Are we in sync during turn switch?

Jieyeon Woo, Liu Yang, Catherine Achard, Catherine Pelachaud

Are we in sync during turn switch?

Jieyeon Woo*, Liu Yang*, Catherine Achard and Catherine Pelachaud
Institut des Systèmes Intelligents et de Robotique (CNRS-ISIR), Sorbonne University, Paris, France

***Abstract*— During an interaction, people exchange speaking turns by coordinating with their partners. Exchanges can be done smoothly, with pauses between turns or through interruptions. Previous studies have analyzed various modalities to investigate turn shifts and their types (smooth turn exchange, overlap, and interruption). Modality analyses were also done to study the interpersonal synchronization which is observed throughout the whole interaction. Likewise, we intend to analyze different modalities to find a relationship between the different turn switch types and interpersonal synchrony. In this study, we provide an analysis of multimodal features, focusing on prosodic features (F0 and loudness), head activity, and facial action units, to characterize different switch types.**

I. INTRODUCTION

Information is communicated through verbal and nonverbal channels in the conversation. Nonverbal behavior refers to "body language" including gestures, facial expressions, body movement, and gaze [8], and form intra-synergies with one's own behavior [14].

While conversing face-to-face, people constantly coordinate and adapt their behavior based on the social signals emitted by their interlocutors [14] to increase the fluidity of the exchange and the engagement level [21]. Interlocutors' behavior coordination is also referred to as interpersonal synchrony which we define as the temporal coordination of interlocutors' behavior signals as in [19]. Synchrony may occur unintentionally during an interaction [37], where interlocutors manage their turn, quickly and frequently exchange roles between being a listener or a speaker. In most cases, speaking turns are exchanged smoothly with no gap or no overlap when the listener and the speaker are ideally following the rules defined in [34]. On the contrary, turns can be forcefully changed by an interruption or through silence.

According to Beattie [3] and Schegloff and Sacks [36], based on simultaneous speech types and willingness to yield the floor, turn switch can be basically classified into three main categories: smooth switch, interruption, and overlap. Overlap is defined as the listener over-anticipating the end of the current speaker [34], resulting in an overlay between the last word or syllable of the current speaker and the first word of the listener [34]. Interruption is when the listener grabs the floor against the speaker's will and is described as a violation of the current speaker's, unlike overlap [29], [36].

For our study, we merge overlap and smooth switch as smooth turn exchange since they are at the end of a turn, while interruption happens usually before the utterance is completed. When trying to detect an interruption in real-time, it can be mistakenly identified as a backchannel due to the fact that both interruption and backchannel, which refers to feedback messages not aimed at taking the floor, may also occur before the completion of an utterance. We aim to use this study as a stepping stone for multiple applications including real-time switch identification, which leads us to also analyze backchannels. Our main interest of this study is to see if a relationship between interpersonal synchrony and switch types (smooth turn exchange and interruption) along with backchannel could be found. We specifically look into dyadic interaction.

In this paper, works related to switch and synchrony will be presented in Section II. Our hypotheses will be introduced in Section III. In Section IV, details of the analyzed corpus and the studied features will be explained. The analysis will be shared in Section V before concluding the paper.

II. RELATED WORKS

Many scholars have been interested in the study of interaction for a long time. Emanuel A. Schegloff [35] defined natural conversations sequencing rules. Interacting partners dynamically collaborate during the course of a conversation to exchange the speaking floor based on rules and maintain the communication [15], [9]. Harvey Sacks [34] thus proposed the idea of conversation analysis, and described its most basic structure as turn-taking.

To understand the coordination during turn switch, multimodal features, such as eye-gaze [17], [25], respiration [24], [26], and head-direction [39] have been analyzed. Moreover, many works analyzed and mentioned the importance of prosodic features variation during switches [22], [27], [40], and found that interruptions are often combined with higher voice energy [38], [23]. The rise of voice from interrupter mat then causes a reduction of interpersonal synchrony, which might not be the case for backchannel and smooth turn exchange. We thus are interested in the relationship between interpersonal synchrony and different switch types.

