Inter-person Intra-modality Attention Based Model for Dyadic Interaction Engagement Prediction

Xiguang Li, Candy Olivia Mawalim, and Shogo Okada(✉)

Japan Advanced Institute of Science and Technology, Nomi, Japan
{s2020425,candylim,okada-s}@jaist.ac.jp

Abstract. With the rapid development of artificial agents, more researchers have explored the importance of user engagement level prediction. Real-time user engagement level prediction assists the agent in properly adjusting its policy for the interaction. However, the existing engagement modeling lacks the element of interpersonal synchrony, a temporal behavior alignment closely related to the engagement level. Part of this is because the synchrony phenomenon is complex and hard to delimit. With this background, we aim to develop a model suitable for temporal interpersonal features with the help of the modern data-driven machine learning method. Based on previous studies, we select multiple non-verbal modalities of dyadic interactions as predictive features and design a multi-stream attention model to capture the interpersonal temporal relationship of each modality. Furthermore, we experiment with two additional embedding schemas according to the synchrony definitions in psychology. Finally, we compare our model with a conventional structure that emphasizes the multimodal features within an individual. Our experiments showed the effectiveness of the intra-modal inter-person design in engagement prediction. However, the attempt to manipulate the embeddings failed to improve the performance. In the end, we discuss the experiment result and elaborate on the limitations of our work.

Keywords: Engagement Modeling · Interpersonal Synchrony · Attention Model

1 Introduction

Researchers have come to realize the importance of engagement prediction in the area of virtual communications and human-robot interaction. The engagement level has been a crucial factor in interaction diagrams. For example, an embodied conversational agent needs to adjust the interaction strategy based on the current engagement level of the subject. Many studies have been conducted based on various modalities via rule-based measurements or machine learning to predict engagement levels [21]. However, synchrony, a prosocial behavior phenomenon [3] which is closely related to high engagement, has not received enough attention in engagement modeling. Despite being a widely observed phenomenon, synchrony

A. Coman and S. Vasilache (Eds.): HCII 2023, LNCS 14025, pp. 91–105, 2023.
https://doi.org/10.1007/978-3-031-35915-6_8

is complex and challenging to define and delimit by rule-based methods. With the rapid development of data-driven deep learning, stunning progress has been made for numerous problems that are also hard to define and delimit with, such image classification tasks and content generation tasks. We believe that the modern deep learning model can capture the synchrony features and improve the prediction accuracy of engagement levels.

This paper introduces an intra-modality attention-based model for dyadic interaction real-time engagement level prediction. We first introduce the related concepts and the most influential works on engagement modeling. Then, we describe our model and evaluate its performance.

2 Related Works

2.1 Engagement and Synchrony

Nadine G. and Catherine P. have conducted an in-depth survey on the engagement for human-agent exchange [11]. There are many definitions in the literature focusing on various perspectives and targets. For example, Dan Bohus and Eric Horvitz describe engagement as "The process subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume or terminate an interaction" [4]. In contrast, Poggi regarded engagement as "The value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" [22]. Yu et al. [28] defined engagement in the voice communication system as "User engagement describes how much a participant is interested in and attentive to a conversation." Engagement can be observed multimodally from both verbal and non-verbal features. For example, engagement detection has been studied on prosodic features and emotions from speech [28], as well as facial expression, smile and gaze [10].

Synchrony is the temporal alignment among the participants during social interaction. Frank J. B and his colleagues defined synchrony as "The coordination of movement between individuals in social interactions" [2]. In loose terms, synchrony is similar to interpersonal coordination - "the degree to which the behaviors in an interaction are non-random, patterned, or synchronized in both timing and form" [3]. In later works, researchers viewed synchrony as a simultaneous synchronization of behaviors [15,23]. There are other similar concepts, such as the chameleon effects [6], co-occurrence [18], mimicry [7]. Synchrony has positively affected building rapport [25], smoothing social interactions [17], and promoting cooperation [27]. Emilie Delaherche et al. [9] have conducted an excellent survey on interpersonal synchrony for more insights.

2.2 Deep Learning Engagement Models

In the early stage of engagement prediction, researchers took a single image or frame from a video to predict the engagement level. For example, Omid Mohamad Nezami et al. [20] proposed a VGG-B [24] style deep neural network.

Their work was trained and evaluated on individual frames sampled from student study videos. Their model showed improved results over the histogram of oriented gradients and support vector machines [8]. Later, researchers included temporal information into consideration. For example, Hadfield et al. [12] studied child-robot attention tasks with long short-term memory (LSTM) models [14] and showed that temporal dynamics are crucial for engagement level prediction as LSTM models outperformed stationary classifiers.

