A Deep Neural Network Approach for Integrating Neural and Behavioral Signals: Multimodal Investigation with fNIRS Hyperscanning and Facial Expressions

Grace Qiyuan Miao q.miao@ucla.edu
Yanru (Joyce) Jiang yanrujiang@g.ucla.edu
Ashley Binnquist abinnquist@ucla.edu
Agnieszka Pluta apluta@psych.uw.edu.pl
Francis Steen steen@comm.ucla.edu
Rick Dale rdale@ucla.edu
Matthew Lieberman lieber@ucla.edu

Abstract

Conversations between people are characterized by complex nonlinear combinations of nonverbal and neurocognitive responses complementing the words that are spoken. New tools are needed to integrate these multimodal components into coherent models of conversation. We present a study and analysis pipeline for integrating multimodal measures of conversation. Data were collected using video recordings and functional near-infrared spectroscopy (fNIRS), a portable neuroimaging technology, during dyadic conversations among strangers (N=70 dyads). Rather than running discrete analyses of neural and nonverbal data, we introduce a pipeline to combine time series data from each modality into multimodal deep neural networks (DNNs) – including channel-based fNIRS signals and OpenFace data that quantifies facial expressions over time – using S2S-RNN-Autoencoders. We explored two measures to examine the resulting t-SNE space: distance and synchrony. We found that across the dimensions integrating neural and nonverbal input features, conversing dyads tend to stay closer together than permuted pairs. Dyads exhibit significantly higher synchrony in their covariation in this space compared to permuted pairs. The results suggest a mixed methodological integration may contribute to a deeper understanding of the dynamics of communication.

Keywords: conversations, multimodal dynamics, deep neural network (DNN), functional near-infrared spectroscopy (fNIRS), hyperscanning, brain-to-brain synchronization, dimension reduction, integrative pluralism.

Introduction

Conversations are complex combinations of verbal, nonverbal, and neurocognitive responses. This sheer complexity of conversation has been recognized by scholars across various disciplines, across cognitive science (Grosz & Hirschberg, 1992; Garrod & Pickering, 2004; Holle et al., 2012; Raczaszek-Leonardi, 2014; Özyürek, 2014; Galati & Brennan, 2014; Paxton et al., 2016; Mondada, 2016; Zima & Bergs, 2017; Rasenberg, Özyürek, & Dingemanse, 2020; Reece et al., 2023), including psycholinguistics (Iverson & Thelen, 1999; Willems et al., 2007; Holler et al., 2013; Pouw et al., 2020) and conversation analysis (Goodwin & Heritage, 1990; Schegloff, 1996; Stivers & Sidnell, 2005; Sidnell, 2006; Schegloff, 2007; Enfield & Sidnell, 2017; Goodwin, 2018; Stivers, 2021). Understanding what underlies and explains this complexity represents a still-evolving domain of research.

Even a single conversational turn is a high-dimensional performance involving behavioral, cognitive, and neural processes, often quite distinct. Speakers manage both the perceptuomotor characteristics of nonverbal behaviors along with more abstract words, phrases, and meanings. These processes unfold at widely varying timescales. Eye movements and social attention can change on the order of milliseconds, while topics of conversation are managed more slowly across minutes. How can we develop novel approaches to integrate all these elements in a way that mirrors how the mind processes multimodal information?

Cognitive neuroscientists have found a neurobiological metric that characterizes whether people are 'in sync' during interactions called neural synchrony, the temporal correspondence in neural activity patterns during interpersonal interaction, with cross-brain alignment indicating the coupling of people's separate neurocognitive systems (Lieberman, 2022). Studies using fMRI have demonstrated that neural synchrony can be an effective neurobiological marker for like-mindedness. For example, higher neural synchrony during video-viewing is associated with more similar interpretations of the video content (Nguyen et al., 2019) and closeness in the real-life social network (Parkinson et al., 2018). However, one major limitation of using fMRI for social interaction research is that MRI scanners isolate participants from the outside world and prevent natural conversations.

Addressing this limitation, functional near-infrared spectroscopy (fNIRS), a portable neuroimaging technology, allows studies of social interactions in their natural environment. Certain fNIRS studies using the neural synchrony metric have been explicitly conducted to investigate verbal communication between dyads (see Kelsen et al., 2022; Jiang et al., 2012; Zhang et al., 2018).

While these fMRI and fNIRS studies revealed that neural synchrony can effectively show whether two people are "on the same page" (Dieffenbach et al., 2021), its application in multimodal analysis presents challenges. Synchrony is clear in activities like watching a video together but less so in conversations that involve turn-taking. Moreover, there is a discernible gap in the literature regarding the integration of neural and behavioral signals during social interactions. This divide in the research highlights a critical oversight in social neuroscience and indicates a pressing need for new approaches to explore interactions between pairs. Our pipeline contributes to multimodal integration and can be extended to other modalities such as body movement and speech.