Various works have studied interpersonal synchrony to check whether the partners are in sync, and numerous methods were proposed and employed.

Pioneer works were based on manual assessment. Observers were trained to perceive synchrony directly in the data on a local scale using behavior coding methods [12], [16] or on a larger time scale via judgment methods [12], [6]. The laborsome workload of manual annotations was relieved by automatic measures which detect synchrony by capturing relevant signals. Among different measures, correlation is the most commonly used one for interpersonal synchrony [11], [18], [32]. Coordinated behaviors are produced as a reaction to the interlocutors' social signals. The perception time

*Both authors contributed equally to this research.

979-8-3503-4544-5/23/$31.00 ©2023 IEEE

results in the need for consideration of a certain time delay (2 to 4 seconds [13], [28]). Taking this time delay into account, several works use the time-lagged cross-correlation [7], [1], [4]. The behavior signals are not only shifted in time but also varied in length. To address such problems, Dynamic Time Warping (DTW) [30], which accounts for time delay and duration variations,has been widely used to find common patterns [5]. Apart from temporal measures, some studies also perform spectral analysis. Information concerning synchrony stability is obtained by measuring the evolution of relative phase [31], [33].

For our work, to study the synchrony during a turn switch, we choose to employ frequently used synchrony measures of correlation (Pearson correlation coefficient), time-lagged cross-correlation, and DTW.

Previous works analyzed interlocutors' behavior synchrony during the whole interaction course, but not especially for turn switches. In addition, studies on conversation analysis did not yet consider interpersonal synchrony which can be a potential feature for switch type characterization. We want to analyze to find if there is a link between synchrony and turn switch using their different characteristics.

III. Hypotheses

Our study focuses on finding a link between switch types and behavior synchrony. Prior to analyzing multimodal signals, we posit the following hypotheses:

- **H1:** Switch types can be distinguished by their synchrony scores (i.e. correlation value) during the switch.
- **H2:** Variation of synchrony scores help to identify different switch types.
- **H3:** Each switch type's synchrony scores evolve differently (before, during, and after the switch).

IV. Corpus

We use the french part of NoXi database [10], containing 21 dyadic interactions of natural conversations, with a total duration of 7h22. Annotated turn switch and backchannel moments (1403 smooth turn exchanges, 1651 backchannels, and 929 interruptions), which followed the annotation schema described in [41], were used for our study.

Each turn switch and backchannel moment (onset point of the listener's voice activity) was noted as t_0. Each moment is segmented into three phases: before ($t_0 - 6\text{s} \sim t_0 - 2\text{s}$), during ($t_0 - 2\text{s} \sim t_0 + 2\text{s}$), and after ($t_0 + 2\text{s} \sim t_0 + 6\text{s}$).

Multimodal features that correspond to the switch and backchannel segments were extracted: facial features via OpenFace [2], acoustic features via openSMILE [20].

For our analysis, the following features were used:

- **Facial features:** AU1, AU2, AU4, AU12, and AU15.
- **Head features:** Head translation and rotation
- **Acoustic features:** F0 and Loudness.

To avoid the bias caused by the interactant's initial position, we use head motion activity with the following equation instead of the absolute position, where x_i, y_i and z_i are the coordinates of the head at time-step i.:

$$v_{Head}(i) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2} \quad (1)$$

All features are normalized using z-scores to be invariant to the quantity of behaviors of interactants.

V. Analysis

With our goal to identify smooth turn exchange and interruption, modality signals were analyzed. For the future provisional extension of our study for real-time switch identification, we also study the synchrony of backchannels. We focus mainly on the features that show significant information on synchrony scores.

