PROSODIC ALIGNMENT IN DIFFERENT CONVERSATIONAL FEEDBACK FUNCTIONS

Carol Figueroa[1,2], Štefan Beňuš[3,4], Gabriel Skantze[1,5]

[1]Furhat Robotics;[2]Aix-Marseille University;[3]Institute of Informatics, Slovak Academy of Sciences;[4]Constantine the Philosopher University in Nitra;[5]KTH Royal Institute of Technology
carol@furhatrobotics.com, sbenus@ukf.sk, skantze@kth.se

ABSTRACT

Prosodic alignment is a phenomenon where the interlocutors' speaking style converge along various (para-)linguistic dimensions. Previous work has suggested that in terms of pitch, local alignment exists between backchannels and the preceding utterance of the interlocutor. In this paper, we propose a new operationalization of local prosodic alignment and investigate whether such alignment exists between short feedback utterances and the preceding, as well as the following, utterance of the interlocutor. Furthermore, we investigate whether this alignment differs between different feedback functions, including continuers, agreement, disagreement, yes/no responses, disapproval, non-understanding, sympathy, and mild/strong surprise. From the Switchboard corpus, we analyzed 2,118 instances of short feedback utterances with the mentioned 10 communicative feedback functions. Although we find significant results, they are much weaker than what has been reported in previous work.

Keywords: prosody, feedback, alignment, backchannels, entrainment

1. INTRODUCTION

The phenomenon where interlocutors become similar in their speaking styles is often referred to as alignment, entrainment or convergence [1]. Interlocutors may align their speaking styles among different dimensions, such as phonetic [2, 3, 4], acoustic-prosodic [5, 6, 7], lexical [8, 9, 10], and syntactic [11, 12]. Acoustic-prosodic alignment can be measured on a global or local level [7]. Global alignment refers to changes that take place over a longer time period (e.g., if speakers gradually converge towards a similar pitch level), whereas local alignment refers to local changes taking place in the vicinity of turn-exchanges. An example of the latter would be similarities of features between the beginning of an utterance and the ending of the preceding utterance of the interlocutor.

Previous work has found that the pitch similarity between backchannels (e.g., *yeah*, *mhm*) and the interlocutor's preceding utterance is greater than the pitch similarity of other turn-shifts [13]. They posit that it is due to this similarity that backchannels are perceived as unobtrusive in conversation. A similar analysis was done by [14], who concluded that pitch similarity between backchannels and the interlocutor's preceding utterance can be seen as an indication of entrainment or alignment between the speakers. However, it is not clear that this is a valid conclusion from this form of analysis. Heldner et al. [13] also found the pitch of backchannels to be higher in general, and it is well known that a rising pitch can serve as a backchannel-inviting cue [15], which could also explain the finding. Thus, we do not know whether the speakers truly align, in the sense that the listener changes the pitch level of the backchannel to 'tune-in' to the pitch level of the preceding utterance.

In this paper, we investigate local prosodic alignment between short feedback utterances and the preceding, as well as the following, utterance of the interlocutor. Following [16], we use a more strict operationalization of local alignment using Pearson's correlation, where we expect correlation between the prosodic features of the two interlocutors. The prosodic features we examine are pitch, pitch slope, and intensity. We also investigate whether local prosodic alignment differs for 10 different feedback communicative functions, namely continuers, agreement, disagreement, yes/no responses, disapproval, non-understanding, sympathy, and mild/strong surprise [17].

The results of our analyses can inform researchers who are interested in implementing feedback in spoken dialogue systems. There has been interest in implementing human alignment behavior in dialogue systems [18, 19]. If there is alignment for certain feedback functions in human-human interactions, this alignment can be implemented in spoken dialogue systems and can contribute to the

systems generating feedback in a more human-like manner. Our results can also be useful for better recognition of the user's feedback function by the system depending on whether the user aligns their prosodic features in their feedback to the system.

2. METHOD

2.1. Corpus and feedback functions

The short feedback utterances were extracted from Switchboard [20], which is a corpus consisting of about 2,500 telephone calls between 500 native speakers of American English. The dyadic pairs did not know each other and spoke between 3-10 minutes. Although the speakers were given a topic to discuss, they sometimes changed the topic of the conversation. The telephone conversations were recorded in two separate channels. The recordings were also transcribed and word level time-alignments were provided.

