*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

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

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

[1]Language Science and Technology, Saarland University, Germany
[2]Mobile Dialog Systems, Institute for Information Technology and Communications, Otto-von-Guericke University, Magdeburg, Germany
[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 demands 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 (HHI) 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 13 500 utterances were recorded, stretch-



http://dx.doi.org/10.21437/Interspeech.2019-1825

Table 1: *Percentage of interaction pairs with significant differences with respect to each target feature with all the interactions together and separated by order tasks.*

| feature | any order | solo first | confederate first |
|---|---|---|---|
| $f_0$ | 67 | 72 | 60 |
| intensity | 67 | 76 | 56 |
| AR | 30 | 31 | 28 |

ing over total recording time of 17 h 7 min (31 min on average per interaction). The permutations of the tasks, conditions, and their order were balanced.

In the Calendar task, the participant made several appointments in pre-defined weeks with the confederate. The participant's calendar was stored online, accessible only via Alexa. In the solo condition, the participants got written information about the confederate's availability, whereas in the confederate condition, the confederate could be asked directly about it, resulting in a *human-human-interaction*. In the Quiz task, the participant was asked to answer trivia questions, like "When was Albert Einstein born?", using Alexa. Although Alexa was not always able to immediately provide a full answer to all the questions, information could be incrementally gathered in multiple steps. Here, the participant solved the quiz alone in the solo condition, or teamed up with the confederate so the two could discuss the question asking strategy. The Quiz task was generally less formal than the Calendar task.

The three interlocutors were arranged at an approximately equal distance from each other. The two human speakers were seated and the Echo device was placed on a table. More information about the recording setup and equipment is described by Siegert *et al.* [10]. Turn times and speakers were manually annotated.

## 3. Method

All interactions from the VACC were used for the analyses. The comparisons were performed on pairs of interactions of the same task in the two conditions presented above. Only the audio signals of the interactions were used for analysis, as recorded by the headset microphone of the participant, which captured the speech of the confederate and Alexa as well. Since the participants sat at an equal distance from Alexa and the confederate, this eliminates any spatial influence on their speech, e.g., in terms of intensity. The turn annotations were used to determine to which of the three speakers the measured values belong.

To increase temporal resolution, the audio signals were split into two-second slices. This duration is short enough to provide a decent temporal resolution shorter than a turn, but still long enough to calculate time-dependent features like articulation rate (AR). Splitting the turn also creates equal, consecutive, and more comparable time units for an interaction without introducing artificial boundaries by dividing it into a pre-defined number of parts [11]. This is especially important for the temporal analysis (Section 4.2). Preliminary experiments with the corpus showed only very small changes in feature measurements with slices of longer duration. The slicing was done per turn, so that a single slice contains the audio of a single speaker. Any remainder of a turn duration got a slice of its own.

The following phonetic features were targeted, as they are in the focus of convergence research [12, 13, 14]:

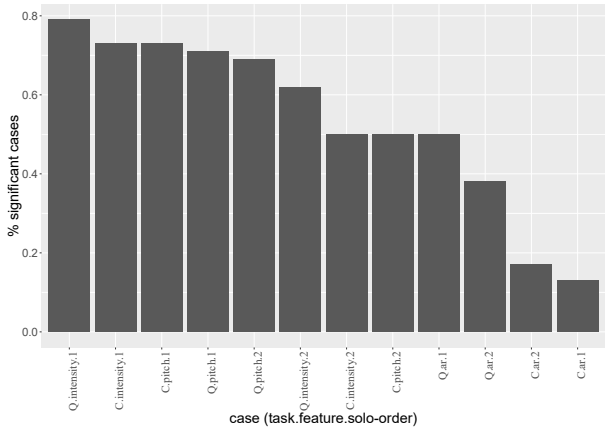


Figure 1: *Percentage of instances with a significant difference between solo and confederate conditions in each case. A case is a combination of the factors* task*,* feature*, and* order*. For example, the case* Q.intensity.2 *contains the comparisons of intensity in interactions of the Quiz task where solo condition was performed second (and the confederate condition first).*

**Fundamental frequency ($f_0$)** mean pitch measured within the audio slice with automatic time step selection and a range between 60 Hz and 350 Hz.

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

**Articulation rate (AR)** ratio of number of syllables to phonation time within the audio slice, as described by De Jong and Wempe [15].

All features were measured in each slice individually using Praat[1] [16] scripts. To filter out noise and concentrate on the more characteristic speech style, only values from the second and third quartiles were taken into account. Furthermore, turns not annotated as speech for either of the speakers (e.g., cross-talk or off-talk) were also ignored.