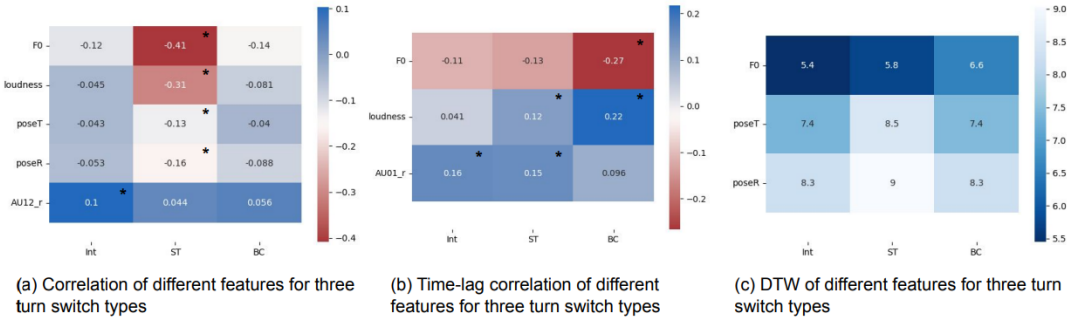
A two-tailed t-test was performed to check the significance of the difference between all pairs of the turn switch types.

We first analyzed the interpersonal synchrony scores during smooth turn exchange, interruption, and backchannel.

In Figure 1, we remark significant differences in synchrony scores with t-test ($p < 0.01$) for the analyzed signals of interruption (Int), smooth turn exchange (ST), and backchannel (BC). To detail, in Figure 1 (a) smooth turn exchange for the acoustic and head features and interruption for AU12 are significant. Also in Figure 1 (b), backchannel is significant for the acoustic features. Backchannel for F0 and smooth turn exchange for head features in Figure 1 (c) are significant.

From (a) in Figure 1, the two acoustic features, F0 and loudness, and head features, rotation and position activity, we can see that the two interlocutors are negatively correlated (behaviors with opposite trends). Via the correlation values of the two sets of features, smooth turn exchange can be differentiated from interruption and backchannel as their values are mostly uncorrelated (close to 0 meaning that the two interlocutors are not synchronized) or comparatively less correlated. With the AU12, a positive correlation is found between the partners during the interruption while no relation is seen for smooth turn exchange and backchannel. This allows interruptions to be identifiable among the others. Smooth turn exchange and interruption thus can be distinguished by the correlation measure. Then from the time-lag correlation presented in (b) of Figure 1, a positive correlation of loudness and AU1 is observed for all three of them. No correlation can be found for backchannel while the two others are positively correlated for AU1. For loudness, the synchrony increases in the order of interruption, smooth turn exchange, and backchannel. A negative correlation for F0 is observed for all, of which backchannel's correlation score is noticeably higher than the other two types. This allows backchannels to be identifiable among the others. In Figure 1 (c), as DTW measures the distance between two signals (the smaller the distance, the more they are synchronized), we can note a low value of synchrony during smooth turn exchange via the head features and also lower synchrony for backchannel with F0 compared to the other two.

As a result, we can validate our hypothesis H1 that synchrony measures, with significant value differences, can be used to identify switch types and backchannel.



(a) Correlation of different features for three turn switch types

(b) Time-lag correlation of different features for three turn switch types

(c) DTW of different features for three turn switch types

Fig. 1. Three synchrony measures of multimodal features for switch types (*: $p < 0.01$)

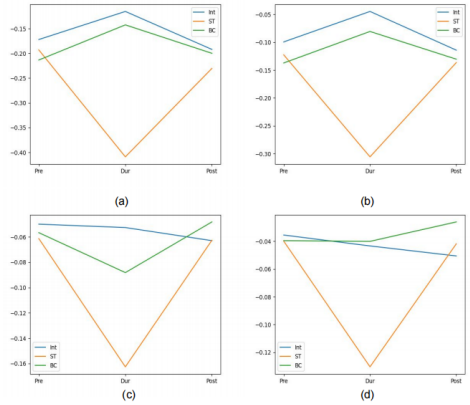


Fig. 2. Correlation of multimodal features during pre/dur/post period for switch types: (a)F0, (b)Loudness, (c)HeadRot, (d)HeadPose

