COMPARING PHONETIC CHANGES IN COMPUTER-DIRECTED AND HUMAN-DIRECTED SPEECH

Eran Raveh[1,2]*, Ingmar Steiner*[2,3]*, Ingo Siegert*[4]*, Iona Gessinger*[1,2]*, Bernd Möbius*[1]

[1]*Language Science and Technology, Saarland University*
[2]*Multimodal Computing and Interaction, Saarland University*
[3]*audEERING GmbH, Gilching*
[4]*Mobile Dialog Systems, Institute for Information and Communications Engineering,*
Otto-von-Guericke University, Magdeburg
raveh@coli.uni-saarland.de

Abstract: This paper presents a study that examines the difference of certain phonetic features between human-directed speech (HDS) and device-directed speech (DDS) in human-human-computer interactions. The corpus used for the analyses consists of tasks performed by participants in cooperation with a human confederate and/or a computer-based interlocutor. This includes distributional and temporal analyses, examining the similarities and differences of the overall distribution of the measured features and time-based changes throughout the interactions. The features fundamental frequency, intensity, and articulation rate were selected for analysis. Results show significant differences in a majority of the cases for two of the three selected features as well as insights regarding the participants' speech behavior during the interaction. These outcomes provide a look into further aspects of HDS and DDS and speech-related features in conversation analysis, which may help studies in topics like addressee detection or human-computer interaction (HCI).

1 Introduction

Nowadays, we are witnessing in our everyday lives an ever-growing presence of devices with spoken interaction capabilities, like personal assistants, speech-activated cars, hands-free medical assistants, and intelligent tutoring systems, to name a few. The question arises, therefore, whether different speech patterns and characteristics emerge in such human-computer interaction (HCI) than in human-human interaction; and if yes, which.

It has been demonstrated that humans may change their speech behavior when interacting with computer-based systems. One way of measuring such changes is in terms of linguistic similarity between the interlocutors. In various HCI experiments, participants have been shown to speak differently to computers in general, and also change their speech during the interaction (see Branigan et al. [1], for examples). However, these are human-computer interactions that emphasize the comparison between different configurations of the system itself [e.g., 2]. Moreover, no direct comparison between human-directed and computer-directed speech was performed.

In this paper, we present conversation-level analyses of speech changes in human-human-computer interactions with an Amazon Alexa device (Echo Dot, 2nd generation) as the computer-based interlocutor. We chose to analyze three phonetic features: fundamental frequency (f_0), intensity, and articulation rate (AR). These analyses show the differences in the human interlocutor's behavior when addressing the confederate human speaker or the computer interlocutor.

Examples of distributional and temporal changes across the interaction are given. These analyses are performed on the Voice Assistant Conversation Corpus (VACC) [3], which comprises two experimental scenarios: the Calendar Module (formal interaction) and the Quiz Module (informal interaction).

Other studies, like Shriberg et al. [4] and van Turnhout et al. [5], used similar corpora to study automatic addressee detection. The present work does not set detection and classification as its goal, but rather aims to provide insights and measures that might be useful for such tasks. The results of the analyses show that the differences between the distributions for the features f_0 and intensity in human-directed speech (HDS) and device-directed speech (DDS) contexts were significant in 74 % and 89 % of the cases, respectively, while for AR in only 13 %. This sheds light on the similarities and differences in speech when addressing humans and computers.

2 Dataset

The VACC [3] was utilized to examine differences in HDS and DDS. This corpus consists of conversations between one or two human interlocutors with the 2nd generation of the commercial smart speaker Amazon Echo Dot, which uses the skills and voice of the virtual assistant Alexa. The conversations comprise a formal and an informal scenario conducted either with the participant alone or together with a confederate accompanying person, which allows investigating how humans address computer-based systems.

2.1 Setting and participants

VACC contains recordings of 27 German native speaking students from Otto von Guericke University Magdeburg. Each speaker participated in four recordings, for a total of 108 interactions (27 participants × 2 scenarios × 2 conditions). The total recording time is 17 h 7 min (31 min on average per interaction) containing $\sim$13 500 utterances. The number of female (14) and male (13) participants is nearly equal, and their ages range from 20 to 32 years (mean 24.11; sd 3.32). The participants came from different study programs, including computer science, engineering, humanities, and medical sciences. Thus, this dataset is not biased towards students with stronger technical background.