Other approaches predicted engagement from modality features instead of video. In [1], a human-agent engagement prediction, employed a RESNet-18 [13] model as the backbone to extract attention signals from the gaze and head pose. Then they predicted the engagement level through rule-based policies on body postures and extracted attention signals. In [16], a student engagement prediction task over online lecture scenarios, the authors took face frames and facial landmarks for a fully connected neural network. In addition, they also used head pose and eye gaze features and fed them into a LSTM model. These works have studied the non-verbal features practical for engagement level prediction. Lastly, Soumia D. and Catherine P. [10] used gaze, head rotation, and facial action unit features and fed them into an LSTM model in a dyadic interaction engagement prediction task on NoXi dataset [5]. Their study consists of three models: target LSTM, partner LSTM, and dyadic LSTM. The result showed that additional information from the interaction partner boosted the prediction accuracy.

3 Methods

The task of our experiments is to predict the target participant's real-time engagement level from both interaction participants' modality features as input. We hypothesize that synchrony manifests as a form of feature similarity. The question becomes what kind of similarity and which time frame to compare the similarity. Different from refined features like the binary features of presence of smiling or other behaviors, measuring the similarity of sensor data such as face mesh is very challenging from the definition. We adopt a multi-stream deep neural network to let data speak for itself to extract similarity (Sect. 3.1 and 3.2). As for the time frame, we manually manipulate the embedding phrase of the network to control the feature grouping (Sect. 3.3) (Fig. 1).

3.1 Overall Structure

The overall model design follows multi-stream late-fusion scheme as shown in Fig. 2. The intuition is to allow the model to extract temporal synchrony features that reside within each modality between two participants. The multi-stream structure process the multi-modalities inputs in isolation. The intuition behind seeking temporal associations within the same modality is from the definitions of synchrony. Running in isolation avoids cross-modality learning that is not related

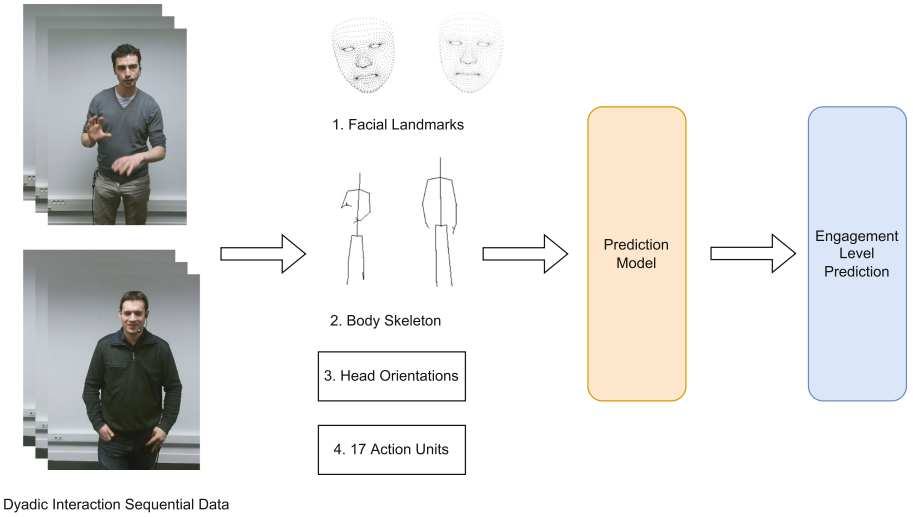

Fig. 1. Task Overview - We obtained facial landmarks, body skeletons, head orientations, and action units from the NoXi dataset as input features. The model will take four modalities sequence from present to past within a preset window length as input to predict the engagement level of the present sample.