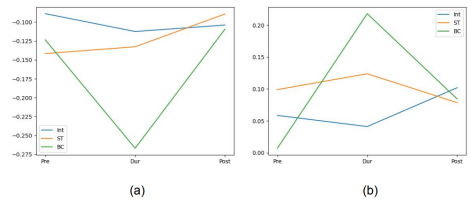


Fig. 3. Time-lag correlation of multimodal features during pre/dur/post period for switch types: (a)F0, (b)Loudness

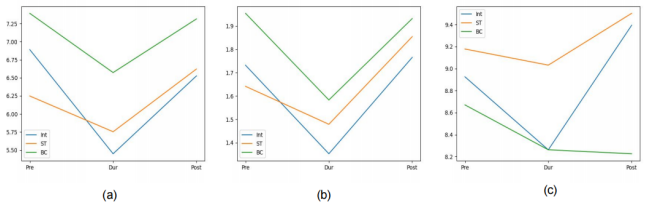


Fig. 4. DTW of multimodal features during pre/dur/post period for switch types: (a)F0, (b)Loudness, (c)HeadRot

Synchrony scores have shown their usefulness in differentiating smooth turn exchange, interruption, and backchannel, but it is still not enough to clearly identify all three them. To reinforce the detection, we look at the variation of synchrony scores before, during, and after the switches.

In Figure 2, using the correlation measure we can remark the sudden increase in negative correlation (signifying an increase in synchrony of signals evolving in the opposite directions) for acoustic and head features during smooth turn exchange which is significantly noticeable compared to the others which show a slight decrease or a stable trend in synchrony during their change. Same as in Figure 1 (b), we can see that backchannels can be noticed thanks to the time-lag correlation in Figure 3 with the variation of synchrony scores using acoustic features. A clear increase of synchrony value (or inverse correlation) can be observed during the transition for acoustic features while the others remain still or augment their synchrony value. The measure of DTW, in Figure 4, shows that the turn switch and backchannel trigger an increase in synchrony value for the acoustic features and head rotation. We also calculate the difference (delta) between the three phases to see the variation significance. The phase transition of before-during and during-after is significant with t-test value of $p < 0.01$ for smooth turn exchange in Figures 2-4, for backchannel in Figures 3 and 4 (a,b), and for interruption in Figures 3 and 4 (a-c).

This confirms our second hypothesis H2 that the variation of synchrony scores can provide further information in differentiating switch types and backchannel. In addition, we have observed that each of them show different scores of synchrony during the change and also possesses different variation trends which support our last hypothesis H3.

VI. CONCLUSIONS

Inspired by previous works of multimodal analysis for different types of turn switches and interpersonal synchrony researches, we came up with the idea to study the relationship between synchrony and turn switch. Analyzing the acoustic features as well as facial expression and head movement signals, we investigate the synchrony scores before, during, and after smooth turn exchange, interruption, and backchannel. Through our study, we were able to validate our hypotheses.

Our study shows that synchrony scores can be an indicator to distinguish smooth turn exchange, interruption, and backchannel. Via this relationship found between synchrony and turn switch and backchannel, as an extension we expect that synchrony measures could serve as an additional feature to predict their type and timing. We also hope to further investigate the long-term effect of the frequency of interruptions and/or backchannels on the overall interpersonal synchrony of the interaction.

VII. ACKNOWLEDGMENTS

This work is performed as a part of IA ANR-DFG-JST Panorama and ANR-JST-CREST TAPAS (19-JSTS-0001-01) projects.

REFERENCES

[1] K. T. Ashenfelter, S. M. Boker, J. R. Waddell, and N. Vitanov. Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 35(4):1072, 2009.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[3] G. W. Beattie. Interruption in conversational interaction, and its relation to the sex and status of the interactants. 1981.

[4] Š. Beňuš, A. Gravano, and J. Hirschberg. Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, 43(12):3001–3027, 2011.

[5] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

[6] F. J. Bernieri, J. S. Reznick, and R. Rosenthal. Synchrony, pseudosynchrony, and dissynchrony: measuring the entrainment process in mother-infant interactions. *Journal of personality and social psychology*, 54(2):243, 1988.