An experiment with a participant consists of four interactions: A formal and an informal scenario, each carried out in solo and confederate conditions. An interaction was finished either by reaching its aim or by stopping it to avoid participant frustration in case no further progress could be made.

In the first scenario, the Calendar Module, each participant was asked to make appointments with a project partner. The participant's calendar was stored online and was only accessible via Alexa's voice commands. In the solo condition, the participants only received written information about the confederate's available dates and had to interact with Alexa on their own. In the confederate condition, the confederate provided the relevant information. Therefore, the participant had to interact with both Alexa and the confederate to find available time slots for the appointments.

In the second scenario, the Quiz Module, the participant answered trivia questions like "When was Albert Einstein born?", which are mostly assumed to be too difficult to answer correctly without Alexa's help. In the solo condition, the participants had to find the correct strategy for formulating questions for Alexa on their own own, whereas in the confederate condition the participant and the confederate teamed up to decide on a strategy.

2.2 Annotations

Each utterance in an interaction was annotated with its speaker, context, and textual transcription. The speaker of each utterance could be the participant, Alexa, or the confederate. The context marked the type of interaction of the utterance, which include HDS, DDS, cross-talk, off-talk, laughter, and more. To deal with clearer data, only HDS and DDS contexts were used for analysis in this paper (see Section 3). The transcription was obtained using the Google Cloud Speech API automatic speech recognition service.

3 Method

To make the comparison between HDS and DDS more direct, we selected only those interactions where the participants talked with both the device and the confederate. The remaining 50 % of the dataset includes one calendar task and one quiz task for each of the 27 participants. This subset was analyzed based on the audio signals of the interactions only. The turn annotations were used to determine to which of the three speakers the measured values should be ascribed. The text transcription was not utilized for this study.

Each interaction was analyzed using the audio from the participant's microphone (where both the confederate and the device are audible as well) and the annotations described in Section 2.2. To increase temporal resolution, the audio signals were cut into two-seconds slices. A single slice always contains audio from a turn of a single speaker, any remainder shorter than 2 seconds gets a slice of its own. For example, a turn of 5.2 s in length was sliced into three slices of 2 s, 2 s, and 1.2 s. Each of the features was then analyzed within a single slice.

The following phonetic features were analyzed:

Fundamental frequency ($\mathbf{f}_0$) – the mean pitch measured within the audio slice with automatic time step selection and a range between 60 Hz and 350 Hz.

Intensity – the mean intensity measured within the audio slice with automatic time step selection.

Articulation rate (AR) – the ratio of number of syllables to phonation time within the audio slice, as described in De Jong and Wempe [6].

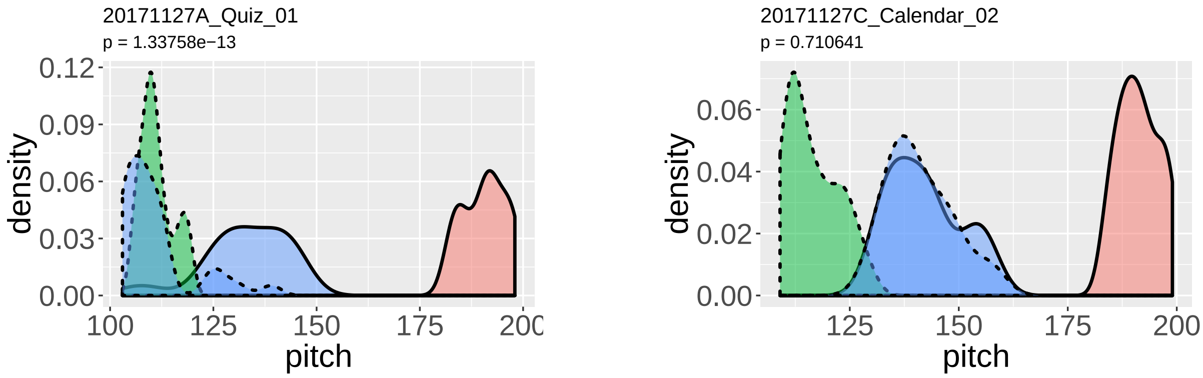
All features were measured using Praat[1] [7] scripts that processed the signal of the participant's microphone. To filter out noise and concentrate on the more characteristic speech style, only values between the first and third quartiles were taken into account for the non time-based analyses. Furthermore, turns not annotated as HDS or DDS (e.g., cross-talk or off-talk) were also ignored.

