Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Pausing strategies in dialogue speech: the interlocutor factor

Tatiana Kachkovskaia[1], *Daniil Kocharov*[2]

[1]Independent Researcher, Finland;
[2]Speech and Cognition Research Group, Tampere University, Finland
tania.kachkovskaya@gmail.com, daniil.kocharov@tuni.fi

Abstract

In everyday conversations, speakers tend to adapt their voices in order to fit the communication situation. One of the factors that seems to influence speech variability is the interlocutor factor: it has been shown that some linguistic features of our speech are dependent on who we are talking to. In our previous work, we observed interlocutor-related changes in speech tempo and frequency of non-speech events. In this research, we are presenting data on interlocutor-induced variability in within-turn pausing strategies, involving frequency of silent intervals, filled pauses and prosodic boundaries, as well as overall percentage of time taken by silent intervals in speech. The results were obtained using the large annotated speech corpus SibLing where each of the 20 core speakers participated in 5 dialogue settings: with a sibling, a friend, 2 strangers (a male and a female), and a person of elder age whose job requires leadership skills.

Index Terms: pause, intra-speaker variability, prosodic phrasing, dialogue speech

1. Introduction

In natural communication, speech flow is often interrupted by relatively short breaks that are called pauses. It is known that pausing strategies are variable across speakers, speaking styles, and types of speech [1, 2, 3]; pausing is also language-dependent [4, 5]. This research is aimed at investigating another source of variability in dyadic conversations—the interlocutor factor. That is, we assume, that a speaker's pausing strategy may change depending on who he/she is talking to.

The idea of interlocutor-induced variability in speech roots in the research area of communication accommodation, also termed, among other, "speech entrainment" and "convergence". It is assumed that in a dialogue the interlocutors may "attune" to each other in a number of features, from gestures and body movements to using similar syntax [6] or lexemes [7], and becoming more similar in the acoustic features: speech tempo, intensity, and fundamental frequency [8, 9]. It is not necessary, though, that in each conversation each pair would entrain, or that it would happen in a particular way. The way entrainment works depends on a number of social and situational factors. There is influence of the type of activity that speakers do (e.g. free conversation vs. collaborative task), speakers' gender (female-female vs. male-male vs. mixed gender pairs), conversational roles (e.g. interviewer vs. interviewee) and other [10].

Given that speech entrainment is socially and situationally dependent, it seems reasonable to infer that some of the within-speaker variation in speech is explained by the relationship between the interlocutors. And if the underlying mechanisms are indeed socially-related, this variation must be observed systematically across speakers from the same social group. This idea is in the centre of the presented research. The main hypothesis is therefore as follows: "There is systematic within-speaker variation in pausing strategies caused by the interlocutor factor".

In order to talk about pausing strategies, we need to elaborate on the notion of a pause. In natural speech, the speech flow is divided into meaningful chunks between which listeners perceive some breaks (interruptions). Sometimes, such breaks are filled with silence, and these are relatively transparent examples of pauses. In practice, however, such breaks are often filled with something else—a filler (e.g. "uhm"), sounds of breathing, clicks, and other non-speech noises. In fact, the way pauses are filled (or not filled) is also variable, and it is therefore a part of pausing strategy.

All the above-mentioned cases have a common feature: they are all measurable in terms of duration, independent of what the filler is. Sometimes, though, a prosodic break is perceived, but there is no measurable "pause" in the speech signal; in such cases the boundary is marked by other cues—mostly, melodic and temporal (lengthening of final segments). Usually, such breaks are not treated as pauses, as there is no actual pause. In this research, however, we are looking at prosodic breaks as part of pausing strategy, because they are very similar to pauses in terms of function, even though manifested by zero duration.

In addition, it is worth noting that measurable pauses do not always occur at boundaries of meaningful chunks; sometimes, especially in unprepared speech, a silent interval or a filler may be produced as a result of poor speech planning, providing the speaker more time to put their ideas into words. This is also a part of pausing strategy, as it indicates how much control the speaker has in the conversational situation, and, therefore, subject to variation.

Below, we will present our analysis of within-speaker variation in terms of components of pausing mentioned above. At this stage, only within-turn pauses will be analysed (although strategies of turn-taking could obviously contribute to the overall picture). In order to test our hypothesis, we are using the SibLing speech corpus, where each of the 20 speakers were recorded in controlled settings so that he/she participated in 5 dialogues with different interlocutors of varying "social distance". This corpus is thus a unique resource for research of within-speaker variation.