Feedback Function	Count
(C) Continue	1004
(U) Non-Understanding	61
(A) Agree	413
(D) Disagree	43
(Y) Yes response	57
(N) No response	107
(S) Sympathy	83
(MS) Mild Surprise	99
(SS) Strong Surprise	186
(Ds) Disapproval	65

Table 1: Feedback communicative functions and count used in analysis.

In order to annotate short feedback utterances, we first searched for lexical tokens that could potentially be instances of feedback; we used the list of lexical tokens provided in [17]. We then listened to these lexical tokens in context and assigned them one of the 10 feedback communicative functions presented in [17]. In total, we annotated 2,118 instances of feedback. The breakdown of the counts per feedback function are listed in Table 1.

2.2. Extraction of prosodic features

We used Parselmouth [21] to extract pitch (F0 Hz) and intensity (dB) every 10 ms from the speech signal. F0 values were first transformed to log scale and then z-score normalized. Intensity values were also z-score normalized. We z-score normalized the features by speaker using each speaker's mean and standard deviation computed from their entire conversation.

In order to obtain features for the preceding 500 ms of the interlocutor, we first extracted features from a window of 2000 ms preceding the start of the feedback. From this list of features we searched for the last voiced frame, then from the last voiced frame we took the preceding 49 frames. Concretely, this means that the preceding 500 ms of the interlocutor consists of the last voiced frame and the preceding 49 frames, 50 frames in total. For the following 500 ms of the interlocutor, we first extracted features from a window of 2000 ms following the end of the feedback. From this list of features we searched for the first voiced frame, then from the first voiced frame we took the following 49 frames, in total 50 frames.

2.3. Measuring local prosodic alignment

We measure local alignment (i.e. local synchrony [7]), as the Pearson's correlation coefficient of prosodic features between the preceding/following utterance of the interlocutor and the short feedback utterance. We decided to use Pearson's correlation to measure local alignment since it captures the following patterns:

- Positive correlation: prosodic features move up or down together, positive alignment
- Negative correlation: prosodic features move in opposite direction, negative alignment

We will interpret the correlations in the following way:

- Correlation between the listener's feedback and the preceding utterance of the interlocutor indicates that the listener aligns to the interlocutor.
- Correlation between the listener's feedback and the following utterance of the interlocutor indicates that the interlocutor aligns to the listener's feedback.
- Correlation between the preceding and the following utterance of the interlocutor does not indicate any alignment between the interlocutors, but the feedback can be seen as less obtrusive.

3. RESULTS AND DISCUSSION

In this section we report the Pearson correlation coefficient of the prosodic features between different feedback functions and the 500 ms preceding and following utterance of the interlocutor. Since we make 30 comparisons for each prosodic feature (10 functions * 3 correlations), a Bonferroni correction

should be applied to avoid Type I errors. In this case, correlations with $p < .0016$ ($\alpha = .05$) can be considered significant. However, since Bonferroni is a very conservative correction, we also report correlation coefficients which could be considered marginally significant ($p < .05$ and $p < .01$) and which could be interesting to explore in future work. We denote the strengths of correlation coefficients as: strong correlation ($r > .5$), moderate correlation ($.3 < r < .5$), and weak correlation ($0 < r < .3$).

We also conducted the analyses by excluding feedback that contained overlapping speech. However, this did not change the observed patterns in the results. Therefore, we report the results using the original 2,118 annotations of short feedback tokens.

3.1. Mean pitch

Figure 1 shows the significant results of the Pearson's correlations of mean pitch between the feedback functions and the preceding, and following 500 ms utterance of the interlocutor.

3.1.1. Feedback vs. preceding utterance

Using Bonferroni correction, we do not find any significant results indicating that the listener aligns their mean pitch of their feedback to the mean pitch of the preceding utterance of the interlocutor. Feedback function (C) *continue* has the same communicative function as backchannels. Heldner et al. [13] reported that backchannels have more similar pitch to the preceding utterance of the interlocutor than other turn-shifts. Similarly, [14] reported that backchannels have more similar pitch to the preceding utterance of the interlocutor than smooth switches, and that this similarity can be seen as an indication of entrainment or alignment. The differences in results could be either due to different methodologies or to the type of corpus. Both [13] and [14] used the same task-oriented corpus where preceding utterances of the interlocutor had a rising pitch [15] and the pitch of backchannels were also found to have a higher pitch in general [13]. It would be interesting to see future work adopting our methodology, to see if the listener 'tunes-in' to the pitch level of the interlocutor in task-oriented conversations.