## 4. Results

Two separate analyses were carried out: distributional and temporal. The first looks at global differences on the interaction level of the participant's and the computer-based interlocutor's productions between solo and confederate conditions. It also checks whether the order in which the tasks were performed had any influence on the changes as well (see Table 1 and Figure 1). The second examines time-based, continuous changes in the proximity between the participant's and the device's productions with emphasis on the *condition* factor, and then also provides additional insights for the factors *sex*, *task*, and *order* (Figures 2 and 3).

### 4.1. Distributional analysis

The distributional analysis examines the differences between the participant and Alexa's speech in solo and confederate conditions in terms of the general behavior of the participants with respect to the target features. This general behavior is determined by the set of values of the target features produced by the participants in each condition. Since this analysis checks whether the participants behave differently as a whole towards

[1] version 6.0.35

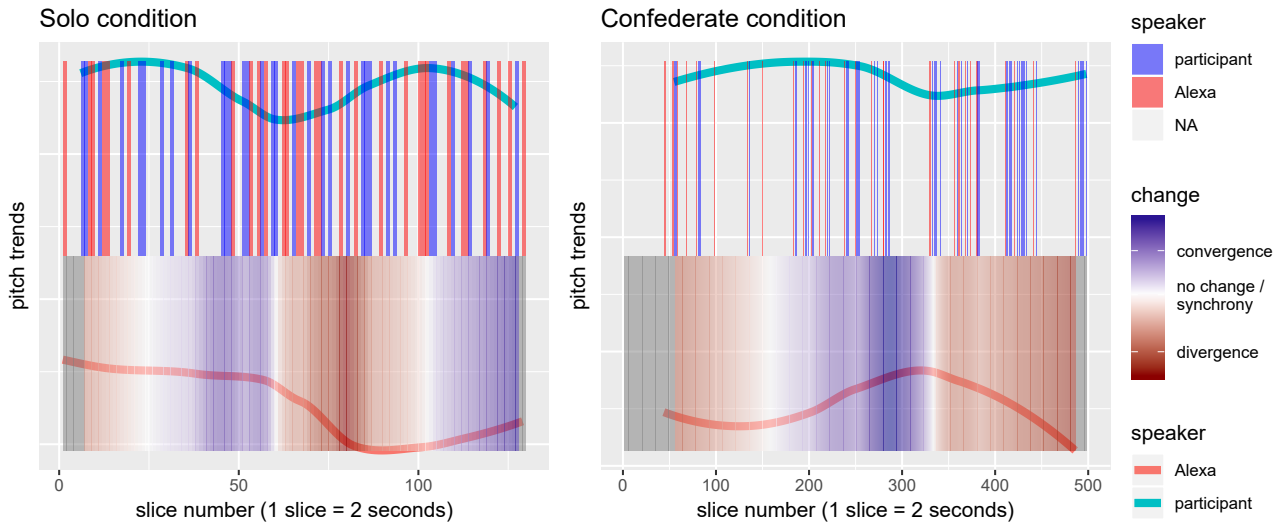


Figure 2: *A comparison between the behavior of the $f_0$ feature in solo condition (left) and confederate condition (right). The lines represent the LOESS smoothing trend lines of the participant (blue) and Alexa (red). Omitted turns, e.g., turns of the confederate and turned removed as explained in Section 3 are not colored (gray). The vertical bars in the upper half represent the turns of the participant (blue) and Alexa (red). The color-scaled vertical bars at the bottom half are the convergence/divergence level of the participant over time as calculated in Equation* (2)*. Blue areas represent convergence while red areas represent divergence. The darker the color, the greater the effect, with white color pointing to points of no change (or synchrony, in segments with both trends moving the same way).*

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 interactions 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

$$change_t = -\Delta_{t,t-1} \mid S_{part} - S_{Alexa} \mid, \quad (1)$$

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 to accommodation is calculated by

$$accomm(participant)_t = change_t - \Delta_{t,t-1} S_{Alexa}. \quad (2)$$

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.

