PHONETIC CONVERGENCE EVALUATION BASED ON FUNDAMENTAL FREQUENCY VARIABILITY

Bistra Andreeva[1], *Grażyna Demenko*[2], *Jolanta Bachan*[2], *Iona Gessinger*[1], *Karolina Jankowska*[2], *Bernd Möbius*[1]

[1]*Department of Language Science and Technology, Saarland University, Germany*
[2]*Faculty of Modern Languages and Literatures, Adam Mickiewicz University, Poznań, Poland*
andreeva@lst.uni-saarland.de

Abstract: There is a growing body of studies investigating whether interlocutors become more similar to each other during a dialogue. The present study contributes to this research line by investigating the pitch profiles, and convergence and synchrony in mean F0 values, in Polish dyadic conversations on provocative art between students and between a student and a teacher. We found different pitch profiles for the different scenarios and identities of the interlocutors. In the student-student and student-teacher conversations where both interlocutors agree (whether or not they accept the provocative art), the students show higher long-term distributional (LTD) F0 values for level, span and standard deviation when interacting with a teacher than with a fellow student. In the student–teacher conversations the students achieve significantly higher LTD F0 values when the two interlocutors do not share the same opinion about the provocative art. Regarding convergence we applied a variety of measures to our data (local, global convergence and synchrony). Using these measures, we found evidence of local and global convergence in the student-teacher conversations and of synchrony in both student-student and student-teacher conversations.

1 Introduction

According to Communication Accommodation Theory [1], individual adjustments of speech characteristics between interlocutors in human-human interaction are assumed to subserve the function of controlling social distance and achieving successful communication. The tendency of interlocutors to become more similar in many aspects of their speaking behavior has been referred to in various disciplines as convergence (e.g. [2]), entrainment (e.g. [3]), alignment (e.g. [4], [5]), accommodation (e.g. [1], [6]), or adaptation (e.g. [7]). In addition to the choice of words (e.g. [8], [9]) or syntactic constructions (e.g. [10], [11]), convergence has been documented for many prosodic and intonational features such as F0, intensity, or temporal aspects including speech rate, rhythm or turn-taking initiation (e.g. [3], [12], [13], [14]). The increase in similarity of speech patterns between interlocutors in the course of a conversation has been studied for many years and for various languages (French in [15]; English in [16], [17], [18]; German in [19], [20], [21], [22]; Italian in [23]; Japanese in [24]; Polish in [25]; Slovak in [26], and Swedish in [27]). The outcome of this research shows that prosodic convergence, especially in collaborative tasks, is rather complex and is affected by the nature of the task, the hierarchy between speaker and interlocutor, the perceived attractiveness and likability of the interlocutor, the visibility of the interlocutor, biological sex, and potentially many other interactional features of spoken conversations. This makes prosodic convergence a challenging but also exciting area of future research.

The present study is concerned with the analysis of within-speaker variations based on long-term distributional (LTD) F0 measures in Polish dyadic conversations, and addresses the following questions:

(1) Do speakers produce different LTD pitch profiles depending on the modality of the dialogue and the social status and sex of their interlocutor?

(2) Do speakers show convergence and synchrony in mean F0, depending on the modality of the dialogue, and the social status and sex of their interlocutor?

2 Materials and method

2.1 Speech corpus

We used a subset of the Polish Harmonia corpus [28] consisting of student-student (6 female and 7 male pairs) and student-teacher (each student with a female teacher) conversations on provocative modern art in three scenarios: (a) both interlocutors accept the provocative art [+like; +agree], (b) both interlocutors do not accept the provocative art [-like; +agree], (c) the teacher does not accept the provocative art while the student accepts it [±like;-agree]. The role of the teacher was always performed by the same female interlocutor (a real university teacher). Table 1 summarizes the dialogue scenarios with respect to the interlocutors and dialogue modalities.