3.1.2. Feedback vs. following utterance

We only find marginally significant results for feedback functions (A) *agree*, (S) *sympathy*, and (N) *no response*. Feedback functions (A), and (N) have

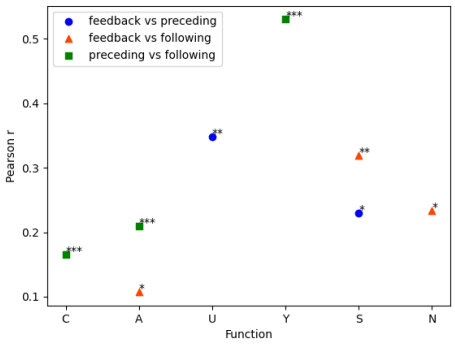

Figure 1: Pearson's correlation coefficients between the mean pitch of feedback functions and preceding/following 500 ms of the interlocutor. $*p < .05$, $**p < .01$, $***p < .001$

positive weak correlations, while (S) is moderately correlated. In these cases we see evidence that the interlocutor may align to the listener.

3.1.3. Preceding utterance vs following utterance

We find significant evidence that the mean pitch of functions (C) *continue*, (A) *agree*, and (Y) *yes response* can be seen as less obtrusive to the conversation. There is a weak positive correlation between the mean pitch of the preceding and following utterances of feedback function (C) ($r = 0.165$, $p < .001$). For the preceding and following utterance of feedback function (A) there is significant positive correlation ($r = 0.209$, $p < .001$). For the preceding and following utterance of feedback function (Y) there is significant and strong positive correlation ($r = 0.53$, $p < .001$). In the case of feedback function (Y), after the listener gives a *yes response*, the interlocutor may continue with the same pitch in a follow-up question or they continue speaking in the case where the listener responded to a rhetorical question.

3.2. Mean intensity

Figure 2 shows the significant results of the Pearson's correlation of mean intensity between the feedback functions and the preceding, and following 500 ms utterance of the interlocutor.

3.2.1. Feedback vs. preceding utterance

We find significant evidence that the listener may align their intensity to that of the interlocutor's for feedback functions (C) *continue* and (A) *agree*. Although we find significant positive correlation between the mean intensity of feedback function (C) and the preceding utterance of the interlocutor, the

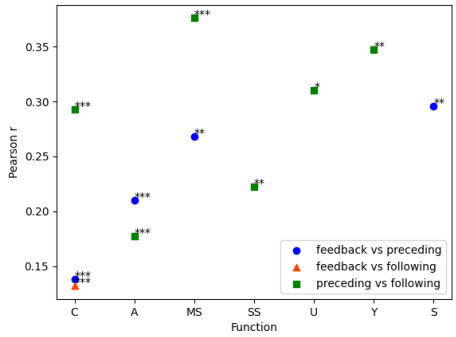

Figure 2: Pearson's correlation coefficients between the mean intensity of feedback functions and preceding/following 500 ms of the interlocutor. $*p < .05$, $**p < .01$, $***p < .001$

correlation coefficient is fairly weak ($r = 0.138$, $p < .001$). The correlation coefficient between feedback function (A) and the preceding utterance of the interlocutor is also significant but weak ($r = 0.21$, $p < .001$). For both function (C) and (A), we can speculate that the listener does not want to take the floor and will therefore 'tune-in' to the intensity of the interlocutor.

3.2.2. Feedback vs. following utterance

We only find evidence that the interlocutor may align their intensity to the listener's feedback function (C) *continue*. However, the positive correlation is weak ($r = 0.132$, $p < .001$). For feedback function (C) we now see the following alignment pattern: interlocutor A says something, listener B then produces a feedback function (C) which aligns its intensity to interlocutor A, then interlocutor A continues speaking and aligns its intensity to the preceding feedback function (C) of listener B.

3.2.3. Preceding utterance vs. following utterance

We find significant evidence that the intensity of the feedback functions (C) *continue*, (A) *agree*, and (MS) *mild surprise* may be perceived as less obtrusive to the conversation. The correlation coefficient between the preceding and following utterance for function (A) is positive and weak ($r = 0.177$, $p < .001$). The correlation between the preceding and following utterance of (C) is also weak, however the results are close to showing moderate correlation ($r = 0.293$, $p < .001$). The preceding utterance and following utterance of feedback function (MS) *mild surprise* are positive and moderately correlated ($r = 0.376$, $p < .001$). It is not surprising that these feedback functions are perceived as less obtrusive since the listener does not try to take the floor.

