Three’s a Crowd? Effects of a Second Human on Vocal Accommodation with a Voice Assistant

Eran Raveh\textsuperscript{1}, Ingo Siegert\textsuperscript{2}, Ingmar Steiner\textsuperscript{3}, Iona Gessinger\textsuperscript{1}, Bernd Möbius\textsuperscript{1}

\textsuperscript{1}Language Science and Technology, Saarland University, Germany
\textsuperscript{2}Mobile Dialog Systems, Institute for Information Technology and Communications, Otto-von-Guericke University, Magdeburg, Germany
\textsuperscript{3}audEERING GmbH, Gilching, Germany
raveh@coli.uni-saarland.de

Abstract

This study examines how the presence of other speakers affects the interaction with a spoken dialogue system. We analyze participants’ speech regarding several phonetic features, viz., fundamental frequency, intensity, and articulation rate, in two conditions: with and without additional speech input from a human confederate as a third interlocutor. The comparison was made via tasks performed by participants using a commercial voice assistant under both conditions in alternation. We compare the distributions of the features across the two conditions to investigate whether speakers behave differently when a confederate is involved. Temporal analysis exposes continuous changes in the feature productions. In particular, we measured overall accommodation between the participants and the system throughout the interactions. Results show significant differences in a majority of cases for two of the three features, which are more pronounced in cases where the user first interacted with the device alone. We also analyze factors such as the task performed, participant gender, and task order, providing additional insight into the participants’ behavior.

Index Terms: vocal accommodation, human-computer interaction, confederate influence

1. Introduction

In recent years, the market for commercial voice assistants (sVAs) has rapidly grown. For example, Microsoft Cortana had 133 million active users in 2016 [1] and Echo Dot was Amazon’s best-selling product between 2016 and 2018 [2]. Furthermore, 72% of people who own a smart speaker say they often use their devices as part of their daily routine [3].

The big advantage of VAs is their simple operation. Using nothing but speech commands, users can perform tasks like playing music, searching the web, shopping online, etc. In the future, we are likely to witness an ever-growing presence of devices with spoken interaction capabilities, like speech-activated cars, hands-free medical assistants, and intelligent tutoring systems. This will increase the demand on voice-activated devices even more, as they will need to support more functionalities in a way that is comfortable and intuitive for the users. Additionally, it can be expected that such devices will be used not only by individuals, but also in more social contexts, i.e., where multiple humans are involved. Therefore, we find it important to investigate not only human-computer interactions, but also human-human-computer interactions.

Besides making the operation of such voice-activated systems simple and user-friendly, VAs also aim to let users interact with them in a familiar, natural manner. One property of natural interactions is the tendency to accommodate to the specific situation and interlocutors to make the interactions more fluent and efficient [4, 5]. Linguistic accommodation is one aspect of this phenomenon, and it is found in various human-human interaction (HII) experiments [e.g., 6, 7]. In various HCI experiments, it has been shown that participants speak differently to computers in general, and also change their speech behavior during the interaction, e.g., by Branigan et al. [8] and Levitan et al. [9]. However, these interactions include only the computer-based interlocutor and emphasize the comparison between different configurations of the system itself. Moreover, they only examine the influence of the system’s speech output on the user, but not the influence of other interlocutors.

The question tackled in this paper is whether and to what extent speaking to a second human interlocutor in addition to Alexa influences the accommodation in interaction with a VA. More generally, we investigate whether users speak differently towards a computer-based system when another human participates in the interactions. This was done using a set of interactions of participants with a VA alone or with a VA and a confederate. Thus, it allows us to analyze the participants’ accommodation to the VA in both social conditions.