Table 1 – Dialogue scenarios and participants.

dialogue ID	interlocutors	modality	pair numbers and sex
d12	student-student	[+like; +agree]	6 female, 7 male
d13	student-student	[-like; +agree]	6 female, 7 male
d14	student-teacher	[+like; +agree]	12 female, 14 mixed-sex
d15	student-teacher	[-like; +agree]	12 female, 14 mixed-sex
d16	student-teacher	[±like; -agree]	12 female, 14 mixed-sex

2.2 Method

Inter-pause stretch (IPS) boundaries were manually marked. IPS containing back-channels were excluded from the analysis. F0 was extracted automatically in all IPS, using the RAPT algorithm [29] implemented in the `get_f0` function from the ESPS software package with time steps of 5 ms for female and 10 ms for male speakers. Irregular voiced stretches of speech caused by laryngealization were excluded from further analysis. The following long-term distributional (LTD) measures were calculated per IPS individually: mean and median F0 values for pitch level (in Hz), interquartile range (IQR, in Hz) and pitch range (in semitones) between maximum and minimum pitch values per IPS for span, and standard deviation (SD) for variation of F0 distribution (in Hz).

To investigate if the interlocutors become more similar to each other we adopt the method described by Edlund et al. [27] and measure convergence and synchrony. Following Levitan et al. [30], we distinguish between global and local convergence. Global convergence corresponds to the similarity between two speakers over the course of a conversation and is measured using feature means over the conversation. Local convergence captures the dynamic similarity between two speakers over time within a conversation, as measured by the turn-level similarity. Turn is defined as a sequence of IPS from a single speaker. For each dialogue we first extracted

mean F0 values for the first and last 30% of the dialogue duration. In a second step, we identified the final (target) IPS of speaker A's turn and the initial (partner) IPS of the corresponding speaker B's turn. For each of these turns mean F0 values were also extracted. The F0 values were normalized by speaker sex using z-scores.

3 Results

3.1 Pitch profiles

3.1.1 Pitch profiles depending on dialogue modality and speaker sex

To answer the first research question, i.e., to what extent male and female speakers produce different pitch profiles depending on the modality of the dialogue, a first analysis compared students' pitch profiles in dialogues with other students or with the teacher in the three modalities. The pitch patterns were analyzed in linear mixed models by means of the JMP software [31] with the respective measure as dependent variable, Speaker as random variable, Modality ([+like; +agree], [-like; +agree], and [±like; +agree]) and Sex (female, male) as fixed factors, as well as the interaction of Modality and Sex. Separate Tukey post-hoc tests were carried out per variable, if appropriate. The confidence level was set at α=0.05.

A systematic comparison of the LTD measures of F0 showed that, predictably, Sex had a significant main effect on all LTD measures, with females having significantly higher F0 values for mean ($F [1, 24] = 163.3$, $p<0.001$), median ($F [1, 24] = 158.3$, $p<0.001$), IQR ($F [1, 24] = 45.2$, $p<0.001$), span ($F [1, 24] = 10.8$, $p<0.01$), and SD ($F [1, 24] = 68.3$, $p<0.001$). However, beyond the expected Sex effect, there was also a significant main effect of Modality on all LTD measurements: for mean ($F [2, 3837] = 119.0$, $p<0.001$), median ($F [2, 3837] = 94.9$, $p<0.001$), IQR ($F [2, 3837] = 39.9$, $p<0.001$), span ($F [2, 3837] = 20.5$, $p<0.001$), and SD ($F [2, 3840] = 62.1$, $p<0.001$). Separate post-hoc tests showed that the F0 measures were significantly higher in the expressive dialogues in which both speakers disagreed than in the dialogues in which both speakers agreed, regardless of whether they both accepted or disliked the provocative art (see Table 2).

Table 2 – LTD measures by dialogue modality (standard deviation in parentheses).

parameter	**sex**	**[±like; +agree]**	**[+like; +agree]**	**[-like; +agree]**
mean (Hz)	m	135.4 (30.1)	121.3 (25.8)	121.8 (23.0)
	f	246.0 (43.5)	225.8 (42.8)	227.1 (46.2)
median (Hz)	m	130.7 (30.2)	118.2 (26.0)	118.6 (22.3)
	f	241.2 (43.5)	222.6 (43.0)	225.1 (45.5)
pitch range	m	16.8 (7.8)	15.2 (7.1)	15.0 (7.1)
(semitones)	f	19.5 (9.2)	17.6 (9.0)	17.8 (8.8)
IQR (Hz)	m	33.0 (31.2)	23.5 (20.9)	24.5 (24.6)
	f	63.5 (40.7)	51.9 (42.6)	53.0 (41.7)
SD (Hz)	m	29.6 (22.0)	22.3 (15.5)	22.5 (18.1)
	f	52.5 (23.8)	44.4 (24.7)	44.4 (22.9)