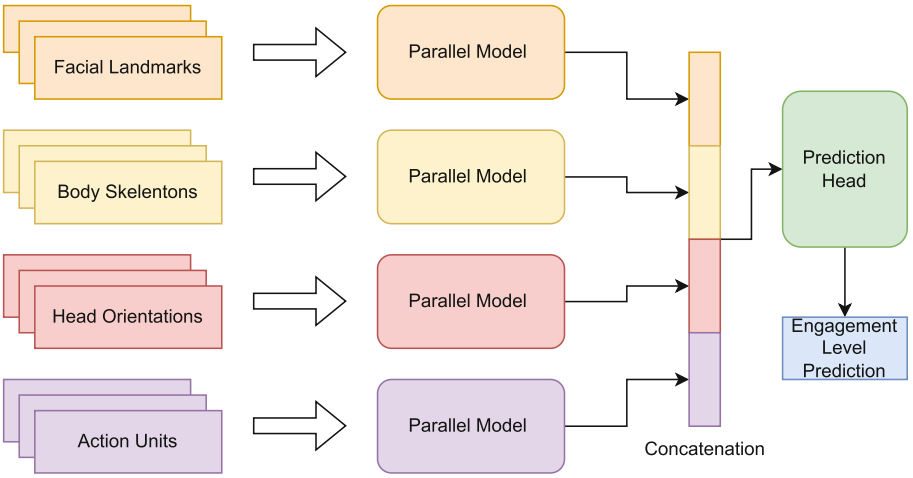

Fig. 2. Model Overview - Each modality sequence is fed into a separate parallel model, and the outputs are concatenated for the final prediction. All parallel models have identical structures (shown in Fig. 3) but with different layer dimensions adjusted for the input modality.

to synchrony. The main mechanism adopted for parallel models is the multi-head self-attention block. Attention models are incredibly flexible in learning temporal relationships. However, it also requires more training data to learn the attention matrix than models with built-in inductive bias. To tackle this issue, we designed two other embedding approaches to reduce the complexity of the attention matrix.

3.2 Multi-head Self Attention Backbone Model

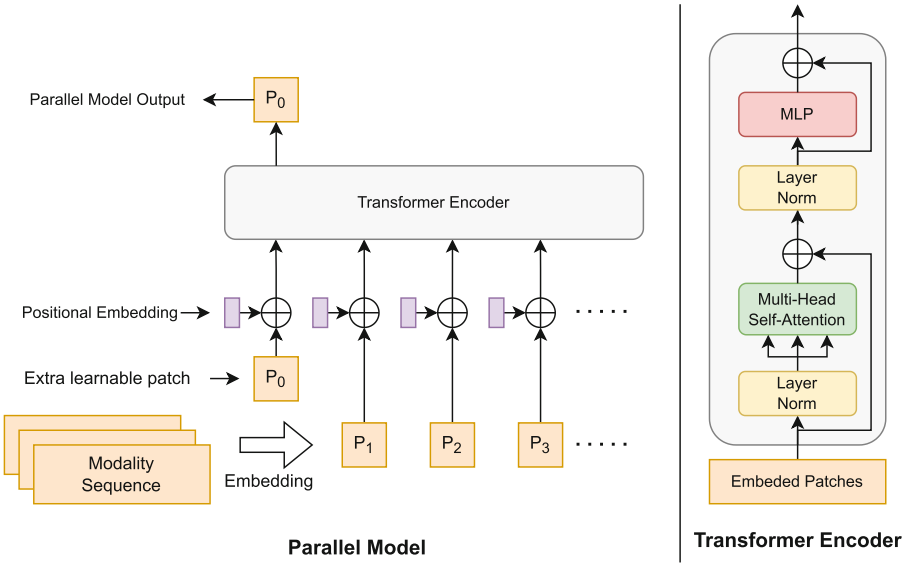

Fig. 3. Parallel Model Structure - The model takes the modality sequence as input, embeds them into patches (Embedding detail in Fig. 4), and processes the patches via a standard transformer block. A typical learnable "class token" patch is added to the sequence, serving as the parallel model fusion output.

Inspired by the fantastic work of ViT [19], we considered modality information as a series of "words" embedded and processed them with a transformer [26] encoder. Figure 2 illustrates the general structure of our parallel models. Modality sequences are first embedded into patches, applied positional embeddings, and processed by attention blocks. The flexibility of the attention mechanism comes from its learnable attention matrix. Embedded patches are first projected into matrices Q, K, and V with the exact dimensions P by L, where P the number of patches and L is the length of the embedding (Eq. 1). Then, the attention matrix is the matrix multiplication of Q and K with softmax and value scale (Eq. 2).

$$Q, K, V = linear(patches) \tag{1}$$

$$attentionMatrix = softmax(\frac{Q \times K^T}{\sqrt{P}}) \tag{2}$$

Each row of the attention matrix indicates the weight of all embedding patches to the corresponding patch. Softmax operation ensures the sum of each row equals 1. By performing matrix multiplication on the attention matrix and V, each row of the final output is the weighted sum of all embeddings.

$$attention = attentionMatix \times V \tag{3}$$

The uniqueness of the attention matrix is that it is learned from scratch without inductive bias. As a result, the nearby and faraway patches have an equal chance of gaining weight. This unbiasedness of the attention matrix is ideal for the vague concept of synchrony. For example, LSTM has an inductive bias that assumes later (newer) input contributes more to the prediction. However, for synchrony, the definitions are either behaviors aligned simultaneously or with unspecified time delay. Neither case fits perfectly with LSTM's bias.