4 Results

As a first measure of speech characteristics, the means, medians, and standard deviations of the selected features in the participants' speech in each of the interactions were calculated in HDS and DDS. The absolute values reflected in the means and medians can shed light on the overall range of values used with each of the two interlocutors, for example, due to different gender (Alexa was always set with a female voice and the confederate was always male) or assumed comprehension capabilities of humans and computers. The standard deviations show the variability of a feature with each interlocutor, which may indicate a different production style.

Each target feature was measured and listed chronologically throughout the interaction. These lists were divided into four distributions based on speaker and context: participant talking

[1] version 6.0.35



(a) An example of HDS and DDS distribution densities with a *significant* difference (p-value$\ll$0.0001, $\alpha = 0.05$) extracted from the f_0 measures of participant 20171127A in the Quiz task.

(b) An example of HDS and DDS distribution densities with a *non-significant* difference (p-value=0.71, $\alpha = 0.05$) extracted from the f_0 measures of participant 20171127C in the Calendar task.

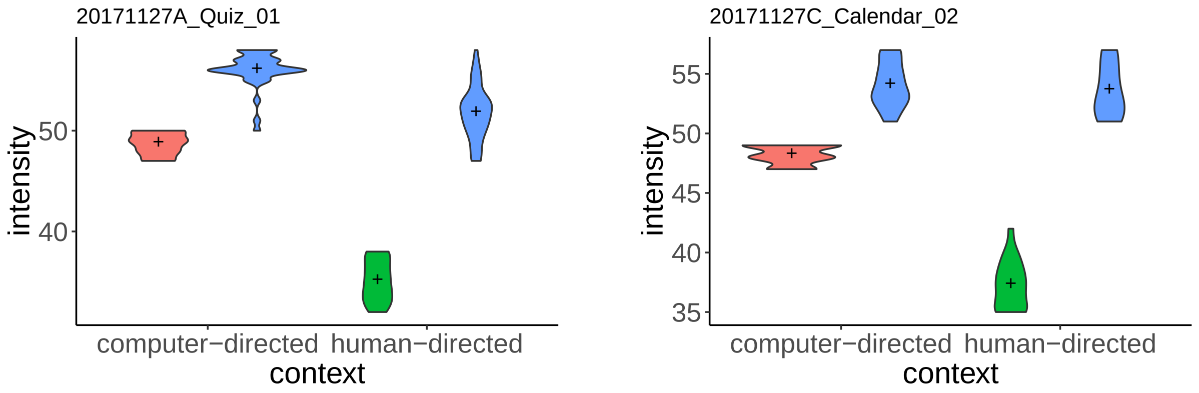
Figure 1 – A *significant* (a) and a *non-significant* (b) difference between distribution densities of participants' HDS and DDS f_0. The colors represent distributions of Alexa (red), the confederate (green), and the participant (blue). The line style differentiates between HDS context (dashed line) and DDS context (solid line).

to the confederate, participant talking to Alexa, confederate talking to participant, and Alexa talking to participant. The contrast between HDS and DDS is observable within the participant's speech only, which was active in both contexts. The significance level of the difference between these two distributions was measured using a two sample t-test with $\alpha = 0.05$.

Figure 1 shows examples of the distributions of the participant's f_0 in HDS and DDS contexts. Since the device always used the default Alexa female voice and the confederate was always male, there is a natural gap between their f_0. This gap leaves room for convergence to occur, i.e., when two interlocutors become more similar to each other over time with respect to some measured feature. In 74 % of the cases out of the 54 analyzed interactions the difference of means of the participant's HDS and DDS f_0 distributions was significant. In 85 % of the cases where the difference was significant, more of the probability mass of DDS's distribution contained higher values than HDS's, and therefore more similar to Alexa.