The statistical analysis revealed a significant interaction between Modality and Sex for mean F0 ($F [2, 3837] = 7.2$, $p<0.001$) and median F0 ($F [2, 3837] = 7.1$, $p<0.001$). This interaction can be explained by the higher F0 register used by the female speakers compared to the male speakers. Male speakers have significantly higher mean and median F0 values in the disagree condition compared to the agree condition. The same pattern is observed for the

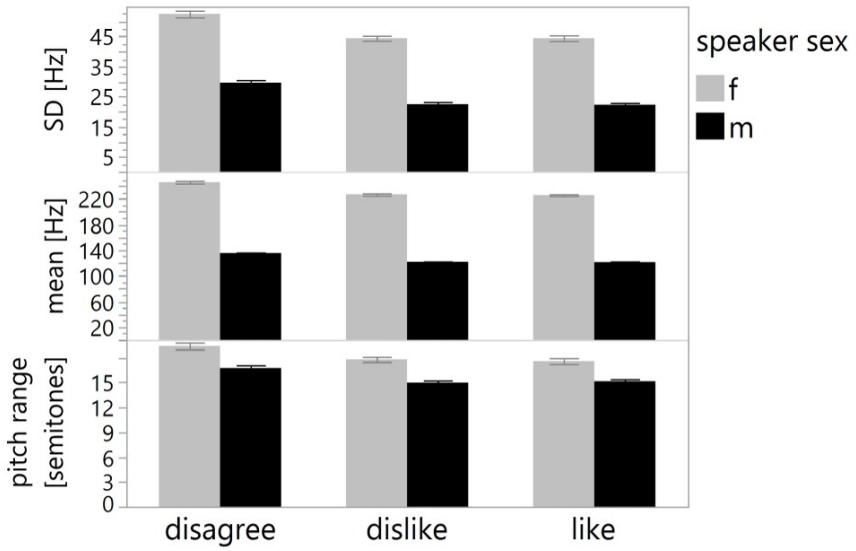

Figure 1 – Pitch range, F0 mean and SD for female and male speakers in the three modalities.

female speakers but in the higher register (see Figure 1).

3.1.2 Pitch profiles depending on dialogue participants and their sex

The second analysis investigated whether student speakers would adapt their F0 values to their interlocutors' F0 values depending on whether the interlocutor is a student or a teacher. The dependent factors were again the LTD F0 measures. Speaker was a random factor. The model included Interlocutors (student-student, student-teacher) and Sex as fixed factors and the interaction between them. In this analysis, as expected, we found again that women had significantly higher F0 values for level, span and SD. A significant main effect of Interlocutors was found on all measures, with students having higher F0 values for level, span and SD when participating in a dialogue with a teacher: mean F0 ($F [1, 3842] = 265.8$, $p<0.001$), median F0 ($F [1, 3842] = 210.4$, $p<0.001$), SD ($F [1, 3842] = 112.5$, $p<0.001$), IQR ($F [1, 3842] = 85.9$, $p<0.001$), span ($F [1, 3842] = 26.9$, $p<0.001$). Figure 2 displays the pitch range, mean F0 values and SD for male and female speakers in the two conditions. Significant interactions between Interlocutors and Sex were found for the mean ($F [1, 3842] = 13.2$, $p<0.001$) and median ($F [1, 3842] = 15.5$, $p<0.001$) F0 values. This interaction can be explained by the higher F0 register used by the female speakers.

4 Convergence and synchrony

We looked for global convergence using a paired t-test to compare the difference between the mean F0 values of speaker A and speaker B extracted for the first 30% and the last 30% of each conversation. We inferred global convergence when the differences in the last 30% were smaller. The results are given in Table 3. There were no significant tendencies towards global convergence in the dialogues between students. Regarding the dialogues between the teacher and the students, only the dialogue scenario in which the opinions of the two interlocutors differed showed significant convergence towards its end.