3.3 Embedding Methods

Since our study focused on a dyadic interaction dataset, the input actually consists of features from both the target participant (blue) and the interaction partner (red) for every sample, shown at the top of Fig. 4. First, the separate embedding, as illustrated in Fig. 4 mid left, is the basic setup that all patches access to each other without imposed limitation. Second, to reduce training difficulty, the same frame embedding embeds the dyadic sample of the same frame into a single patch, shown at the mid right of Fig. 4. This embedding is based on strict case synchrony that defines the behavior aligned at exact timing. This method reduces the number of patches by half and simplifies the attention matrix.

If the separate embedding model is adequately trained, the embeddings will be fit towards the engagement task. In another sense, these embeddings are projections of the initial modality on a particular embedding domain. Furthermore, this domain is trained to represent necessary information for the engagement prediction task. We hypothesize that the cosine similarity of these projections is akin to the modality similarity for measuring synchrony. Therefore, the third method takes the embeddings from a trained separate embedding model and performs a dynamic warping algorithm (DTW) over the target and the partner sequence to obtain a matching table from the target to the partner. This matching represents the most synchronized partner sample of each target sample within the input sequence. Finally, the matching pair embedding embeds the matching pairs into patches. This method also reduced the number of patches by half.

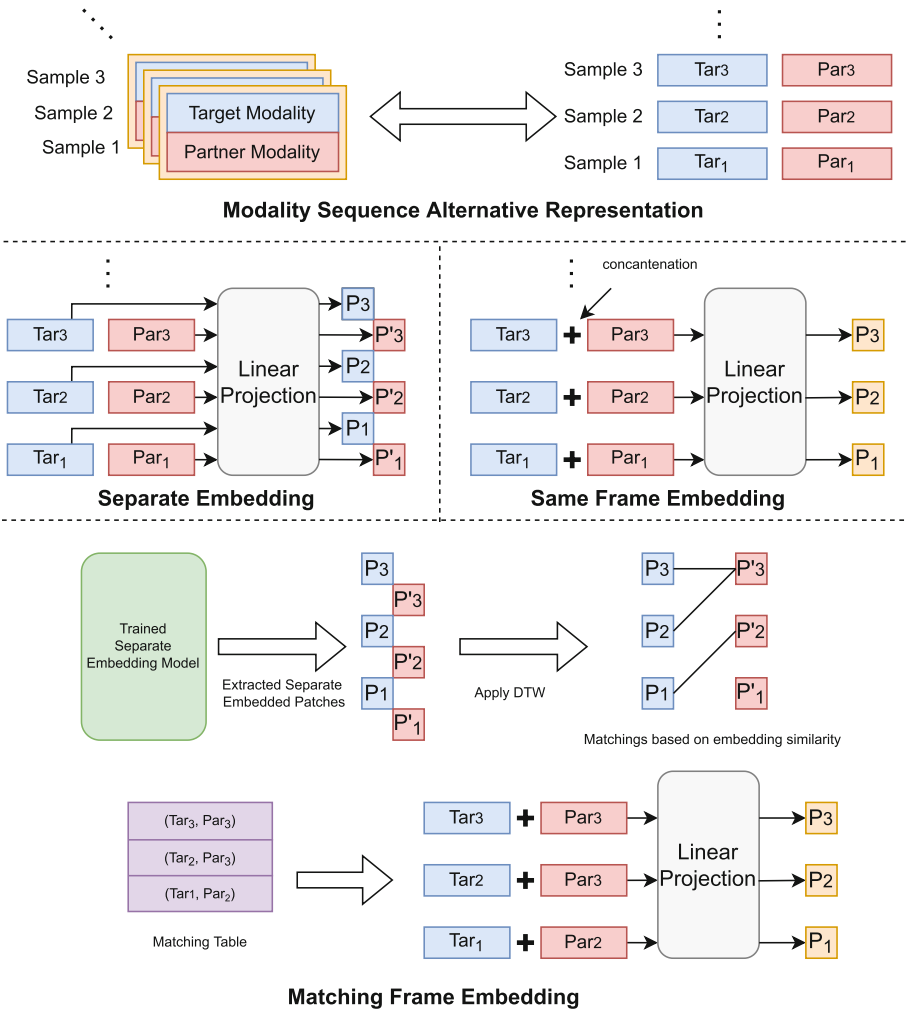

Fig. 4. Embedding Details - The topmost section describes an alternate representation of the dyadic modality input sequence. In this alternative representation, Tar_k stands for the kth input sample of the target participant. Similarly, Par_k stands for the kth input sample of the partner. In the mid-left section, Tar_k is embedded into P_k, where P represents embedded patches. Likewise, Par_k is embedded into P'_k, which means each sample will be embedded into two separate patches. In the mid-right section, both the target participant and partner of the same sample k will be concatenated first, then embedded into one P_k, where the yellow color of P indicates a mix of both participants. Finally, in the bottom section, the trained separate embedding model serves as a feature extractor. The matching table contains matched pairs from the dynamic time-warping algorithm (marked as DTW in the figure). Dyadic samples are concatenated based on the matching table and embedded input patches. In this schema, P_k always contains Tar_k, but not necessarily Par_k. (Color figure online)

4 Experiments

4.1 Data

We experimented with our models on the NoXi dataset, a dyadic interaction corpus of an expert sharing knowledge with a novice [5]. One great feature of the NoXi dataset is the open-source database, which provides frame-level annotations and the original sensor data. We mainly used the engagement labels under the annotator gold standard, which annotated the data with continuous engagement values between 0 and 1. Additionally, we downloaded all other samples with available annotations of the same engagement scale to extend the sample size. In total, we downloaded 27 sessions with the face, body skeleton, action unit, head orientation sensor data, and continuous engagement annotation ranging from 0 to 1.