Progress in deep neural networks (DNNs) has facilitated the conversion of complex data, including images (Caron et al., 2018), audio (Cramer et al., 2019; Purwins et al., 2019), and text (Rosen & Dale, 2023) into numerical representations known as embeddings. These condensed numerical

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

representations facilitate analysis and uncover underlying patterns that may not be immediately discernible. For instance, they can be employed to establish correlations between dense vectors and brain signals or other dependent variables (Goldstein et al., 2022; Schrimpf et al., 2021). DNNs are valuable when grappling with sparse and high-dimensional data, and address the alignment challenge across multiple modalities and time scales.

Numerous studies have employed DNNs to transform low-level features, including visual (McMahon et al., 2023), semantic (Huth et al., 2016; Heilbron et al., 2022), and phonetic elements (Gong et al., 2023) into vectorized representations for brain-predictive modeling. However, previous research has often focused on each modality independently, stemming from the divergent data structures and temporal resolutions inherent to each modality, thus limiting the exploration of multimodal integration.

Our approach to modality alignment uses a recurrent neural network (RNN) architecture capable of processing time-series data by considering temporal dependencies (Tealab, 2018). RNN solves the challenge particularly by converting sequential data into one dense embedding (e.g., representing a sequence of word tokens with a sentence embedding), allowing multimodal signals with different frequencies to be integrated at the embedding level. To generate a dense embedding that represents the input signals without conditioning on a specific task, we constructed a Sequence-to-Sequence RNN Autoencoder (S2S-RNN- Autoencoder), utilizing the same input and output (Strobelt et al., 2019; Lyu et al., 2018). Autoencoders enable us to integrate multiple signal channels (e.g., facial movement and neuro signals) by projecting them into the latent space through self-supervised vectorization (Jiang et al., 2024).

Integrating the methods from cognitive science, social neuroscience, and machine learning, this paper presents a new pipeline that tackles the challenge of modality integration by utilizing DNNs for both unimodal and multimodal representations. These dense integrative representations, in combination with other downstream dimensionality reduction analyses (such as t-SNE), allow researchers to analyze sequential data in a latent space that can be theoretically meaningful (e.g., synchrony and proximity). Together, this pipeline enables scholars to better triangulate how social interactions are supported by different modalities such as the brain and facial expressions.

Data

Procedure

Two strangers engaged in a get-to-know-you conversation while seated face-to-face, with topics of discussion displayed on a computer screen one-by-one (Fig 1). The original design of this experiment contrasted depth of topic (Kardas et al., 2022). Example topics include: How's the weather today? How often do you get your hair cut? What is one of the more embarrassing moments in your life? Participants are instructed to try to stay on topic, and when they are done with one topic, click the button to move on to the next one. Every session is designed to last for 20 minutes and occurs without the presence of experimenters, thereby allowing a natural conversation flow.

Figure 1: Experiment setup. Stranger dyads are equipped with fNIRS. Three GoPro cameras are placed in the room – two capturing the two participants' facial expressions, and one capturing the scene from a third-person perspective.

During the experiment, participants wore a functional near-infrared spectroscopy (fNIRS) rig with coverage of cortical regions implicated in social interactions (Fig 2), such as mentalizing (Gallagher et al., 2000, Wang et al., 2018). Three GoPro cameras are placed in the room to record conversations and nonverbal behaviors. Specifically, one camera is placed in front of each participant to record their facial expressions and the third camera captures both participants together. Participants also complete questionnaires about their personal traits and experiences with the conversation.

Participants

We recruited 70 dyads (21 male-male, 25 female-female, and 24 male-female) from the UCLA Departments of Psychology and Communication subject pools as well as flyers on the UCLA campus. The study was approved by the UCLA IRB (#22-001209) and informed consent was obtained from all subjects.

Neural Data Acquisition

Participants were scanned using a mobile fNIRS system (NIRSport2 by NIRx Medical Technologies, LLC, NY).

The probe layout was comprised of 16 light sources and 16 detectors with a 3-cm average source-detector separation distance, which forms 42 channels (source-detector pairs) for partial-brain coverage across mentalizing (i.e., medial prefrontal cortex (mPFC) and temporo-parietal junction (TPJ)) and working memory regions (i.e., lateral prefrontal cortex (lPFC) and superior parietal lobule (SPL)) (Fig 2). The montage layout (Fig 2) was created in accordance with the 10-10 UI external positioning system to ensure consistency across head sizes. We measured participants' head sizes and then fitted them with caps of appropriate sizes that affix the optodes to the scalp. Raw light intensity data was collected at a sampling rate of 5.09 Hz at wavelengths of 760 and 850 nm.

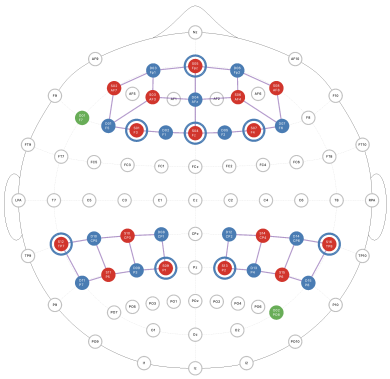

Figure 2: fNIRS montage consists of 42 channels for partial-brain coverage of cortical regions implicated in social interactions (i.e., medial prefrontal cortex (mPFC), temporo-parietal junction (TPJ), lateral prefrontal cortex (lPFC) and superior parietal lobule (SPL)).