At the present moment, there are not many publications that addressed variability due to the factor of social distance, mainly because it is hard to collect speech material that would cover this aspect. Still, there are some publications that compared two degrees of familiarity: familiar vs. unfamiliar [11], strangers vs. friends [12], strangers vs. spouses [13]. There are differences in the amount of disfluencies (lower in unfamiliar pairs); the degree of speech entrainment (higher in friends than strangers, and even higher in couples); the frequency of turn overlaps (lower in married couples). Previous publications based on the SibLing corpus have reported on interlocutor-induced differences in speech tempo, disfluencies, frequency of filled pauses, and laughter (for more detail, see Section 4).

10.21437/SpeechProsody.2024-124

2. Material

This research is based on the corpus of Russian dialogue speech SibLing [14]. This large corpus was designed for research on communication accommodation (entrainment) in speech, with particular focus on variability in the social distance between the interlocutors.

The Sibling is composed of 90 dialogues that are organized in the following way. The basic set of speakers are 10 pairs of same-gender siblings aged 23 to 40 (10 male speakers and 10 female). Each of the speakers from the basic set was recorded in five dialogue settings: (1) with a sibling, (2) with a close friend of the same gender and similar age, (3) with a stranger of the same gender and similar age, (4) with a stranger of the opposite gender and similar age, and (5) with a stranger of the same gender and greater age whose job requires leadership skills. To enable this setup, more speakers were involved as interlocutors; their age ranged between 20 and 70.

As a result, this corpus enables to compare the same speaker's features in dialogues with different social distance between the interlocutors. The social distance factor is composed of the degree of familiarity (siblings-friends-unfamiliar), gender (in dialogues with strangers, male vs. female interlocutor of similar age), and social hierarchy (in dialogues with strangers, same vs. elder age, with potential influence of job position).

The conversations were organized in the form of collaborative tasks: first, a card-matching game, and then the classical map task. The card-matching game consisted in searching for similar items in two decks of ten whimsical cards. After finding a suitable pair of cards, the interlocutors were instructed to switch roles. The game was limited in time by approx. 12 minutes. In the map task, the speakers were required to guide each other through a set of 4 schematic maps, switching roles after each map. The task took 15 to 60 minutes to complete.

The corpus contains orthographic, phonetic and, partly, expert prosodic annotation; it is segmented into inter-pausal units, turns and conversational roles (leader vs. follower). Within the orthographic tier, the annotators manually registered all non-speech events (hesitations, false-starts, within-word interruptions) and marked lexical stress. Prosodic annotation is only provided with the card game recordings, and includes segmentation into intonational phrases, determining the nucleus placement and its melodic type (see [14]). Subsequent alignment of the annotation with speech signal was performed automatically.

3. Method

Based on the corpus annotation, the following features were calculated automatically.

Within-turn silent intervals (SI): frequency and average duration (median) of silent intervals occurring within each turn of a given speaker in a given dialogue; also, the percentage of time taken by silent intervals within all speaker's turns. SI detection was performed automatically with the threshold of 200 ms. To avoid bias caused by turns without internal pauses, we excluded the latter from the analysis.

Prosodic breaks: frequency of intonational-phrase (IP) boundaries (per minute), average IP duration. The calculations are based on prosodic annotation provided with the corpus, and as a result, were obtained for the card-matching game only.

The ratio between the number of silent pauses and the total number of IP boundaries (SI to IP ratio): approx. how many IP boundaries are marked by silent intervals. (Similar to the previous feature, it is based on prosodic annotation and, therefore, was calculated for the card-matching game only.) In read speech, the values of this parameter are about 60 % [15]. In unprepared speech, this ratio has high variation (in [15], 39 to 90 %; in [3], similar variation is observed for the percentage of clause endings marked by pauses), which leads us to assume that the interlocutor factor may also manifest itself in this parameter. High values of this ratio, in general, occur in non-fluent speech with significant self-interruptions.

Filled pauses: frequency of pauses filled with vowel-like segments. Filled pauses are typical for unprepared speech and are related to speech planning. Their frequency is variable, and may reach the value of approx. 0.3 filled pauses per second [15].

Statistical analysis was performed in R using rstatix[16]. Repeated measures ANOVA was run using "anova_test" and included within-subject factors of social distance (5 categories), conversational role (Leader vs. Follower) and type of task (card game vs. map task), and a between-subject factor of speaker's gender (f vs. m). Extreme outliers were detected using "identify_outliers" and removed from the analysis. Post-hoc pairwise t-testing was performed with "pairwise_t_test", and the Bonferroni adjustment method was used.