4.2 Models

All models consist of 4 parallel streams (face, skeleton, action unit, and head orientation) and a regression head. The backbone model consists of one linear embedding layer, an attention block, and one linear layer for resizing output. The attention block is the standard attention block introduced in the transformer. The late fusion prediction head is a two-layer multi-layer perceptron with hidden layer ratio of 4.

The differences across different models reside in the embedding layer. This part follows the three methods of separate embedding, same-frame embedding, and matching frame embedding. As a baseline, we experimented with a two stream LSTM model with each stream process all four modalities of the target or the partner, similar to the model described in [10]. Additionally, we experimented with three attention blocks instead of one for same frame embedding and separate embedding to test if a deeper and larger model improves the result. For the matching frame embedding, we conducted an extra experiment that directly embeds the extracted embeddings instead of embedding original modality The window constraint for time dynamic warping is set from the present to 3 s in the past. Additionally, the algorithm cannot skip the target participant, details shown in Algorithms 1 and 2. In Algorithm 1, X and Y are target patches and partner patches respectively, and W is the window constraint which is 75 samples (3 s). The output of Algorithm 1, the DTW cost table, is the input for Algorithm 2 to calculate the optimal path which consists of optimal matching pairs.

We used mean square error as loss function. The training adopted the leave-one-out strategy. The first session from Paris serves as the testing data, and the training utilized the remaining 26 sessions. We set the initial learning rate as 1e−4, the default attention blocks as 1, dropout rate as 0.3. To reduce the training time, the input is limited to 250 frames with a striding of 5 frames and we embed 5 frames as a single patch. All experiments ran for 50 epochs on GPU with a fixed random seed of 22718.

Algorithm 1. CostTable(X,Y,W)

Ensure: $|X| = |Y|$
  $N \leftarrow |X| + 1$
  $dtw[] \leftarrow new[N \times N]$
  for $i \leftarrow 0; i < N; i++$ **do**
    for $j \leftarrow 0; j < N; j++$ **do**
      $dtw[i, j] \leftarrow \infty$ ▷ Initialize costs to infinity
    end for
  end for
  $dtw(0, 0) \leftarrow 0$
  for $i \leftarrow 1; i < N; i++$ **do**
    for $j \leftarrow max(1, i - W); j < i + 1; j++$ **do** ▷ loop with window constraint
      $cost \leftarrow distance(X[i], Y[j])$
      $prev \leftarrow min(dtw[i - 1, j], dtw[i - 1, j - 1]$ ▷ no skipping for target
      $dtw[i, j] \leftarrow prev + cost$
    end for
  end for
  return dtw

Algorithm 2. TracePath(dtw)

Ensure: $rows(dtw) = columns(dtw)$
  $path \leftarrow new[]$
  $N \leftarrow rows(dtw)$
  $min \leftarrow \infty$
  for $j \leftarrow 0; j < N; j++$ **do**
    if $dtw[N - 1, j] < min$ **then**
      $min \leftarrow dtw[N - 1, j]$
      $J \leftarrow j$ ▷ find the index of minimum total cost
    end if
  end for
  $i \leftarrow N - 1$
  $j \leftarrow J$
  while $i \neq 1$ **do**
    $prev \leftarrow min(dtw[i - 1, j], dtw[i - 1, j - 1])$ ▷ no skipping for target
    if $dtw[i - 1, j] = prev$ **then**
      $i \leftarrow i - 1$
    else if $dtw[i - 1, j - 1] = prev$ **then**
      $i \leftarrow i - 1$
      $j \leftarrow j - 1$
    end if
    $path$ add (i, j)
  end while
  return $path$