Data Processing

Facial Video Data Processing

For behavioral data, we focused on facial keypoint dynamics under the framework of Facial Action Coding System (FACS), an anatomically based system for describing all visually discernible facial movements (Ekman & Friesen, 1978). This comprehensive system breaks down facial expressions into individual components of muscle movement, called Action Units (AUs). This project employs OpenFace, a computer vision system that automates the detection and analysis of AUs (Baltrušaitis et al., 2016)

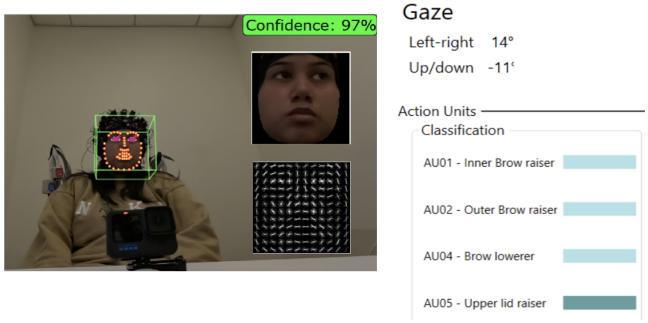

Figure 3: Facial Video Data Processing. Demonstration for facial movement analysis using OpenFace. *Left*: Feature extraction based on the Facial Action Coding System (FACS), an anatomically based system for describing visually discernible facial movements. *Top right*: The left-right and up/down angles of gaze is reported per video frame. *Bottom right*: FACS-based individual components of muscle movement, called Action Units, are reported.

Neural Data Processing

Collected NIRS data underwent a comprehensive preprocessing pipeline. This pipeline[1] utilized custom scripts in MATLAB alongside the Homer2 software suite (Huppert, Diamond, Franceschini, & Boas, 2009), adhering to established fNIRS best practices (Yücel, 2021). Emphasis was placed on analyzing oxyhemoglobin (HbO) concentrations, which prior research has indicated are more responsive to changes in cerebral blood flow than deoxyhemoglobin (HbR) levels work (Pan et al., 2017).

The preprocessing began with removing unrelated data – each time-course was truncated based on a trigger that indicated the start of the conversation. Noisy and oversaturated channels were identified and excluded using a modified quartile coefficient of dispersion (Bonett, 2006), with specific thresholds adjusted for the sampling rate (Cthresh = 0.6 – 0.03*sampling rate).

Further refinement of the data included corrections for motion and non-neural changes in blood oxygenation. To address motion artifacts, discrete wavelet transform techniques (Molavi & Dumont, 2012) were performed to remove spike artifacts. To address non-neural physiological influences (e.g., cardiac and respiratory rhythms) and baseline drift, a conservative bandpass filter (0.008-0.2 Hz) was applied. Past work suggests that the cognitive dynamics of interest in this study are primarily manifested in lower frequency ranges (Sasai et al., 2011; Zuo et al., 2010).

Filtered data were then transformed from optical density to hemoglobin concentration values. This conversion used the modified Beer Lambert Law (MBLL) with a standard differential path length filter [6, 6], commonly applied to adult cortical tissue to account for light dispersion.

The final quality control step involved an autocorrelation change assessment to gauge the impact of motion correction. Channels displaying a substantial change in autocorrelation (exceeding a threshold of $r = 0.1$) were deemed significantly influenced by motion and thus excluded from subsequent analyses.

Multimodal Data Processing

Model Architectures An RNN is a neural network capable of modeling sequential data and time-dependent tasks (Tealab, 2018), such as text generation, speech recognition, and stock market prediction. RNN represents an iterative function that takes an input sequence (x) and an internal state (h) from the previous timestep (t - 1) to predict the current timestep (t), then updates the state as follows:

$$h_t = f(x_{t-1}, h_{t-1}) \quad t\ in\ \{0, 1, 2, \ldots T-1\}$$

We selected the RNN model for representing integrated multimodality because it can process temporal information under the assumption that the facial AUs and fNIRS signals in each timestep depend on signals in the previous timesteps (Jiang, 2023). A Long-Short-Term-Memory (LSTM) RNN was chosen over the vanilla RNN because the latter experienced the vanishing-gradient problem during model training, which inhibited it from effectively leveraging context between elements by maintaining its internal state