In tables and figures, we will be using simplified terms for the 5 dialogue settings reflecting:

- sibling: with a same-gender sibling;
- friend: with a same-gender friend of similar age;
- stranger1: with a stranger of the same gender and similar age;
- stranger2: with a stranger of the opposite gender and similar age;
- boss: with a stranger of the same gender and greater age whose job requires leadership skills.

4. Results and discussion

4.1. Average duration of silent intervals

Median silent interval duration did not reveal any influence of the factor of social distance ($F(4, 28) = 1.65$, $p = 0.19$, $\eta_g^2 = 0.05$). The only factor that significantly influenced the SI median value was the type of task ($F(1, 7) = 5.74$, $p = 0.048$, $\eta_g^2 = 0.09$), with greater values for the card game: 0.616 seconds (SD=0.143) vs. 0.519 (SD=0.133). It probably means that in the card game, the participants needed more time to look through their pictures and decide which objects to discuss, while in the map task the items were easier to find as there were less of them and they were located along the route.

From previous research using the same data [17] we know that speech tempo, measured in syllables per second and excluding pauses (articulation rate), is slightly higher in map task compared with card game. Given this information, shorter silent intervals in map task are another indicator of slower tempo.

4.2. Percentage of time taken by silent intervals

The percentage of time taken by silent intervals, again, did not reveal any statistically significant influence of the factor of social distance ($F(2.08, 20.75) = 2.49$, $p = 0.11$, $\eta_g^2 = 0.06$). The average values were as follows: sibling 20.2 % (SD=7.3); friend 21.9 % (SD=7.9); stranger1 21.5 % (SD=5.5); stranger2 24.3 % (SD=8.3); boss 22.4 % (SD=8.6). In general, we might be observing a tendency of slightly higher values with the growth of social distance.

Among other factors, the type of task had borderline statistical significance ($F(1, 10) = 4.8$, $p = 0.053$, $\eta_g^2 = 0.04$); in

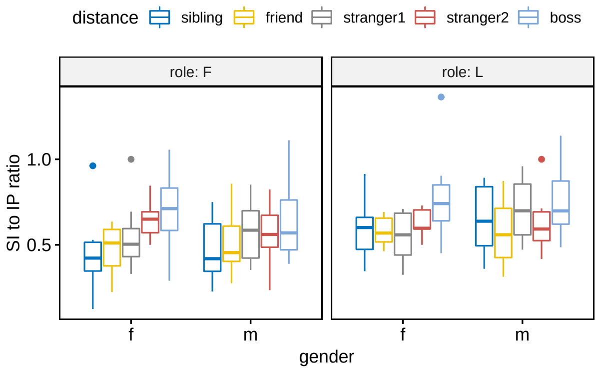

Figure 1: *The ratio between the number of silent intervals (SI) and the number of IP boundaries grouped in accordance with the speaker's role (leader vs. follower), speaker's gender and the interlocutor*

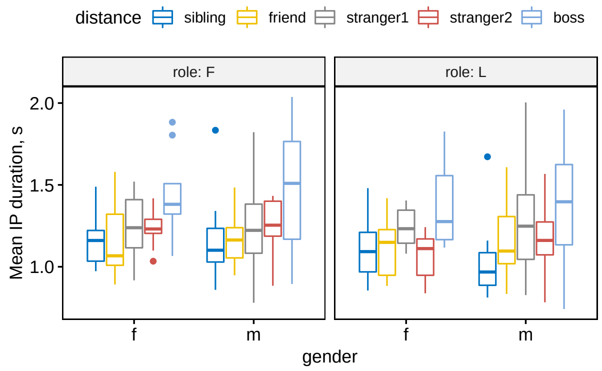

Figure 2: *Mean duration of intonational phrase (in seconds) grouped in accordance with the speaker's role (leader vs. follower), speaker's gender and the interlocutor*

addition, we observed interaction between the factors of task and role, as well as between role, task, and social distance. On average, card game had slightly more silence in it: 23.9 % (SD=8.4), compared with 20.3 % (SD=6.4) for map task.