Local convergence was computed as the Pearson's correlation coefficient between time and the absolute difference between each target and its corresponding partner IPS: the more negative the correlation, the stronger the convergence. As shown in Table 3, we observed local convergence for student-teacher pairs in conversations where the two participants did not accept the provocative art or where the opinions of the two interlocutors differed. Again, student-

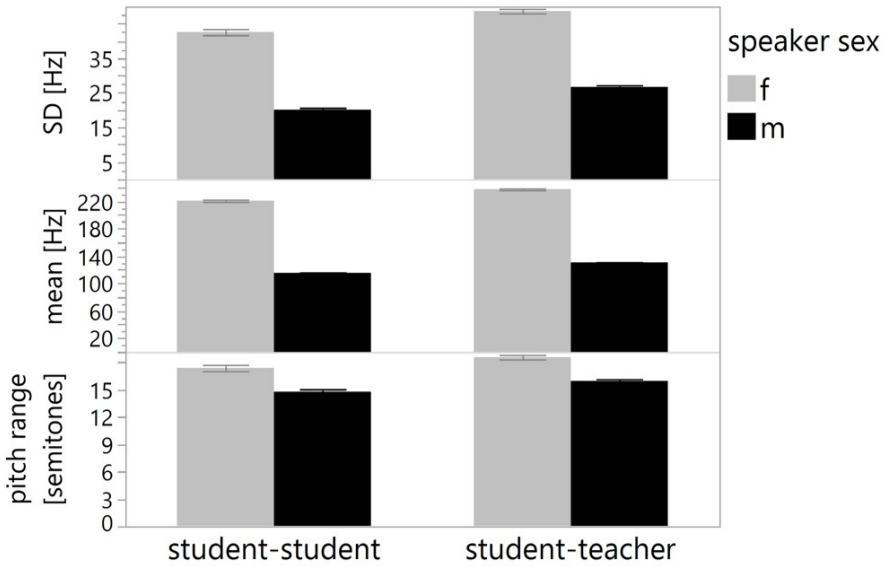

Figure 2 – Pitch range, F0 mean and SD for female and male speakers in student-student and student-teacher dialogues.

student pairs did not exhibit convergence behavior. This is most likely due to the fact that during the first part of the recordings they also participated in other dialogues with their peers that were not used in this study [32], and convergence may have already occurred there.

Pearson's correlation coefficient between each target IPS and its corresponding partner IPS captures the local synchrony. Positive correlation implies that speakers have similar speech patterns over time. Unlike local and global convergence, there was evidence of synchrony for all scenarios.

In general, the correlation coefficients are low for local convergence and low to moderate (the highest being .39) for synchrony, indicating a lack of strong trends across pairs.

Table 3 – Results for global and local convergence and for synchrony. T-statistics are reported for global convergence and Pearson's r coefficients are reported for local convergence and synchrony (* p<.05, ** p<.01, *** p<.001).

interlocutors	**modality**	**global convergence t**	**local convergence r**	**synchrony r**
student-student	[+like; +agree]	n.s.	n.s.	.21*
	[-like; +agree]	n.s.	n.s.	.39***
student-teacher	[+like; +agree]	n.s.	n.s.	.24*
	[-like; +agree]	n.s.	-.11*	.35***
	[±like; -agree]	-4.4***	-.13**	.21***

5 Discussion

The present study was concerned with pitch profiles and similarity in Polish conversational interactions in different scenarios between interlocutors with the same or different degrees of social status and the same or different sex.

With respect to our first research question, whether speakers systematically produce different LTD pitch profiles in different conversations, our results show that in the student-student and student-teacher conversations where both interlocutors agree whether or not they accept the provocative art, the students show higher LTD F0 values for level, span and standard deviation

when interacting with a teacher than with a fellow student. Additionally, in the student-teacher conversations the students achieve significantly higher LTD F0 values when the two interlocutors do not share the same opinion about the provocative art.

Regarding our second research question, whether interlocutors become more similar in mean F0 in the course of the conversation, we found evidence of local and global convergence only in the student-teacher conversations (see also [33]). The lack of convergence between students could be explained by the fact that convergence may already have taken place in previous conversations between the students. The fact that we observe convergence in the disagreement condition is in line with the literature. On the one hand, research in the area of conversational analysis has established the general tendency that there is a strong preference for agreement (or to avoid disagreement) between interlocutors [34], [35]. On the other hand, speakers at the lower end of the hierarchy tend to converge to the hierarchically higher interlocutor [16]. We observe synchrony in both student-student and student-teacher conversations. However, the weak to moderate correlation coefficients for convergence and synchrony indicate the variation in, and the complexity of, the overall speech coordination process. Detailed prosodic analyses of additional features known to be affected at the level of individual phrases seems particularly important. This will be the next stage of our research.