[1] https://github.com/abinnquist/fNIRSpreProcessing

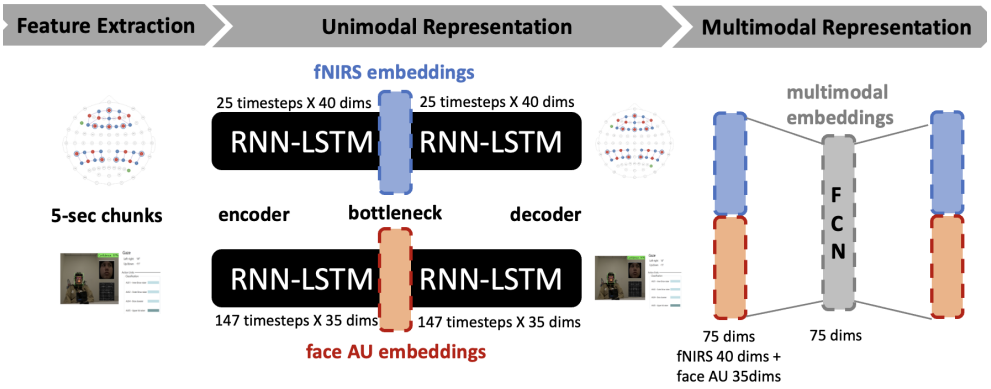

Figure 4: Flowchart and Architectures for Multimodal Data Processing. Through feature extraction, we obtain multidimensional sequences for both fNIRS and facial AUs, and chunk them into 5-second intervals. Then, each modality is vectorized using an S2S-RNN-Autoencoder. Finally, the vectorized modalities are concatenated to form larger embeddings to achieve integrative multimodal embeddings using an FCN-Autoencoder.

throughout the sequence (Sherstinsky, 2020). LSTM, which is represented by the function f in the equation, was introduced here as an additional state variable (i.e., the cell state) for controlling specific information that needed to be kept or updated while processing the entire sequence (Joo et al., 2019). As a result, LSTM effectively reduced the vanishing-gradient problem encountered by RNN (Sherstinsky, 2020).

Next, we constructed multiple autoencoders to independently vectorize facial AUs and fNIRS signals, and then integrated them. An autoencoder is a neural network architecture that contains three components: encoder, bottleneck, and decoder (Michelucci, 2022). The model learns to reconstruct the input data by compressing (encoder) it into a lower-dimensional embedding (bottleneck), then reconstructing it back into its original form or any target form (decoder). When the output is the same as the input, this process allows the network to learn a compressed, latent representation of the input data that captures the most salient features of the original data. This method is commonly used for dimensionality reduction and self-supervision. In this case, because h_t represents a lower-dimensional compression that can nevertheless reconstruct temporal sequences of multimodal behavior, this embedding quantitatively summarizes how these channels go together in that moment of the interaction (details shown in Fig. 4).

Multimodal Integration In our pipeline[2], time-series data for both fNIRS and facial AUs were initially segmented into approximately 5-second chunks (approximation due to frequency misalignment between multiple modalities). Due to the different temporal resolutions between video data (around 30 FPS) and fNIRS (5.09 Hz), we first constructed separate S2S-RNN-Autoencoders for each modality and then employed another fully-connected-network (FCN)-Autoencoder to integrate these modalities at the chunk-level. While downsampling the high-frequency modality is also a viable option for enforcing temporal alignment across modalities, this approach would unavoidably lead to information loss during the sampling process. Therefore, we opted to vectorize facial movement and fNIRS independently and then integrate them at each 5-second interval.

All autoencoders were trained using the Adam optimizer with a learning rate of 0.001. The loss function used was the sum of mean squared errors (MSE) that measure the average squared difference between the input data and the corresponding reconstructed output across all dimensions. Intuitively, this MSE calculates the loss to ensure that similar input patterns are mapped to similar representations. The batch size was set to 32, and all models were trained for 20 epochs.

Each 5-second data point had a shape of 147 timesteps (i.e., frames) x 35 dimensions for facial AUs and 25 timesteps (i.e., frames) x 40 dimensions for fNIRS signals. The S2S-RNN-Autoencoder comprised one layer of LSTM for both the encoder and decoder, with each LSTM layer having input, output, and hidden dimensions matching the dimension of each data point (i.e., 35 dimensions for AUs and 40 dimensions for fNIRS).

After vectorizing each modality at the 5-second interval, we constructed an FCN-Autoencoder with one layer of FCN for both the encoder and decoder. The input, output, and encoding dimensions were all set to 75, which is the combination of dimensions from facial AUs and fNIRS channels. This autoencoder allowed us to obtain embeddings that vectorize the modality integration for time-series data.

Finally, in order to interpret the high-dimensional data, we applied t-SNE, a technique for visualizing high-dimensional data in a lower-dimensional space (van der Maaten & Hinton, 2008). Applying the Rtsne function (Krijthe et al., 2018) to these vectorized multimodal embeddings, we obtained a 3-dimensional representation of dyads' conversations. Figure 5 plots the first and second dimensions of the t-SNE space as a demonstration of the resulting 3-dimensional multimodal data. The visualization indicates that the clustering of embeddings (by dyad) is preliminarily aligned with our expectations: chunks from the same dyad tend to be closer to each other, without obvious outlier data points. A more sophisticated analysis was conducted in the following section.