4.3. Frequency of silent intervals

The number of silent intervals (SI) per 1 second of speech including pauses did not reveal statistically significant influence of social distance ($F(4, 44) = 1.36$, $p = 0.26$, $\eta^2_g = 0.02$). The average values were: sibling 0.276 (SD=0.067); friend 0.289 (SD=0.057); stranger1 0.3 (SD=0.053); stranger2 0.309 (SD=0.061); boss 0.302 (SD=0.075). The general trend is as follows: in dialogues with unfamiliar people, speakers tend to insert silent intervals slightly more frequently.

Our analysis has revealed some influence of conversational role ($F(1, 11) = 6.23$, $p = 0.03$, $\eta^2_g = 0.03$) as well as the interaction of role and task. In card game, leaders tend to insert SIs more frequently: compare mean values 0.323 (SD=0.063) and 0.263 (SD=0.058). In map task, the situation is reversed: 0.286 (SD=0.046) vs. 0.309 (SD=0.07).

4.4. Silent intervals to IP boundaries ratio

The ratio between the number of silent intervals and the number of intonational-phrase (IP) boundaries shows variation due to the factor of social distance ($F(2.5, 42.42) = 5.77$, $p = 0.004$, $\eta^2_g = 0.14$). Another statistically significant factor was speaker's role ($F(1, 17) = 19.46$, $p < 0.001$, $\eta^2_g = 0.06$). The factor of speaker's gender was not statistically significant ($F(1, 17) = 0.005$, $p = 0.9$, $\eta^2_g = 0.0001$), and the effect of task type could not be measured as this parameter could only be calculated for the card game. No significant interactions were observed. The distribution of the data is presented in Figure 1.

Table 1 provides average values of the SI to IP ratio across social distance situations. Pairwise comparison enabled us to reveal more detailed systematic variation: dialogues with the "boss" differed from each of the other dialogue settings. Statistically significant differences were observed in the settings "boss vs. sibling" (difference of means -0.189, $p = 0.002$, $CI[-0.283, -0.095]$), "boss vs. friend" (difference of means -0.185, $p < 0.001$, $CI[-0.278, -0.108]$), "boss vs. stranger1" (difference of means -0.132, $p = 0.008$, $CI[-0.205, -0.059]$), "boss vs. stranger2" (difference of

Table 1: *Ratio between the number of silent intervals and the number of IP boundaries depending on the interlocutor*

	Mean	SD
sibling	0.536	0.208
friend	0.54	0.164
stranger1	0.593	0.179
stranger2	0.613	0.138
boss	0.725	0.229

means -0.112, $p = 0.037$, $CI[-0.199, -0.042]$).

We may conclude that the conversational setting with an unfamiliar person of greater age whose job requires leadership skills this parameter is higher than in other settings. This means that speakers either mark more IP boundaries with silence, or use more silent breaks within IPs; that is, they either speak more distinctly in terms of pausing, or, on the opposite, less fluently.

In our previous work [17], we already addressed the question of fluency using the same material. We calculated the frequency of false-starts (e.g. "to- tomorrow" or "to- yesterday") and word splits (e.g. "to [break] morrow"), i.e. those disfluencies that cause IP-internal silent intervals. The factor of social distance had borderline significance (p=0.056), but, more importantly, the most disfluencies were observed in conversations with a friend, while the least—with unfamiliar interlocutors, including "boss". For this research, it means that higher SI to IP ratio in the "boss" setting could not be explained by more disfluencies, but, on the opposite: speaking with a "boss", one uses more distinct phrasing (along with less disfluencies).

The general trend for this parameter is that it gets higher with the growth of social distance (see Table 1). The lowest amount of silences at IP boundaries is observed in the sibling-to-sibling and sibling-to-friend settings, which are most comfortable in terms of mutual understanding.

4.5. Mean IP duration

The average (mean) duration of an intonational phrase (IP) reveals systematic variation due to the factor of social distance ($F(4, 64) = 10.79$, $p < 0.0001$, $\eta^2_g = 0.2$) and some influence of conversational role ($F(1, 16) = 4.88$, $p = 0.042$, $\eta^2_g = 0.02$, with followers having slightly greater values); the factor of gender did not reveal statistical significance, and the effect of task could not be measured as this parameter could only be calculated for the card game.