6 Acknowledgements

This research was supported by the Polish National Science Centre, project no.: 2014/14/M/HS2/00631, "Automatic analysis of phonetic convergence in speech technology systems" and the Bulgarian National Science Fund, project no. KP-06-40/11/12.12.2019, "Prosodic aspects of Bulgarian in comparison with other languages with lexical stress". This research was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID: MO 597/6-2.

References

[1] GILES, H., N. COUPLAND, and J. COUPLAND: *Accommodation theory: Communication, context, and consequence*. In H. GILES, J. COUPLAND, and N. COUPLAND (eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pp. 1–68. CUP, 1991.

[2] PARDO, J. S.: *On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America*, 119(4), pp. 2382–2393, 2006. doi:10.1121/1.2178720.

[3] LEVITAN, R. and J. HIRSCHBERG: *Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions*. In *Interspeech*, pp. 3081–3084. Florence, 2011.

[4] PICKERING, M. J. and S. GARROD: *Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences*, 27(2), pp. 169–190, 2004. doi:10.1017/S0140525X04450055.

[5] PICKERING, M. J. and S. GARROD: *An integrated theory of language production and comprehension. Behavioral and Brain Sciences*, 36(4), pp. 329–347, 2013. doi:10.1017/s0140525x12001495.

[6] SHEPARD, C. A., H. GILES, and B. A. LE POIRE: *Communication accommodation theory*. In W. P. ROBINSON and H. GILES (eds.), *The New Handbook of Language and Social Psychology*, pp. 33–56. Wiley, 2001.

[7] BELL, L., J. GUSTAFSON, and M. HELDNER: *Prosodic adaptation in human-computer interaction*. In *International Congress of Phonetic Sciences (ICPhS)*, pp. 2453–2456. Barcelona, 2003.

[8] BRENNAN, S. and H. CLARK: *Conceptual pacts and lexical choice in conversation. Journal of experimental psychology. Learning, memory, and cognition*, 22 6, pp. 1482–93, 1996.

[9] DANESCU-NICULESCU-MIZIL, C., L. LEE, B. PANG, and J. M. KLEINBERG: *Echoes of power: language effects and power differences in social interaction*. In *WWW*, pp. 699–708. ACM, 2012.

[10] BRANIGAN, H., M. PICKERING, and A. CLELAND: *'syntactic co-ordination in dialogue*. *Cognition*, 75, pp. B33–25, 2000.

[11] REITTER, D., F. KELLER, and J. D. MOORE: *Computational modelling of structural priming in dialogue*. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 121–124. Association for Computational Linguistics, New York City, USA, 2006.

[12] NATALE, M.: *Convergence of mean vocal intensity in dyadic communication as a function of social desirability*. *Journal of Personality and Social Psychology*, 32(5), pp. 790–804, 1975.

[13] STREET, R. L.: *Speech convergence and speech evaluation in fact-finding interviews*. *Human Communication Research*, 11(2), pp. 139–169, 1984. doi:10.1111/j.1468-2958.1984.tb00043.x.

[14] WLODARCZAK, M., J. SIMKO, and P. WAGNER: *Pitch and duration as a basis for entrainment of overlapped speech onsets*. In *INTERSPEECH*, pp. 535–538. ISCA, 2013.

[15] BAILLY, G. and A. MARTIN: *Assessing objective characterizations of phonetic convergence*. In *Interspeech*, pp. 2011–2015. Singapore, 2014.

[16] GREGORY, S. W. and S. WEBSTER: *A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions*. *Journal of personality and social psychology*, 70(6), pp. 1231–1240, 1996. doi:10.1037/0022-3514.70.6.1231.

[17] LEVITAN, R., A. GRAVANO, L. WILLSON, S. BENUS, J. HIRSCHBERG, and A. NENKOVA: *Acoustic-prosodic entrainment and social behavior*. In *NAACL Conference on Human Language Technologies*, pp. 11–19. 2012.