Figure 2 shows examples of the distributions of the participant's intensity in HDS and DDS contexts. Unlike in the case of f_0, absolute measured values may not be as meaningful due to the device's and the confederate's location relative to the participant's microphone. As explained in Section 3, the signal from the participant's microphone was used for the difference analyses. This means that the absolute values of the participant's intensity in HDS and DDS can be compared directly, but only indirectly with Alexa's and the confederate's. Therefore, in Figure 2 the differences in distribution and frequency can be compared within a context, but the values should only be compared within the participant's speech (in blue). In 89 % of the cases out of the 54 analyzed interactions the difference of means of the participant's HDS and DDS intensity distributions was significant. Moreover, participants tended to speak to Alexa with a louder voice than to the confederate.

The differences of articulation rate (AR) distributions in HDS and DDS were calculated as well. In 13 % of the cases out of the 54 analyzed interactions the difference of means of the participant's HDS and DDS AR distributions was significant. This shows that the participants largely spoke at the same speed with the confederate and the device. It was found that the articulation rate was lower in some specific cases where the participant tried to improve her/his intelligibility to the system, specifically when the system's output indicated that it could not understand the participant's utterance. While such utterance-level changes are interesting and may point to a temporary change in behavior, a more detailed analysis is outside the scope of



(a) An example of HDS and DDS values with a *significant* difference (p-value≪0.0001, $\alpha = 0.05$) extracted from the intensity measures of participant 20171127A in the Quiz task.

(b) An example of HDS and DDS values with a *significant* difference (p-value=0.55, $\alpha = 0.05$) extracted from the intensity measures of participant 20171127C in the Calendar task.

Figure 2 – A significant (a) and non-significant (b) difference between extracted values of participants' HDS and DDS intensity. The colors represent distributions of Alexa (red), the confederate (green), and the participant (blue). HDS is plotted in the right, and DDS in the left of the plot. The width of the box represents the frequency of the values and the '+' sign marks their respective means.

this study, which concentrates on interaction-level behavior.

Looking at the distribution differences of the selected features in HDS and DDS sheds light on the general speech behavior in these contexts. However, this analysis leaves out an important aspect of spoken interaction, namely the time dimension. While the static measures of distributions show the overall range and frequency of the values, temporal analysis adds the information as to how they changed over time. Adding the time dimension gives an overview of the interaction's structure and reveals fine-grained information regarding its characteristics, such as turn lengths, turn switching, pauses, density of a speaker's utterances, convergence or divergence effects, etc. For example, Figure 3(a) shows a case where the the absolute f_0 values are roughly the same in HDS and DDS, namely around 150 Hz, but the behavior of the participant is different. In the DDS context, the participant generally keeps a stable distance from Alexa's voice, whereas in the HDS context the f_0 values are closer to the confederate. In both contexts, the participant's f_0 starts around 150 Hz. However, in HDS the minimum f_0 is only slightly below this initial value, whereas in DDS it drops as far as 25 Hz lower. An example for the intensity feature is shown in Section 4. Unlike the previous example, here the absolute values steadily differ by about 5 dB, but the overall change is similar. That is, in both cases the intensity rises from the beginning to around a quarter of the interaction's duration, and then decreases again until the end (in HDS more quickly than in DDS) down to approximately the same value as at the beginning.

Since Figure 3 shows two examples of the Quiz task performed by two different participants, it is possible to compare the structure of these interactions as well. As described in Section 2, the Quiz task in the confederate condition is designed so that the two human speakers needs to find an effective way to solve the questions using Alexa. After improving their strategy, the lead should ultimately be taken by the participant, who interacts with Alexa to solve the questions as quickly and correctly as possible. In both examples, the interaction starts with relatively short turns and rapid context changes. This might be ascribed to the fact that the participant is still trying to figure out the best way to interact with Alexa and the confederate. Then, sometime after the middle of the interaction, there is a larger block of DDS only, followed by some more turns of HDS. Finally, the interactions end with another, shorter, block of DDS, in which the participants finish the last questions of the quiz. This structure was found to be