[2] github.com/JoyceJiang73/Multimodal-Integration-Autoencoders

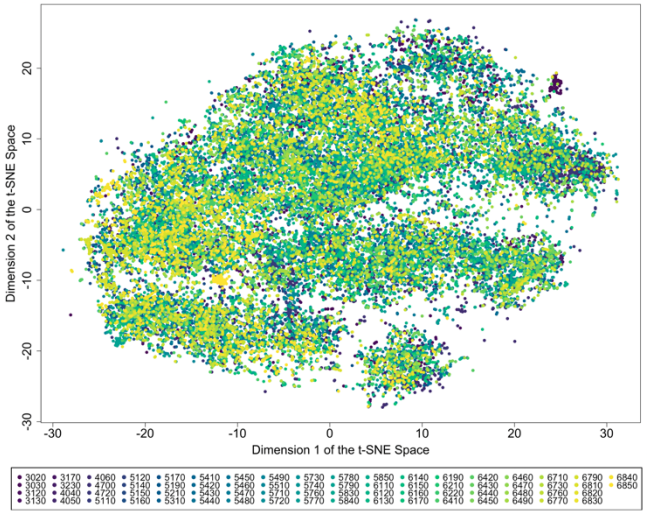

Figure 5: Visualizing the t-distributed Stochastic Neighbor Embedding (t-SNE) Space. Each color represents one of the 70 dyads in the dataset. Every point corresponds to a five-second segment of input. The x and y axes represent the first and second dimensions of the 3-dimensional representation of dyadic conversations within the t-SNE space.

Dyadic Interactions in t-SNE Space

We define each dyad of two individuals engaging in a conversation as a distinct cluster (of time points) to examine questions about dyadic interactions: Do dyads in different conversations share unique interactive signatures? How do these patterns differ from those observed in permuted pairs (i.e. surrogate baseline)?

We first chose a random time slice for an individual, then chose comparison slices at the same time point from two different individuals: (1) the other individual within the same dyad and (2) an individual randomly sampled from a different dyad. The first comparison aimed to capture the dynamics of a dyad. In contrast, the second comparison involving permuted pairs served as a control to represent non-interacting pairs. This process was iteratively conducted 10,000 times across the entire dataset, comprising 33,997 time slices and 70 dyads. As a result, we generated 10,000 within-cluster (i.e. dyads) and 10,000 across-cluster (i.e. permuted pairs) comparisons.

We propose two measures to examine the resulting t-SNE space: proximity and synchrony. Results show that conversing dyads score differently on these measures than permuted pairs, indicating unique interactive signatures.

Distance

The first measure is inspired by Pickering and Garrod (2004), whose prominent framework predicts that interacting participants tend to align their behaviors, which may result from a priming mechanism that drives probabilistic structure of interaction to be more behaviorally similar. In the lower-dimensional t-SNE space we explore, this manifests as the proximity of facial expressions and neural signals, which we define as the average distance between members of a dyad in the t-SNE space computed by Euclidean distance.

Linear regression analysis for distance yielded a clear result. Comparing across the first two dimensions of the t-SNE space, the average distance was significantly lower in within-cluster analyses compared to across-cluster analyses ($\beta = -2.28$, $t = -14.46$, $p < .00001$). In other words, the average distance between dyads engaging in conversations is significantly shorter than permuted pairs.

These tentative results are confirmed by a multilevel regression analysis. In all three dimensions of the t-SNE space, we conducted multilevel analysis with fixed effects as the distinction between dyads and permuted pairs, and random effects as variance within clusters. Results show that individuals in dyads engaging in conversations are significantly closer to each other ($t = 15.78$, $p < 0.00001$).

Our analysis demonstrates that across the fundamental dimensions integrating neural and nonverbal input features, dyads tend to stay closer together than permuted pairs. These results extend ideas about alignment (Pickering & Garrod, 2004) to a framework that highlights neuro-behavioral integration.

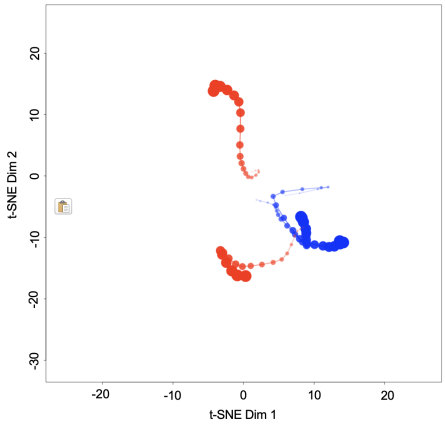

Figure 6: Illustration of proximity. Each dot represents one 5-second segment of data in the t-SNE space. Colors represent dyads. The blue dyad has high proximity at the selected time slice, and the red has low proximity.

Synchrony

Various researchers have suggested that humans approximate coupled oscillators while interacting (Strogatz & Stewart, 1993; Wilson & Wilson, 2005; Wiltshire et al., 2020; Miao et al., 2023). In the course of a conversation, participants attending to each other show a tendency to synchronize their movements (Dahan et al., 2016; Wiltshire et al., 2020; Sabharwal et al., 2022). Similarly, neural synchrony has been associated with a tendency towards greater social connection (Parkinson et al., 2018). In our analysis, synchrony appears as concurrent movements of two individuals in the same direction at the same time in the t-SNE space. We quantify dyadic synchrony as two individuals' co-variation of position across the t-SNE space measured by Pearson's correlation (r).