5 Results and Discussion

5.1 Experiment Results

Table 1. Experiment results - Mean Square Error and pseudo accuracy

Experiment Models	MSE	Pseudo Accuracy (±0.1)
Two stream LSTM (baseline)	0.0480	0.2750
Separate Embedding	**0.0278**	**0.3998**
Same Frame Embedding	0.0310	0.3228
Matching Frame Embedding	0.0361	0.3035
Separate Embedding with 3 Attention Blocks	0.0392	0.2754
Same Frame Embedding with 3 Attention Blocks	0.0303	0.3092
Matching Frame Embedding with Embedding as input	0.0389	0.2757

Since the annotations are continuous, we cannot simply calculate the accuracy of our results. Instead, we evaluate the performance by MSE and pseudo accuracy. First, MSE indicates the overall deviation of the prediction from the ground truth. Second, we set predictions within a ±0.1 tolerance range of the ground truth as positive predictions to calculate pseudo accuracy.

Table 1 lists each experiment's testing MSE and pseudo accuracy. The pseudo accuracy aligned with MSE, which showed no unexpected exceptions. This result indicated that the predictions from all models generally followed that lower MSE had higher pseudo accuracy. In other words, no particular model had most of its predictions accurate but had a small number of severely erroneous results that contributed to the majority of the MSE.

Our results showed that the intra-modality structure did improve the engagement level prediction accuracy. All models that adopted inter-person intra-modality structure outperformed the two-stream LSTM baseline model in MSE. However, except the separate embedding model, there is no significant improvement in pseudo accuracy.

The embedding methods also created distinct differences in the results. The separate embedding model, which had to learn the largest attention matrix, turned out to be the best-performed model. The same frame embedding model followed as the intermediate result. The worst result was the matching frame embedding models. Additionally, given our experiment setup, the deeper models with three attention blocks did not outperform their simpler counterparts. Finally, for matching-frame embedding, directly matching the extracted embedding patches resulted in an even worse result than matching the original modality features.

5.2 Discussions

In this section, we first discuss the reasons behind the experiment results. Then we highlight the limitations of this work.

Result Discussion

Modality Independent and Person Independent. The two-stream baseline processes each participant as an independent entity. In each stream, time series features of all modalities are processed in the sequential layers and contribute to the output. This structure properly utilizes the multi-modal features for each participant. However, the output of each stream contains mixed information about all input modalities. Once the information is intertwined, it would be improbable to learn the synchrony phenomenon as synchrony is observed within the same modality. A similar issue also applies to our multi-stream inter-person intra-modal model. Our models allow each modality stream to learn from two participants, which can also be considered treating each modality independently. The output of each stream contains mixed information from both participants, which can hinder the learning of cross-modality information of each person. In other words, the two-stream model prioritizes the cross-modal learning of each participant, while our model prioritizes the interpersonal synchrony of each modality. Our results showed that the inter-person intra-modality models all had better results than the two-stream baseline. That is interpersonal information outweighs cross-modal features in the dyadic scenarios. But different results may apply in different scenarios. Extensive experiments with different settings would be desired to validate our findings. Furthermore, these designs are not necessarily mutually exclusive. Developing models covering both structures as submodels with a weighted fusion can be promising.

Issues for Hand-Crafted Pair Embedding. Among our models, the two hand-crafted pair embedding models failed to outperform the baseline model by a considerable margin, especially for the pseudo accuracy. This is because no matter which hand-crafted method, we limited the possible cases of synchrony. Notably, we observed that the matching frame embedding model was harshly underperforming. There are two major reasons for this. First, the matching frame embedding depends on the matching algorithm which requires a reliable similarity function. In our hypothesis, a trained separate embedding model can serve as a feature extractor so that the cosine similarity of the extracted feature serves as the similarity function. However, our feature extractor was severely undertrained compared with commonly used feature extractors. As a result, the features could not be adequately projected into the new domain, and the cosine similarity of the extracted features could not appropriately reflect the modality similarity. Second, there is a missing sample problem created by the constraints for dynamic time warping. The constraints are that the target participant samples cannot be skipped or matched to future partner samples. As a result, the partner samples close to the present are discarded in nearly all cases. The only possible matching that uses the present partner sample is the same frame matching. Otherwise, the present partner sample will be the future sample for the rest of the target samples, which is prohibited by constraints. That is, the matching frame model can never obtain the newest features of the partner. Therefore,

we need a better solution for the modality feature extraction and an improved matching algorithm.