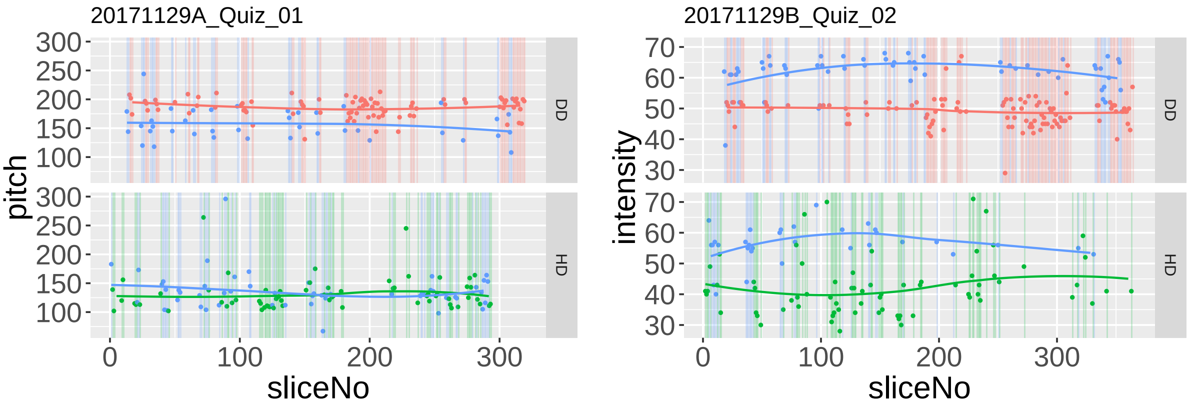


Figure 3 – The changes in pitch (left) and intensity (right) over time in DDS (upper part) and HDS (lower part). The time spans on the x-axis are represented by turn slices, as explained in Section 3, and the y-axis shows the value of the feature. A slice's background color indicates the speaker in this slice and the circle with the same color, the measured value of the feature in it. Alexa's voice is shown in red, the confederate in green, and the participant in blue. The lines are smoothed values calculated by LOESS [8].

typical for the Quiz task.

Table 1 summarizes the results.

5 Discussion

Section 4 presented the results for the features f_0, intensity, and articulation rate (AR). The first two show a greater degree of difference between HDS and DDS, and the latter, a smaller one. One possible explanation for the different f_0 distributions is the natural difference in male and female f_0 (Alexa used a female voice while the confederate was always male) and the fact that humans sometimes tend to match their f_0 to the interlocutor. In that case, the results shown here point to the fact that the participants generally treated Alexa as a human interlocutor with regards to f_0 behavior, as opposed to, for example, matching only the human interlocutor and talking to the computer-based interlocutor using the same f_0. A similar effect was found for intensity. Since the device and the confederate were spaced approximately at the same distance from the participants, there was no apparent reason for the participants to speak more loudly with either interlocutor. Therefore, an explanation of the tendency to speak more loudly to the device may come from the intuition that a computer-based system has a harder time to understand human speech and therefore needs a clearer signal. Another explanation may be the illusion that Alexa feels more distant than the human interlocutor, because Alexa is not an embodied agent. Keeping in mind that an interaction aims to be as efficient as possible using minimum amount of energy, it seems like changing these features helped – or at least felt like helping – the participant to interact more efficiently with the device. This is not the case with AR, which shows a lower degree of differences. Slower, more carefully articulated speech, occurs less often in regular speech than louder speech or higher pitch. Such enhanced articulation not only takes longer to produce, but also requires more effort, making it a less preferred way to communicate, unless it is necessary. In a somewhat formal, experimental setting, participants are likely to speak more slowly than usual, and the motivation to complete the task in a short time does not encourage them to speak even more slowly. This supports the hypothesis that extra slow speech would only be used when necessary, e.g., when a repetition is required due to a misunderstanding of an utterance. Even in that case, the overall AR tends to increase afterwards, to make the interaction more fluent again.