Given the inherent clustering of our data by dyads and the presence of individual differences, multilevel analysis was

deemed most suitable for examining dyadic synchrony. The dependent variable in our analysis was the average Pearson's *r* across three dimensions within the t-SNE space. The model accounts for fixed effects based on the distinction between dyads and permuted pairs, while the random effects were attributed to the variance within clusters, thereby addressing the issue of data non-independence within these clusters.

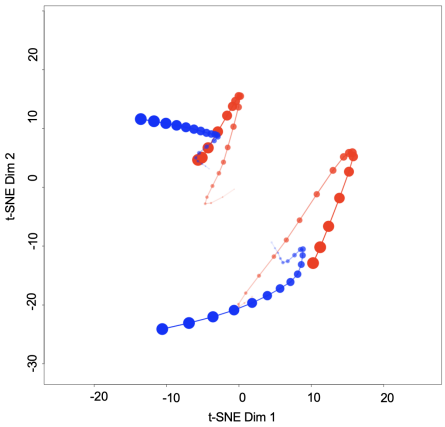

Figure 7: Illustration of synchrony. Each dot represents one 5-second segment of multimodal data in the t-SNE space. Colors represent distinct dyads. Individuals in both dyads are highly synchronous with their partner.

Results suggest that when pairs were permuted, the resulting mean correlation coefficient approached zero, indicating a significant decrease in synchrony compared to dyads (t = 4.94, p < 0.00001). On the contrary, for dyads engaging in conversations, the average correlation coefficient was significantly greater than zero at 0.05 (t = 6.03, p < 0.00001). This finding substantiates that dyadic synchrony is higher in real interactions within a dyad than baseline.

General Discussion

We introduced a novel data processing pipeline that integrates video and fNIRS recordings during natural conversations across 70 dyads. To explore interactive signatures inside the resulting multimodal, we compared conversing dyads with permuted pairs (i.e., surrogate baseline). We found that conversing dyads are significantly closer to each other than permuted pairs. Dyads also exhibit significantly higher synchrony in their covariation in this space compared to permuted pairs.

Using this pipeline, future papers could examine such datasets to investigate differences amongst conversing dyads focusing on a variety of theoretical questions. We aim to pursue such questions, including: Do interactive signatures differ across dyads that reported varied levels of connection after conversations? How does the depth of conversational topics affect interactive signatures and reported connection?

Our approach uses statistical models to integrate multimodal data in a way that may reflect related human cognitive processes. Many multimodal investigations tend to disassemble complex interactions into parts (e.g., words, gestures, facial expressions, and neural fluctuations) and analyze them separately. Yet, in an important sense, the human brain does not have the privilege of the scientist – to disassemble everything and carefully analyze it in parts. Our approach pursues how such integration may be conducted on the fly, during natural interaction.

Compared to other multimodal investigations that use innovative approaches to combine modalities without DNNs – such as multidimensional recurrence quantification analysis (MdRQA) (Wallot et al., 2016; Amon et al., 2019) and multivariate Surrogate Synchrony (mv-SUSY) (Tschacher & Meier, 2019) – our approach enables the conversion of sequential multimodal signals that can have different frequencies into dense embeddings through DNN architectures. This allows integration at the embedding level without substantial information loss. Additionally, the introduced pipeline serves as the first step for more comprehensive multimodal integration and investigation, which can be extended to other modalities such as body movement and speech. Given that semantic signals are usually not aligned with behavioral signals in a uniformly consistent way, using the DNN-based approach can circumvent this constraint by incorporating DNN-based language models (such as RNN or BERT) as an additional layer.

This paper describes our methodological pipeline. Our future work will examine questions related to dimensional interpretations and modality comparisons, including: What are interpretations of the three compressed multimodal dimensions in a qualitatively meaningful way? How do the results from multimodal analyses differ from the results of unimodal analyses? Do lower-dimensional spaces give clues to the nature of mechanisms of multimodal integration?

This work raises the prospect that we could gain a deeper understanding of the dynamics of communication by a mixed methodological approach, incorporating independent sources of data across sensory modalities. Using this pipeline, we can better triangulate how social interactions are supported and accomplished by different modalities, and identify cognitive mechanisms underlying social goals, including establishing interpersonal connections and beyond.

Acknowledgments

We thank research assistants Monica Yingke Mao, Laura Hongye Li, Tracy Jiahe Mao, Brandon Ha, Malia Groth, Joyce Li, Brandon Lustgarten, Ian Lieberman, Kate Kunitz, and Howard Fung for their help with data collection.

References

Amon, M. J., Vrzakova, H., & D'Mello, S. K. (2019). Beyond dyadic coordination: Multimodal behavioral irregularity in triads predicts facets of collaborative problem solving. *Cognitive science, 43*(10), e12787.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.

