoya, Aichi, Japan.

Tesch, R. (1990). *Qualitative research analysis types and software tools*. Palmer Press, New York.

Tilley, A. (2017). Neighbor unlocks front door without permission with the help of apple's siri. Forbes. [Online; 17-Sep-2017].

Toyama, S., Saito, D., and Minematsu, N. (2017). Use of global and acoustic features associated with contextual factors to adapt language models for spontaneous speech recognition. In *Proc. of the INTERSPEECH'17*, 543–547.

Tsai, T., Stolcke, A., and Slaney, M. (2015). Multimodal addressee detection in multiparty dialogue systems. In *Proc. of the 40th ICASSP*, 2314–2318. Brisbane, Australia.

Valli, A. (2007). Notes on natural interaction. Technical report, University of Florence, Italy.

van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proc. of the 7th ACM ICMI*, 175–182. Torento, Italy.

Vinyals, O., Bohus, D., and Caruana, R. (2012). Learning speaker, addressee and overlap detection models from multimodal streams. In *Proc. of the 14th ACM ICMI*, 417–424. Santa Monica, USA.

Zhang, R., Lee, H., Polymenakos, L., and Radev, D.R. (2016). Addressee and response selection in multi-party conversations with speaker interaction RNNs. In *Proc. of the EMNLP 2016*, 2133–2143.

How do we speak with ALEXA

Siegert, Ingo; Krüger, Julia

In: Kognitive Systeme / 2018 - 1

DOI: https://doi.org/10.17185/duepublico/48596

URN: urn:nbn:de:hbz:464-20190417-120012-5

Link: https://duepublico.uni-duisburg-essen.de:443/servlets/DocumentServlet?id=48596