The influence of social distance was observed mainly because of the outstanding difference for the “boss” setting, which can be seen clearly in Figure 2. The average values were as follows: sibling 1.12 (SD=0.22); friend 1.16 (SD=0.2); stranger1 1.25 (SD=0.25); stranger2 1.18 (SD=0.18); boss 1.43 (SD=0.32). Pairwise comparisons have shown statistically significant differences in the settings “boss vs. sibling” (difference of means -0.31, $p < 0.0001$, $CI[-0.392, -0.185]$), ”boss vs. friend” (difference of means -0.27, $p < 0.0001$, $CI[-0.366, -0.175]$), ”boss vs. stranger1” (difference of means -0.18, $p = 0.001$, $CI[-0.253, -0.0895]$), ”boss vs. stranger2” (difference of means -0.25, $p < 0.0001$, $CI[-0.324, -0.167]$). In addition, there was significant difference for the settings ”sibling vs. stranger1”: in dialogues with a sibling, one uses shorter IPs than with an unfamiliar person of the same age and gender (difference of means -0.13, $p = 0.025$, $CI[-0.181, -0.042]$).

In general, there is a slight tendency for longer IP mean with the growth of social distance (see Figure 2). When the distance is the greatest (unfamiliar person, age difference and, potentially, hierarchy caused by job position of the “boss”), the IP duration reaches significantly higher values.

4.6. Frequency of filled pauses

Figure 3 presents the frequency of filled pauses for the conversational role of Leader (for the other role, there were not enough data). Only the factor of social distance was statistically significant ($F(4, 44) = 3.32$, $p = 0.019$, $\eta_g^2 = 0.09$). The average values were as follows: sibling 0.094 (SD=0.054); friend 0.117 (SD=0.064); stranger1 0.112 (SD=0.061); stranger2 0.125 (SD=0.058); boss 0.144 (SD=0.069). Pairwise comparisons revealed difference between the ”boss” setting and two other settings: “sibling” (difference of means -0.05, $p = 0.0002$, $CI[-0.066, -0.027]$) and ”friend” (difference of means -0.028, $p = 0.027$, $CI[-0.048, -0.011]$).

As we can see in Figure 3, there is a general trend for this parameter to get higher along with social distance. Thus, there is a possibility that greater social distance may cause speakers to fill more of their pauses with vowel-like segments. Since filled pauses are often associated with hesitations, this might mean that in settings with strangers, especially the ”bossy” ones, speakers feel uncomfortable and, as a result, speech planning requires more time. On the other hand, higher number of fillers could be a consequence of using longer intonational phrases (as seen in Figure2). It is worth noting, though, that this parameter only reflects vowel-like fillers, but completely dismisses e.g. the lexicalized ones (e.g. ”well”), that are quite frequent in this corpus.

5. Conclusions

This research has enabled us to reveal some systematic variation in the way the interlocutor factor influences the use of pauses in speech. With the growth of social distance between the interlocutors, we observe a slight increase in the frequency of silent intervals and the percentage of time taken by silent intervals. Often, the lowest values are observed in dialogues between siblings and friends, and the highest - in dialogues with strangers of the same gender and greater age whose job requires leadership skills (”bosses”).

More statistically reliable data were obtained for those features that are related to division into intonational phrases. With

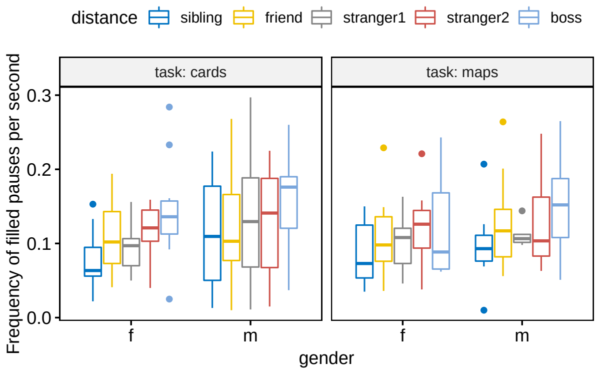

Figure 3: *Frequency of filled pauses (per second) grouped in accordance with the type of collaborative task (card game vs. map task), speaker’s gender and the interlocutor*

the growth of social distance, we observe an increase in average IP duration and the ratio between the number of silent intervals and IP boundaries. The percentage of filled pauses with vowel-like fillers shows the same tendencies. At the same time, self-repairs and interrupted words are less frequent in dialogues with strangers. This means that (1) the pausing strategy for closely-related interlocutors is to use more IP-internal pauses, but less silent intervals at IP boundaries, shorter IPs and slightly less silent intervals; in addition, more laughter is used in such conversations [17]. On the other hand, (2) the pausing strategy for dialogues with stranger is of the opposite nature. (3) The pausing strategy for dialogues with bosses is an intensified variant of the latter one.