Bonett, D. G. (2006). "Confidence interval for a coefficient of quartile variation". *Computational Statistics & Data Analysis. 50*(11): 2953–2957

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep Clustering for Unsupervised Learning of Visual Features. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 132-149).

Cramer, A. L., Wu, H.-H., Salamon, J., & Bello, J. P. (2019). Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3852–3856). (ISSN: 2379-190X)

Dahan, A., Noy, L., Hart, Y., Mayo, A., & Alon, U. (2016). Exit from synchrony in joint improvised motion. *PloS one, 11*(10), e0160747.

Dieffenbach, M. C., Gillespie, G. S., Burns, S. M., McCulloh, I. A., Ames, D. L., Dagher, M. M., Falk, E. B., & Lieberman, M. D. (2021). Neural reference groups: A synchrony-based classification approach for predicting attitudes using fNIRS. *Social cognitive and affective neuroscience, 16*(1-2), 117-128.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior.*

Enfield, N. J., & Sidnell, J. (2017). On the concept of action in the study of interaction. *Discourse Studies, 19*(5), 515–535.

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture. *Language, Cognition and Neuroscience, 29*(4), 435–451.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind'in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11-21.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8–11.

Goldstein, A., Zada, Z., Buchnik, E., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*, 369-380.

Gong, X. L., Huth, A. G., Deniz, F., Johnson, K., Gallant, J. L., & Theunissen, F. E. (2023). Phonemic segmentation of narrative speech in human cerebral cortex. *Nature Communications, 14*(1), 4309.

Goodwin, C. (2018). *Co-Operative Action.* Cambridge University Press. (Google-Books-ID: Jg44DwAAQBAJ)

Goodwin, C., & Heritage, J. (1990). Conversation Analysis. *Annual Review of Anthropology, 19*(1), 283–307.

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In *2nd International Conference on Spoken Language Processing (ICSLP 1992)* (pp. 429–432). ISCA.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A., Ward, J., & Gunter, T. (2012). Gesture Facilitates the Syntactic Analysis of Speech. *Frontiers in Psychology, 3,* 74.

Holler, J., Turner, K., & Varcianna, T. (2013). It's on the tip of my fingers: Co-speech gestures during lexical retrieval in different social contexts. *Language and Cognitive Processes, 28*(10), 1509–1518.

Huppert, T. J., Diamond, S. G., Franceschini, M. A., & Boas, D. A. (2009). HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied optics, 48*(10), D280-D298.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453.

Iverson, J., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies, 6*(11-12), 19–40.

Jiang, Y. (2023). Automated Nonverbal Cue Detection in Political-Debate Videos: An Optimized RNN-LSTM Approach. In *International Conference on Human-Computer Interaction* (pp. 32-40). Cham: Springer Nature Switzerland.

Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., & Lu, C. (2012). Neural Synchronization during Face-to- Face Communication. *Journal of Neuroscience, 32*(45), 16064–16069.

Jiang, Y., Dale, R., & Lu, H. (2024). Transformability, generalizability, but limited diffusibility: Comparing global vs. task-specific language representations in deep neural networks. *Cognitive Systems Research*, *83*, 101184.

Kardas, M., Kumar, A., & Epley, N. (2022). Overly shallow?: Miscalibrated expectations create a barrier to deeper conversation. *Journal of Personality and Social Psychology*, *122*(3), 367.

Kelsen, B. A., Sumich, A., Kasabov, N., Liang, S. H., & Wang, G. Y. (2022). What has social neuroscience learned from hyperscanning studies of spoken communication? A systematic review. *Neuroscience & Biobehavioral Reviews, 132*, 1249-1262.

Krijthe, J., van der Maaten, L., & Krijthe, M. J. (2018). Package 'Rtsne'. *R package version 0.13*.

Lieberman, M. D. (2022). Seeing minds, matter, and meaning: The CEEing model of pre-reflective subjective construal. *Psychological Review, 129*(4), 830–872.

Lyu, X., Hueser, M., Hyland, S. L., Zerveas, G., & Raetsch, G. (2018). Improving clinical predictions through unsupervised time series representation learning. *arXiv preprint arXiv:1812.00490*.

McMahon, E., Bonner, M., & Isik, L. (2023). Hierarchical organization of social action features along the lateral visual pathway. *PsyArXiv preprint*.

Miao, G. Q., Dale, R., & Galati, A. (2023). (Mis) align: a simple dynamic framework for modeling interpersonal coordination. *Scientific Reports, 13*(1), 18325.

Michelucci, U. (2022). An introduction to autoencoders. *arXiv preprint arXiv:2201.03898*.

Molavi, B., & Dumont, G. A. (2012). Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiological measurement, 33*(2), 259.

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics, 20*(3), 336–366.

Nguyen, M., Vanderwal, T., & Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage, 184*, 161–170.

Pan, Y., Cheng, X., Zhang, Z., Li, X., & Hu, Y. (2017). Cooperation in lovers: an f NIRS-based hyperscanning study. *Human brain mapping, 38*(2), 831-841.