2. Dataset

To analyze the influence of a confederate speaker in HCI, we used the Voice Assistant Conversation Corpus (VACC) [10], which is freely-available for research. This corpus comprises balanced human-computer (solo condition) and human-human-computer (confederate condition) interactions with a 2nd generation Amazon Echo Dot that uses the skills and female voice of the virtual assistant Alexa. The first human speaker is the participant, which always takes part in the interactions. In the confederate condition, the confederate interacted only with the human speaker, never with Alexa. The interactions consist of Calendar and Quiz tasks, where the former simulates a formal situation and the latter a rather informal situation. This corpus, which alternately introduces a confederate speaker into otherwise similar conversations, allows investigating the influence of the confederate on the behavior of the participant.

VACC contains recordings of 27 (14 female) German native speakers with a mean age of 24 years (sd 3.3). Each participant performed the Quiz and Calendar tasks in both solo and confederate conditions, for a total of 108 interactions. An interaction was finished either by completing the task or by stopping it prematurely in case no further progress could be made, to avoid participant frustration. The latter, however, happened only a few times. Approximately 13500 utterances were recorded, stretch-
In the focus of convergence research [12, 13, 14]:...
The non-human speaker, the temporal order of the values is not considered (cf. Section 4.2).

To detect these differences, the distribution of their respective values in the solo and confederate conditions in each interaction pair were compared. This was done by using the two-sample Wilcoxon test [17], with $\alpha = 0.05$ with the null hypothesis that similar value distributions of the target feature were used in both conditions. A significant result of the test means that the participant produced the respective feature differently when interacting with Alexa alone compared to when the confederate participated as well. Table 1 shows the percentage of interaction pairs, in which the null hypothesis was rejected, i.e., that the feature was utilized differently by the participant in each condition. Since chronologically, one of the conditions needed to precede the other, the percentages were also calculated separately for the cases where tasks were performed first in the solo condition (and then in the confederate condition), and vice versa. This separation shows whether interacting first with Alexa alone, without any human input, influenced the vocal behavior of the participants. As there were no breaks between the tasks, the only factors for change were the order of the conditions and the involvement of another human speaker. Indeed, the percentages of significant differences when interacting first only with Alexa were higher by 12%, 20%, and 3% for $f_0$, intensity, and AR, respectively. Noticeably, the participants’ AR was sometimes temporarily lower when repetitions were required due to false recognition by Alexa.

Figure 1 further breaks down the differences between interaction pairs and introduces the factor of the performed task. In line with the tendency shown in Table 1, the features $f_0$ and intensity have the highest percentages of significant cases regardless of the performed task, and the tasks performed first show higher percentages of different distributions. In the lower percentages, it is the task, rather than the target feature, that shows differences between the cases. And last, for AR, with the lowest percentages, there is a clear difference between the Quiz and the Calendar tasks. All in all, the task factor was a good indicator only for the feature with the lowest difference percentage and the order factor was more informative for the features with higher percentages.

4.2. Temporal analysis

Another way to look at accommodation in an interaction is in the temporal dimension. In this analysis, the same raw measured values were used to examine changes that occur over time. That is, unlike the analysis presented in Section 4.1, here the order of the values plays a major role, and effects may be found in specific time windows.

To perform such an analysis over the entire interaction, two additional computation steps are required. First, each point in time must have a corresponding value for each feature produced by all speakers. This was achieved by smoothing the measured value using LOESS [18], a non-parametric regression method that deterministically fits a function to a localized subset of the data. The fitting was done for each speaker separately over all slices of human-directed speech/device-directed speech with measured values of the features. This results in a predicted value for each slice of the conversation. Figure 2 shows an example of these smoothed measures of one participant and Alexa for the $f_0$ feature. The lower part of the figure shows the accommodation changes of the participant during the interaction (blue for convergence and red for divergence), and the upper part shows the turn-taking events. The confederate condition has fewer turn events, as the analysis concentrates on the participant and Alexa, and the confederate turns are not shown. Secondly, the relationship between a feature’s values in each slice needs to be determined to describe their temporal changes. Since we are interested in accommodative behavior, a measure for the relative change between slices was used. It calculates the participant’s contribution in the overall change of distances.
between the participant and Alexa. Alexa’s contribution is considered as a static effect, as it is not able to change based on the user’s speech input, and is therefore not taken into account. The change tendency between two slices is calculated by