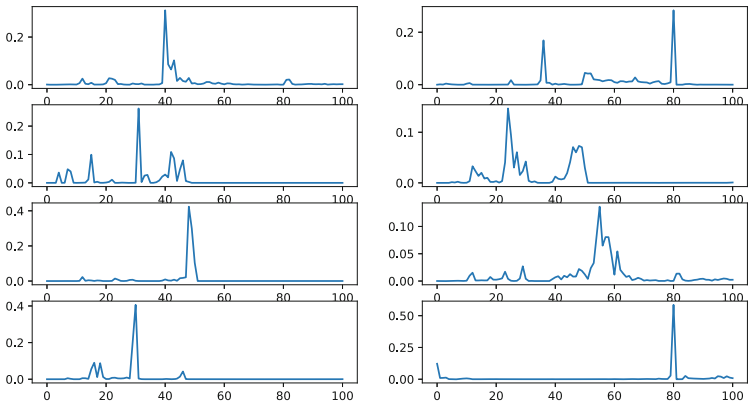

Fig. 5. A sample of attention weights for the extra patch from all 8 heads in the separate embedding model - Each figure represents an attention head. The x-axis indicates the index of the patches, where 0 is the extra patch, 1 to 50 are the target participant from past to present, and 51 to 100 are the partner from past to present. The y-axis shows the weights of which the sum equals 1.

Other Limitations

Explainability. We were unable to model synchrony directly. In the early stage of this work, we attempted to model the synchrony itself. For example, we used high-level features such as smiles and head nods and tried to define a successful case of synchrony. However, we found these high-level behaviors very individual-based. For example, some participants smile habitually, and some rarely smile. Eventually, we adopted a data-driven approach using base-level sensor data, which is less interpretable. Figure 5 shows a sample of attention weights for the extra patch, illustrated in Fig. 3, which is the only patch used for prediction. Therefore, in the case of one block of attention, we only need to examine the attention matrix's first row, i.e., the row for the extra patch. The figure indicates most heads are trained to get information from the target participant only, but some heads, three heads on the right side, partially get weights from the partner. As all heads in multi-head attention contribute to the output with learned weights, temporal information from both participants affected the prediction. This matches the our hypothesis that interpersonal information benefits

the prediction, but getting any further explanation is challenging. We cannot be certain that these weights are a manifestation of synchrony.

Data and Annotation. We used continuous annotation because it was the most viable type of annotation across all sessions. However, for engagement level prediction, such precision is unnecessary. Moreover, training a regression task is significantly more problematic than a classification task. Another aspect of annotation limitation is the difficulty of obtaining frame-level annotations. In NoXi dataset, each session contains tens of thousands of frames. Annotating on such a scale is a daunting task for either crowd annotating or expert annotating.

Limited Optimization. Many machine learning techniques, such as hyperparameter grid search, can help improve the results. However, since our experiments are on the frame level, the training takes much longer than the conversation level tasks. As a result, our experiment could not optimize each model; instead adopted similar hyperparameters for all experiments. There is a possibility that some results can be significantly improved if supported by proper optimization techniques.

Individual Modality Effects. A common aspect of multi-modality research is to experiment with the contribution of each modality and different combinations of modalities. During our experiments, we encountered distinct attention distributions between parallel models. However, we considered this aspect beyond the scope of this paper. Which modality is better suited for the inter-person intra-modality structure remains an undiscussed topic.

6 Conclusion

This work explored the gap between engagement modeling and interpersonal synchrony. To enable models to capture the behavior synchrony between dyadic partners, we developed an inter-person intra-modality attention based model with different embedding schemas. Our experiments verified the positive impact of inter-person intra-modality features in engagement level prediction. We showed that time series feature processing grouped by each modality produced better results in the dyadic interaction scenario than those grouped by each participant. In future work, we plan to extend the model to cover both intra-modal inter-person and grouped-by-person submodels, explore different methods to assist training, and expand the training data to support more complex models.

Acknowledgement. This work was also partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 22K21304, No. 22H04860 and 22H00536), JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6) and JST Moonshot R&D program (JPMJMS2237-3).