[7] S. M. Boker, J. L. Rotondo, M. Xu, and K. King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, 7(3):338, 2002.

[8] J. K. Burgoon, L. K. Guerrero, and V. Manusov. Nonverbal signals. *The SAGE handbook of interpersonal communication*, pages 239–280, 2011.

[9] J. K. Burgoon, L. A. Stern, and L. Dillman. *Interpersonal adaptation: Dyadic interaction patterns.* Cambridge University Press, 1995.

[10] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. Andre, and M. Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. pages 350–359, 11 2017.

[11] N. Campbell. Multimodal processing of discourse information; the effect of synchrony. In *2008 Second International Symposium on Universal Communication*, pages 12–15. IEEE, 2008.

[12] J. N. Cappella. Behavioral and judged coordination in adult informal social interactions: Vocal and kinesic indicators. *Journal of personality and social psychology*, 72(1):119, 1997.

[13] T. L. Chartrand and J. A. Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

[14] W. S. Condon and W. D. Ogston. Sound film analysis of normal and pathological behavior patterns. *Journal of nervous and mental disease*, 1966.

[15] W. S. Condon and W. D. Ogston. A segmentation of behavior. *Journal of psychiatric research*, 5(3):221–235, 1967.

[16] W. S. Condon and L. W. Sander. Neonate movement is synchronized with adult speech: Interactional participation and language acquisition. *Science*, 183(4120):99–101, 1974.

[17] I. De Kok and D. Heylen. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 91–98, 2009.

[18] E. Delaherche and M. Chetouani. Multimodal coordination: exploring relevant features and measures. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 47–52, 2010.

[19] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

[20] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[21] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[22] P. French and J. Local. Turn-competitive incomings. *Journal of Pragmatics*, 7(1):17–38, 1983.

[23] A. Gravano and J. Hirschberg. A corpus-based study of interruptions in spoken dialogue. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[24] M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[25] R. Ishii, K. Otsuka, S. Kumano, M. Matsuda, and J. Yamato. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 79–86, 2013.

[26] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato. Using respiration to predict who will speak next and when in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(2):1–20, 2016.

[27] E. Kurtić, G. J. Brown, and B. Wells. Resources for turn competition in overlapping talk. *Speech Communication*, 55(5):721–743, 2013.

[28] N. P. Leander, T. L. Chartrand, and J. A. Bargh. You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry. *Psychological science*, 23(7):772–779, 2012.

[29] M. Moerman and H. Sacks. Appendix b. on "understanding" in the analysis of natural conversation. In *Talking Culture*, pages 180–186. University of Pennsylvania Press, 2010.

[30] M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[31] O. Oullier, G. C. De Guzman, K. J. Jantzen, J. Lagarde, and J. Scott Kelso. Social coordination dynamics: Measuring human bonding. *Social neuroscience*, 3(2):178–192, 2008.

[32] D. Reidsma, A. Nijholt, W. Tschacher, and F. Ramseyer. Measuring multimodal synchrony for human-computer interaction. In *2010 international conference on cyberworlds*, pages 67–71. IEEE, 2010.

[33] M. J. Richardson, K. L. Marsh, R. W. Isenhower, J. R. Goodman, and R. C. Schmidt. Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human movement science*, 26(6):867–891, 2007.

[34] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

[35] E. A. Schegloff. Sequencing in conversational openings 1. *American anthropologist*, 70(6):1075–1095, 1968.

[36] E. A. Schegloff and H. Sacks. *Opening up closings*. Walter de Gruyter, Berlin/New York Berlin, New York, 1973.

[37] R. C. Schmidt and M. J. Richardson. Dynamics of interpersonal coordination. In *Coordination: Neural, behavioral and social dynamics*, pages 281–308. Springer, 2008.

[38] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[39] G. Skantze, M. Johansson, and J. Beskow. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 67–74, 2015.

[40] K. P. Truong. Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee. In *Interspeech*, pages 1404–1408, 2013.

[41] L. YANG, C. ACHARD, and C. PELACHAUD. Annotating interruption in dyadic human interaction. *Listener*, 600:600ms.