$$\text{change}_t = -\Delta_{t, t-1} | S_{part} - S_{Alexa} |,$$

where the index $t$ refers to the current slice and $S_{part}$ and $S_{Alexa}$ are the smoothed values of the participant and Alexa, respectively. The minus sign at the front flips the result so that increased proximity (convergence) is represented by positive values and distancing (divergence) by negative values (see Figure 2). Subsequently, the participant’s contribution toward accommodation is calculated by

$$\text{accomm}[^{\{\text{participant}\}]_t = \text{change}_t - \Delta_{t, t-1} S_{Alexa}.$$  

The sum of the proximity changes of each target produced by all participants in every sex-task-condition-order combination was calculated, resulting in a single value that represents the overall change. A value greater than zero means that more convergence was observed, and a negative value points to more divergence. There were only two instances where this value was exactly zero, both for the AR feature. These instances were treated as cases of divergence. Using this approach, only a few interactions had no feature convergence in them, and several had all three features showing convergence. However, we took a stricter approach, where a feature was considered as converging only if its overall accommodation value was higher than one standard deviation from its mean. Based on that, all interactions were categorized by the number of features that showed more convergence in them.

Figure 3 summarizes the categorization with respect to each factor individually. Each line represents a single interaction. The strata labels through which each line passes, indicate the group it belongs to with respect to each factor. The number of features that showed more convergence than divergence overall are marked by the color of the line. Some tendencies emerge from this categorization: first, in 35% of the interactions, there was at least one feature that showed convergence, but in none of them did all three features do so. In seven interactions, two features showed convergence, twice by males and five times by females. In total, males converged in 5% of all measured features and females in 7%. Furthermore, of all the instances of converged features, 58% occurred in the solo condition, compared to 42% in the confederate condition. However, no difference between the Calendar and Quiz tasks was found, with 49% and 51% of the cases, respectively. The same holds for the comparison between the two orders in which the tasks could be performed.

These results support the addition of the confederate to the interaction as the factor for less convergence.

5. Discussion and conclusion

We have presented distributional and temporal analyses of differences in convergence of three vocal features in HCl, with emphasis on changes resulting from simultaneous, spoken HHI. The distributional analysis examined the difference in behavior of the features between interactions only with a computer-based device and interactions with both the device and a confederate, while the temporal analysis investigated the participants’ account in the overall changes in proximity between the interlocutors. To also consider the point in time during a session in which the confederate was speaking, the order factor was considered as well. For all three features, it was found that (a) the distributions differed more when the participant first interacted with the device alone, and (b) more convergence was aggregated in the task that was performed first. From these two findings, it can be concluded that chronological order of interactions affected the speech behavior of the participants. Further analyses took the factors sex and task into account and indicated that female participants showed less convergence than male participants, but the task performed did not play any role in increasing the amount of convergence.

The first speech input a participant encounters may cause a priming effect that, together with the natural tendency to converge to an interlocutor, results in more change in interactions that occur first. However, the interchangeability of input here, both HHI and HCI) seems to hinder the ability of the participants to converge to Alexa. One explanation for this may be that it is more natural for humans to accommodate to other humans, so once another human is involved, the accommodation towards the computer interlocutor is neglected. Another possible explanation is that due to the multiple interlocutors, the participants do not have a steady target towards which to accommodate, which leads to a weakened convergence effect. This is confirmed by the higher rate of convergence in the solo condition compared to the confederate condition. Since HCI still lacks the mutuality of a comprehensive accommodation effect, the question arises whether these tendencies would be stronger in interactions with a single human versus interactions with two different human speakers simultaneously. The difference in convergence instances between female and male speakers may be ascribed to the VA using a female voice and could be further investigated by using a VA with a male voice.
6. References