3.3. Pitch slope

Unlike mean pitch and intensity, we did not find any significant or marginally significant results across all the feedback function categories. This could be due to the way pitch slope was calculated. For example, in [15] they were computed by fitting least-squares linear regression models to the f0 values.

3.4. Limitations and future work

Our analysis does not account for temporal information that global alignment captures; there may be differences with how feedback functions align in the beginning of the conversation compared to the end of the conversation. Future work should also explore the difference in task-based conversations compared to spontaneous conversations and how this affects alignment. Different normalization methods for speech features might also affect the results of alignment. Levitan et al. [14] showed that latency can negatively affect local entrainment. Future work should investigate the difference between the end time of the preceding utterance of the interlocutor and the start of the feedback, as well as the difference between the end time of the feedback and the start time of the following utterance for all feedback functions.

4. CONCLUSION

We find that listeners align their intensity for feedback functions (C) *continue* and (A) *agree* to the intensity of the preceding utterance of the interlocutor. In terms of intensity for feedback function (C) we also find a pattern of alignment: when interlocutor A says something, listener B aligns the intensity of their feedback to that of interlocutor A, then interlocutor A continues speaking and aligns their intensity to the feedback of listener B. This can be useful for spoken dialogue designers who would like to implement human-like feedback. Our results also show that in terms of pitch and intensity feedback functions (C) *continue*, (A) *agree* can be perceived as less obtrusive in conversation.

5. ACKNOWLEDGEMENTS

This work was funded by the European Union's Horizon 2020 research and innovation program

under the Marie Skłodowska Curie grant agreement No 859588 and in part by the Slovak Granting Agency grant VEGA 2/0165/21, Slovak Research and Development Agency grant APVV-21-0373.

6. REFERENCES

[1] C. J. Wynn and S. A. Borrie, "Classifying conversational entrainment of speech behavior: An expanded framework and review," *Journal of Phonetics*, vol. 94, p. 101173, 2022.

[2] T. Biro, J. C. Toscano, and N. Viswanathan, "The influence of task engagement on phonetic convergence," *Speech Communication*, pp. 50–66, 2022.

[3] A. J. Olmstead, N. Viswanathan, T. Cowan, and K. Yang, "Phonetic adaptation in interlocutors with mismatched language backgrounds: A case for a phonetic synergy account," *Journal of Phonetics*, vol. 87, p. 101054, 2021.

[4] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[5] S. I. Levitan, J. Xiang, and J. Hirschberg, "Acoustic-Prosodic and Lexical Entrainment in Deceptive Dialogue," in *Proc. Speech Prosody 2018*, 2018, pp. 532–536.

[6] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 325–334.

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

[8] J. Kruyt, Š. Beňuš, C. Faget, C. Lançon, and M. Champagne-Lavau, "Prosodic and lexical entrainment in adults with and without schizophrenia," in *Proc. Speech Prosody 2022*, 2022, pp. 125–129.

[9] A. Weise, V. Silber-Varod, A. Lerner, J. Hirschberg, and R. Levitan, "Talk to me with left, right, and angles": Lexical entrainment in spoken Hebrew dialogue," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 2021, pp. 292–299.

[10] H. Friedberg, D. Litman, and S. B. F. Paletz, "Lexical entrainment and success in student engineering groups," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 404–409.

[11] M. L. Allen, S. Haywood, G. Rajendran, and H. Branigan, "Evidence for syntactic alignment in children with autism," *Developmental Science*, vol. 14, no. 3, pp. 540–548, 2011.

[12] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic co-ordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13–B25, 2000.

[13] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Proc. Interspeech 2010*, 2010, pp. 3054–3057.

[14] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *2015 AAAI Spring Symposium Series*, 2015.

[15] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Proc. Interspeech 2009*, 2009, pp. 1019–1022.

[16] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, 2011, pp. 113–117.

[17] C. Figueroa, A. Adigwe, M. Ochs, and G. Skantze, "Annotation of communicative functions of short feedback tokens in switchboard," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1849–1859.

[18] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An end-to-end conversational style matching agent," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 111–118.

[19] R. Levitan, Š. Beňuš, R. H. Galvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," in *Proc. Interspeech 2016*, 2016, pp. 1166–1170.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[21] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.