References

1. Abdelrahman, A.A., Strazdas, D., Khalifa, A., Hintz, J., Hempel, T., Al-Hamadi, A.: Multimodal engagement prediction in multiperson human-robot interaction. IEEE Access **10**, 61980–61991 (2022). https://doi.org/10.1109/ACCESS.2022.3182469
2. Bernieri, F., Reznick, J., Rosenthal, R.: Synchrony, pseudosynchrony, and dissynchrony: measuring the entrainment process in mother-infant interactions. J. Pers. Soc. Psychol. **54**, 243–253 (1988). https://doi.org/10.1037/0022-3514.54.2.243
3. Bernieri, F., Rosenthal, R.: Interpersonal coordination: behavior matching and interactional synchrony. Fundamentals of Nonverbal Behavior. Studies in Emotion and Social Interaction, January 1991
4. Bohus, D., Horvitz, E.: Learning to predict engagement with a spoken dialog system in open-world settings. In: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 244–252. SIGDIAL 2009. Association for Computational Linguistics, USA (2009)
5. Cafaro, A., et al.: The NoXi database: multimodal recordings of mediated novice-expert interactions. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, pp. 350–359. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3136755.3136780
6. Chartrand, T.L., Bargh, J.A.: The chameleon effect: the perception-behavior link and social interaction. J. Pers. Soc. Psychol. **76**(6), 893–910 (1999)
7. Chartrand, T.L., Dalton, A.N.: Mimicry: its ubiquity, importance, and functionality (2009)
8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
9. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: a survey of evaluation methods across disciplines. IEEE Trans. Affect. Comput. **3**(3), 349–365 (2012). https://doi.org/10.1109/T-AFFC.2012.12
10. Dermouche, S., Pelachaud, C.: Engagement modeling in dyadic interaction. In: 2019 International Conference on Multimodal Interaction, ICMI 2019, pp. 440–445. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3340555.3353765
11. Glas, N., Pelachaud, C.: Definitions of engagement in human-agent interaction, pp. 944–949, September 2015. https://doi.org/10.1109/ACII.2015.7344688
12. Hadfield, J., Chalvatzaki, G., Koutras, P., Khamassi, M., Tzafestas, C.S., Maragos, P.: A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1251–1256 (2019). https://doi.org/10.1109/IROS40897.2019.8968443
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, pp. 770–778 (06 2016). https://doi.org/10.1109/CVPR.2016.90
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
15. Hu, Y., Cheng, X., Pan, Y., Hu, Y.: The intrapersonal and interpersonal consequences of interpersonal synchrony. Acta Psychologica **224**, 103513 (2022). https://doi.org/10.1016/j.actpsy.2022.103513
16. Kaur, A., Mustafa, A., Mehta, L., Dhall, A.: Prediction and localization of student engagement in the wild. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–8 (2018). https://doi.org/10.1109/DICTA.2018.8615851

17. Kendon, A.: Movement coordination in social interaction: some examples described. Acta Psychologica **32**, 101–125 (1970). https://doi.org/10.1016/0001-6918(70)90094-6
18. Kimura, R., Okada, S.: Personality trait classification based on co-occurrence pattern modeling with convolutional neural network. In: Stephanidis, C., et al. (eds.) HCII 2020. LNCS, vol. 12427, pp. 359–370. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60152-2_27
19. Kolesnikov, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale (2021)
20. Nezami, O.M., Dras, M., Hamey, L., Richards, D., Wan, S., Paris, C.: Automatic recognition of student engagement using deep learning and facial expression (2018). https://doi.org/10.48550/ARXIV.1808.02324
21. Oertel, C., et al.: Engagement in human-agent interaction: an overview. Front. Robot. AI **7** (2020). https://doi.org/10.3389/frobt.2020.00092
22. Poggi, I.: Isabella Poggi Mind, Hands, Face and Body A Goal and Belief View of Multimodal Communication, March 2022. https://doi.org/10.1515/9783110261318.627
23. Reddish, P., Fischer, R., Bulbulia, J.: Let's dance together: synchrony, shared intentionality and cooperation. PLOS ONE **8**(8), 1–13 (2013). https://doi.org/10.1371/journal.pone.0071182
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). https://doi.org/10.48550/ARXIV.1409.1556
25. Sun, X., Nijholt, A.: Multimodal embodied mimicry in interaction. In: Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A. (eds.) Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues. LNCS, vol. 6800, pp. 147–153. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25775-9_14
26. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 6000–6010. Curran Associates Inc., Red Hook, NY, USA (2017)
27. Wiltermuth, S., Heath, C.: Synchrony and cooperation. Psychol. Sci. **20**, 1–5 (2009). https://doi.org/10.1111/j.1467-9280.2008.02253.x
28. Yu, C., Aoki, P.M., Woodruff, A.: Detecting user engagement in everyday conversations (2004). https://doi.org/10.48550/ARXIV.CS/0410027