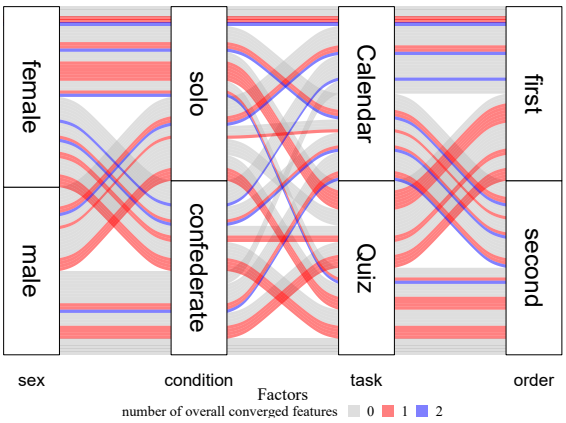


Figure 3: *Overview of the relation between the factors* sex, condition, task, *and* order *and the number of features that showed more convergence in total across the interactions. Each line represents one interaction. The* sex *stratus refers to the sex of the participant, and the* order *stratus to the position of an interaction in the order in which the task-condition combination was performed. The color of a line stands for the number of target features that showed overall convergence in this interaction, from none (zero features, in gray), through one (red), and up to two (blue). For example, a blue line going through the strata sequence female → solo → Quiz → first represents an interaction with a female participant performing the Quiz task in solo condition first, where the participant converged in two out of the three target features.*

## 5. Discussion and conclusion

We have presented distributional and temporal analyses of differences in convergence of three vocal features in HCI, 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


[1] J. Osborne. (Jul. 2016). Why 100 million monthly Cortana users on Windows 10 is a big deal, [Online]. Available: https://www.techradar.com/news/software/operating-systems/why-100-million-monthly-cortana-users-could-be-a-bigger-deal-than-350-million-windows-10-installs-1325146.

[2] M. R. Dickey. (Dec. 2017). The Echo Dot was the best-selling product on all of Amazon this holiday season, [Online]. Available: https://techcrunch.com/2017/12/26/the-echo-dot-was-the-best-selling-product-on-all-of-amazon-this-holiday-season/.

[3] S. Kleinberg. (Jan. 2018). 5 ways voice assistance is shaping consumer behavior, [Online]. Available: https://www.thinkwithgoogle.com/_qs/documents/5604/1178-CES-Voice-Research-PDF.pdf.

[4] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: Communication, context, and consequence", in *Contexts of Accommodation: Developments in Applied Sociolinguistics*, H. Giles, J. Coupland, and N. Coupland, Eds., Cambridge University Press, 1991, pp. 1–68. DOI: 10.1017/CBO9780511663673.001.

[5] C. Gallois and H. Giles, "Communication accommodation theory", in *The International Encyclopedia of Language and Social Interaction*, K. Tracy, C. Ilie, and T. Sandel, Eds., Wiley, 2015, pp. 1–18. DOI: 10.1002/9781118611463.wbielsi066.

[6] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers", *Attention, Perception, & Psychophysics*, vol. 79, no. 2, pp. 637–659, Feb. 2017. DOI: 10.3758/s13414-016-1226-0.

[7] A. Schweitzer, N. Lewandowski, and D. Duran, "Social attractiveness in dialogs.", in *Interspeech*, Aug. 2017, pp. 2243–2247. DOI: 10.21437/Interspeech.2017-833.

[8] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean, "Linguistic alignment between people and computers", *Journal of Pragmatics*, vol. 42, no. 9, pp. 2355–2368, Sep. 2010. DOI: 10.1016/j.pragma.2009.12.012.

[9] R. Levitan, 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*, San Francisco, CA, USA, Sep. 2016, pp. 1166–1170. DOI: 10.21437/Interspeech.2016-985.

[10] I. Siegert, 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 Language and Body in Real Life & Multimodal Corpora*, Miyazaki, Japan, May 2018. [Online]. Available: http://lrec-conf.org/workshops/lrec2018/W20/pdf/13_W20.pdf.

[11] V. Silber-Varod, A. Lerner, and O. Jokisch, "Prosodic plot of dialogues: A conceptual framework to trace speakers role", in *International Conference on Speech and Computer*, Sep. 2018, pp. 636–645. DOI: 10.1007/978-3-319-99579-3_65.

[12] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions", in *Interspeech*, Aug. 2011, pp. 3081–3084. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2011/i11_3081.html.

[13] M. Natale, "Convergence of mean vocal intensity in dyadic com-munication as a function of social desirability", *Journal of Personality and Social Psychology*, vol. 32, no. 5, pp. 790–804, 1975. DOI: 10.1037/0022-3514.32.5.790.

[14] S. Gregory, S. Webster, and G. Huang, "Voice pitch and amplitude convergence as a metric of quality in dyadic interviews", *Language & Communication*, vol. 13, no. 3, pp. 195–217, 1993. DOI: 10.1016/0271-5309(93)90026-J.

[15] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically", *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, May 2009. DOI: 10.3758/BRM.41.2.385.

[16] P. Boersma, "Praat, a system for doing phonetics by computer", *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[17] F. Wilcoxon, "Individual comparisons by ranking methods", *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, Dec. 1945. DOI: 10.2307/3001968.

[18] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting", *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988. DOI: 10.1080/01621459.1988.10478639.