[18] BABEL, M., G. MCGUIRE, S. WALTERS, and A. NICHOLLS: *Novelty and social preference in phonetic accommodation*. *Laboratory Phonology*, 5(1), pp. 123–150, 2014. doi:10.1515/lp-2014-0006.

[19] SCHWEITZER, A. and N. LEWANDOWSKI: *Social factors in convergence of F1 and F2 in spontaneous speech*. In *International Seminar on Speech Production*. Cologne, 2014.

[20] MICHALSKY, J. and H. SCHOORMANN: *Pitch convergence as an effect of perceived attractiveness and likability*. In *Interspeech*, pp. 2253–2256. Stockholm, 2017. doi:10.21437/Interspeech.2017-1520.

[21] SCHWEITZER, K., M. WALSH, and A. SCHWEITZER: *To see or not to see: interlocutor visibility and likeability influence convergence in intonation*. In *Interspeech*, pp. 919–923. Stockholm, 2017. doi:10.21437/Interspeech.2017-1248.

[22] GESSINGER, I., E. RAVEH, I. STEINER, and B. MÖBIUS: *Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing*. *Speech Communication*, 127, pp. 43–63, 2021. doi:10.1016/j.specom.2020.12.004.

[23] SAVINO, M., L. LAPERTOSA, A. O. CAFFÒ, and M. REFICE: *Measuring prosodic entrainment in italian collaborative game-based dialogues*. In *SPECOM*, vol. 9811 of *Lecture Notes in Computer Science*, pp. 476–483. Springer, 2016.

[24] DE LOOZE, C., S. SCHERER, B. VAUGHAN, and N. CAMPBELL: *Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction*. *Speech Communication*, 58, pp. 11 – 34, 2014. doi:https://doi.org/10.1016/j.specom.2013.10.002.

[25] KARPINSKI, M., K. KLESSA, and A. CZOSKA: *Local and global convergence in the temporal domain in Polish task-oriented dialogue*. In *Proc. 7th International Conference on Speech Prosody 2014*, pp. 743–747. 2014. doi:10.21437/SpeechProsody.2014-136.

[26] BENUS, S., A. GRAVANO, R. LEVITAN, S. I. LEVITAN, L. WILLSON, and J. HIRSCHBERG: *Entrainment, dominance and alliance in supreme court hearings*. *Knowl. Based Syst.*, 71, pp. 3–14, 2014.

[27] EDLUND, J., M. HELDNER, and J. HIRSCHBERG: *Pause and gap length in face-to-face interaction*. In *Interspeech*, pp. 2779–2782. ISCA, 2009.

[28] BACHAN, J., M. OWSIANNY, and G. DEMENKO: *Creation of a dialogue corpus for automatic analysis of phonetic convergence*. In Z. VETULANI and P. PATRICK PAROUBEK (eds.), *Proceeding of 8th Language & Technology Conference, 17-19 November 2017, Poznań, Poland*, pp. 246–250. 2017.

[29] TALKIN, D.: *A robust algorithm for pitch tracking (RAPT)*. In *Speech Coding and Synthesis*, pp. 497 – 518. Elsevier, New York, 1995.

[30] LEVITAN, S. I., J. XIANG, and J. HIRSCHBERG: *Acoustic-prosodic and lexical entrainment in deceptive dialogue*. In *Proc. 9th International Conference on Speech Prosody 2018*, pp. 532–536. 2018. doi:10.21437/SpeechProsody.2018-108.

[31] JMP: *Version <13>. SAS Institute Inc., Cary, NC*. 1989-2020.

[32] DEMENKO, G. and J. BACHAN: *Annotation specifications of a dialogue corpus for modelling phonetic convergence in technical systems.* 2017.

[33] DEMENKO, G. (ed.): *Phonetic Convergence in Spoken Dialogues in View of Speech Technology Applications*. Akademicka Oficyna Wydawnicza EXIT, 2020.

[34] POMERANTZ, A.: *Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes*, p. 57–101. Studies in Emotion and Social Interaction. Cambridge University Press, 1985. doi:10.1017/CBO9780511665868.008.

[35] SACKS, H.: *On the preferences for agreement and contiguity in sequences in conversation*. In J. R. L. GRAHAM BUTTON (ed.), *Talk and Social Organisation*, chap. 2, pp. 54–69. Multilingual Matters, Clevedon, 1987.