The use of less word interruptions, more pauses, longer IPs and marking more IP boundaries with silent intervals—these are all features of more careful, elaborate speech. The higher number of filled pauses, though, is indicative of disfluency. It may be the case that the social situation of collaborating with a stranger requires speaking slowly and carefully, but, if the speaker needs more time for planning, a filled pause is a perfectly appropriate instrument.

In casual conversations, other instruments are used for planning, such as word interruptions and shorter IP duration. It is also possible that in dialogues with closer people, speech planning goes easier, as the interlocutors feel more comfortable.

As already mentioned, settings with ”bosses” stand apart from other dialogues with strangers. The other two settings—dialogues with unfamiliar people of the same vs. opposite gender (but similar age)—never differ from each other significantly. However, as we can see in Figures 1 and 2, it is sometimes the case that the values are slightly greater in same-gender dialogues. This is the opposite of what was expected, as mixed gender pairs have greater social distance (unfamiliarity + gender difference). But it is possible that in mixed-gender conversations, other social and situational factors may interfere, such as e.g. likability.

In a similar way, dialogues between friends sometimes reveal slightly greater values than dialogues between siblings, violating the general trend. Initially, we assumed that sibling-to-sibling conversations represent a case of the smallest social distance. But in fact, we do not choose our siblings, but we do choose our friends; between siblings, there is often competition that may increase the social distance. This could mean that friend-to-friend relationship is actually the closest.

6. References

[1] S. Gustafson-Capkova and B. Megyesi, "A comparative study of pauses in dialogues and read speech," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[2] P. Šturm and J. Volín, "Occurrence and duration of pauses in relation to speech tempo and structural organization in two speech genres," *Languages*, vol. 8, no. 1, p. 23, 2023.

[3] M. E. Van Donzel and F. J. Koopmans-van Beinum, "Pausing strategies in discourse in Dutch," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2, 1996, pp. 1029–1032.

[4] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Proc. Speech prosody 2002*, 2002.

[5] R. Werner, J. Trouvain, and B. Möbius, "Optionality and variability of speech pauses in read speech across languages and rates," in *Proc. Speech Prosody 2022*, 2022, pp. 23–26.

[6] D. Reitter and J. D. Moore, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 28, no. 28, 2006.

[7] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker, "Language style matching predicts relationship initiation and stability," *Psychological science*, vol. 22, no. 1, pp. 39–44, 2011.

[8] R. Levitan and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *INTERSPEECH*, 2011, pp. 3081–3084.

[9] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Attention, Perception, & Psychophysics*, vol. 79, pp. 637–659, 2017.

[10] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and speech*, vol. 44, no. 2, pp. 123–147, 2001.

[11] J. Reverdy and C. Vogel, "Measuring synchrony in task-based dialogues." in *INTERSPEECH*, 2017, pp. 1701–1705.

[12] A. Lelong and G. Bailly, "Study of the phenomenon of phonetic convergence thanks to speech dominoes," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues: COST 2102 International Conference, Budapest, Hungary, September 7-10, 2010, Revised Selected Papers*. Springer, 2011, pp. 273–286.

[13] J. M. Twenge, "Changes in women's assertiveness in response to status and roles: A cross-temporal meta-analysis, 1931–1993." *Journal of personality and social psychology*, vol. 81, no. 1, p. 133, 2001.

[14] T. Kachkovskaia, T. Chukaeva, V. Evdokimova, P. Kholiavin, N. Kriakina, D. Kocharov, A. Mamushina, A. Menshikova, and S. Zimina, "SibLing corpus of Russian dialogue speech designed for research on speech entrainment," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6556–6561.

[15] T. Kachkovskaia and P. Skrelin, "Prosodic phrasing in Russian spontaneous and read speech: evidence from large speech corpora," in *Proc. Speech Prosody 2020*, 2020, pp. 166–170.

[16] A. Kassambara, *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2023, r package version 0.7.2. [Online]. Available: https://rpkgs.datanovia.com/rstatix/

[17] T. Kachkovskaia, A. Mamushina, D. Kocharov, P. Kholiavin, A. Menshikova, D. Guseva, A. Portnova, V. Evdokimova, and S. Zimina, *Kommunikativnaja adaptatsija v dialoge: vzgl'ad fonetista (Communication Accommodation in Dialogues. A Phonetician's Perspective) [in Russian]*. Saint Petersburg: Saint Petersburg University Publishing House, 2022.