Future work may go in two main directions, both concerning a temporal aspect of interac-

Table 1 – Summary of results. The percentage of interactions in which the difference of distribution means was significant for each feature, and their mean and standard deviation (sd).

	f_0	intensity	AR
signif. diff.	74 %	89 %	13 %
HDS mean (sd)	10.5 Hz	2.95 dB	0.627
DDS mean (sd)	10 Hz	2.61 dB	0.634

tion. The first has do to with analysis on the speech signal level, where the changes in measures over time can capture phenomena like convergence or divergence. In a more comprehensive analysis in this direction, more detailed patterns may emerge. Such an analysis can concentrate on one context or on comparing patterns in both HDS and DDS. Additionally, more features can be measured to reveal more details regarding speech behavior. The second potential direction may highlight behavioral patterns of the conversation and turn levels. This can include a closer examination of the interaction structure as a whole, the dynamics of turn changes, pauses and repetitions, etc. Such an analysis can be performed on interactions in solo and confederate condition to inspect whether humans deal with the same task differently with a computer alone, than when another human is involved.

6 Conclusion

In this paper, we presented an analysis of phonetic features in a study based on a subset of a human-human-computer corpus, which includes two tasks with 108 interactions in total. Three interlocutors participate in the interactions: the participant, a confederate, and an Amazon Alexa device. The features f_0, intensity, and AR were analyzed for each of the three interlocutors across each interaction. Based on the dataset's turn annotations, the utterances of the speakers were categorized as either HDS or DDS context. First, the participant's speech in both contexts was examined by comparing the distributions of the measured values in each context. Then, the difference of the distributions was checked for significance. Finally, the temporal changes of the features across the interaction were examined as well.

The difference between the distributions was significant in 74 % and 89 % of the interactions for f_0 and intensity, respectively, and in 13 % for AR. As for the temporal analysis, different patterns of changes were observed, like cases where the participant accommodated to the human interlocutor but not the computer and cases where a similar behavior was observed in both contexts.

References

[1] Branigan, H. P., M. J. Pickering, J. Pearson, and J. F. McLean: *Linguistic alignment between people and computers*. *Journal of Pragmatics*, 42(9), pp. 2355–2368, 2010. doi:10.1016/j.pragma.2009.12.012.

[2] Levitan, R., S. Benus, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg: *Implementing acoustic-prosodic entrainment in a conversational avatar.* In *Interspeech*, pp. 1166–1170. San Francisco, CA, USA, 2016. doi:10.21437/Interspeech.2016-985.

[3] Siegert, I., J. Krüger, O. Egorow, J. Nietzold, R. Heinemann, and A. Lotz: *Voice assistant conversation corpus (VACC): A multi-scenario dataset for addressee detection in human-computer-interaction using amazon's ALEXA*. In *Workshop on Lan-*

guage and Body in Real Life & Multimodal Corpora. Miyazaki, Japan, 2018. URL http://lrec-conf.org/workshops/lrec2018/W20/pdf/13_W20.pdf.

[4] SHRIBERG, E., A. STOLCKE, and S. RAVURI: *Addressee detection for dialog systems using temporal and spectral dimensions of speaking style*. In *Interspeech*, pp. 2559–2563. Lyon, France, 2013. URL https://www.isca-speech.org/archive/interspeech_2013/i13_2559.html.

[5] VAN TURNHOUT, K., J. TERKEN, I. BAKX, and B. EGGEN: *Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features*. In *7th ACM International Conference on Multimodal Interfaces (ICMI)*, pp. 175–182. Trento, Italy, 2005. doi:10.1145/1088463.1088495.

[6] DE JONG, N. H. and T. WEMPE: *Praat script to detect syllable nuclei and measure speech rate automatically*. *Behavior Research Methods*, 41(2), pp. 385–390, 2009. doi:10.3758/BRM.41.2.385.

[7] BOERSMA, P.: *Praat, a system for doing phonetics by computer*. *Glot International*, 5(9/10), pp. 341–345, 2001.

[8] CLEVELAND, W. S. and S. J. DEVLIN: *Locally weighted regression: an approach to regression analysis by local fitting*. *Journal of the American Statistical Association*, 83(403), pp. 596–610, 1988. doi:10.1080/01621459.1988.10478639.