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications, 9*(1), 332.

Paxton, A., Dale, R., & Richardson, D. C. (2016). Social coordination of verbal and nonverbal behaviours. In P. Passos, K. Davids, & J. Y. Chow (Eds.), *Interpersonal coordination and performance in social systems*. Routledge.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences, 27*(2), 169-190.

Pouw, W., Harrison, S. J., & Dixon, J. A. (2020). Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General, 149*(2), 391–404.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., & Sainath, T. (2019). Deep Learning for Audio Signal Processing. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 206–219.

Raczaszek-Leonardi, J. (2014). Multiple Systems and Multiple Time Scales of Language Dynamics: Coping with Complexity. *Cybernetics & Human Knowing, 21*(1-2), 37–52.

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in Multimodal Interaction: An Integrative Framework. *Cognitive Science, 44*(11).

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., & Marin, S. (2023). The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances, 9*(13), eadf3197.

Rosen, Z. P., & Dale, R. (2023). BERTs of a feather: Studying inter-and intra-group communication via information theory and language models. *Behavior Research Methods*, 1-21.

Sabharwal, S. R., Varlet, M., Breaden, M., Volpe, G., Camurri, A., & Keller, P. E. (2022). huSync-A model and system for the measure of synchronization in small groups: A case study on musical joint action. *IEEE Access, 10*, 92357-92372.

Sasai, S., Homae, F., Watanabe, H., & Taga, G. (2011). Frequency-specific functional connectivity in the brain during resting state revealed by NIRS. *Neuroimage, 56*(1), 252-257.

Schegloff, E. A. (1996). Issues of Relevance for Discourse Analysis: Contingency in Action, Interaction and Co-Participant Context. In E. H. Hovy & D. R. Scott (Eds.), *Computational and Conversational Discourse* (Vol. 151, pp. 3–35). Berlin, Heidelberg: Springer Berlin Heidelberg.

Schegloff, E. A. (2007). *Sequence Organization in Interaction*. Cambridge University Press. (Google-Books-ID: 5XbJRFQ4dhsC)

Schrimpf, M., Blank, I. A., Tuckute, G., Fedorenko, E., & Kanwisher, N. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45),

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena, 404*, 132306.

Sidnell, J. (2006). Coordinating Gesture, Talk, and Gaze in Reenactments. *Research on Language & Social Interaction, 39*(4), 377–409.

Stivers, T. (2021). Is Conversation Built for Two? The Partitioning of Social Interaction. *Research on Language and Social Interaction, 54*(1), 1–19.

Stivers, T., & Sidnell, J. (2005). Introduction: Multimodal interaction. *Semiotica, 2005*(156), 1–20.

Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., & Rush, A. M. (2019). Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 353–363.

Strogatz, S. H., & Stewart, I. (1993). Coupled oscillators and biological synchronization. *Scientific American, 269*(6), 102-109.

Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, *3*(2), 334–340.

Tschacher, W., & Meier, D. (2020). Physiological synchrony in psychotherapy sessions. *Psychotherapy Research, 30*(5), 558-573.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(11), 2579-2605.

Wallot, S., Roepstorff, A., & Mønster, D. (2016). Multidimensional Recurrence Quantification Analysis (MdRQA) for the analysis of multidimensional time-

series: A software implementation in MATLAB and its application to group-level data in joint action. *Frontiers in psychology, 7*, 224211.

Wang, M. Y., Luan, P., Zhang, J., Xiang, Y. T., Niu, H., & Yuan, Z. (2018). Concurrent mapping of brain activation from multiple subjects during social interaction by hyperscanning: a mini-review. *Quantitative imaging in medicine and surgery, 8*(8), 819–837.

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When Language Meets Action: The Neural Integration of Gesture and Speech. *Cerebral Cortex, 17*(10), 2322–2333.37.

Wilson, M. & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12*, 957–968.

Wiltshire, T. J., Philipsen, J. S., Trasmundi, S. B., Jensen, T. W., & Steffensen, S. V. (2020). Interpersonal coordination dynamics in psychotherapy: A systematic review. *Cognitive Therapy and Research, 44*, 752-773.

Yücel, M. (2021). Lühmann A v, Scholkmann F, Gervain J, Dan I, Ayaz H, et al. Best practices for fNIRS publications. *Neurophotonics, 8*(1), 012101.

Zhang, Y., Meng, T., Hou, Y., Pan, Y., & Hu, Y. (2018). Interpersonal brain synchronization associated with working alliance during psychological counseling. *Psychiatry Research: Neuroimaging, 282*, 103-109.

Zima, E., & Bergs, A. (2017). Multimodality and construction grammar. *Linguistics Vanguard, 3*(s1).

Zuo, X.-N., Di Martino, A., Kelly, C., Shehzad, Z. E., Gee, D. G., Klein, D. F., . . . Milham, M. P. (2010). The oscillating brain: complex and reliable. *Neuroimage, 49*(2), 1432-1445.

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130296.