Motion Energy Alignment Analysis in Dialogue

Zohreh Khosrobeigi, Maria Koutsombogera, Carl Vogel
Computational Linguistics Group
Trinity College Dublin, the University of Dublin
Dublin 2, Ireland
{khosrobz, koutsomm, vogel}@tcd.ie

***Abstract*—We study physical alignment in conversations and motion energy (ME) validation. ME is estimated through counts of pixel changes in video regions of interest. ME of interlocutors is quantified with the MEA system, which uses a frame difference algorithm, and it is designed and widely deployed in situations defined to involve a natural dialogue. In the spirit of replication and system validation, we report our efforts to validate motion energy analysis on collaborative dialogues. Within successive windows on ME in video recordings (c. 0.3 seconds duration; 9 frames), and over entire videos without windowing, we correlate the simultaneous ME of paired interlocutors. Then, we compare these values with those that arise where ME for one party is either randomized or reversed. We see expected differences between actual and randomized interlocutors, but not reversals. We demonstrate relations among ME values, ME correlations, and ME validation.**

***Index Terms*—Conversational Synchrony, Alignment, Windowing, Motion Energy Validation, MEA.**

I. Introduction

Given the importance in cognitive infocommunications research, which focuses on developing and analyzing technologies that have the possibility of enhancing human cognitive capabilities [1]–[4], it is important to understand the extent of human abilities and the parameters that influence them. Human communicative behaviours in interaction are among these important parameters [5]–[7]. Others in this field study, for example, abilities with affordances of virtual reality [8], and at face value, immersion in such systems will be enhanced if encountered synthetic avatars respond in a natural way to humans participating in such systems. From a cognitive infocommunications perspective, understanding the dynamics of human dialogue is crucial. Here, we focus on an aspect of such dynamics: the coordination of bodily motions. The ability to measure and quantify alignment through bodily motion has significant implications for cognitive infocommunications and human-computer interaction. It provides a quantitative method to assess the quality of interactions.

We address the measurement of bodily motion as recorded in video within a multi-modal dialogue communications corpus, quantified in terms of pixel change in regions of interest. Our intent is to confirm that widely used motion energy (ME) analysis systems [9], [10] are valid and effective for the materials we are studying. Our approach is two-pronged: we compare ME correlation between dialogue participants with that obtained in matched synthetic data; we compare the measurements of the widely available system with measurements obtained by using open-source video processing libraries.

In support of this work, we adopt the widely deployed MEA system [9], [10], developed for dialogue situations in which there is a natural dialogue leader. The method is based on counts of pixel changes in regions of interest in videos. In the same spirit of research that addresses multi-modal emotion recognition systems attempting to assess the extent to which they agree on high-level and low-level classifications [11], we validate the use of this system through a comparison of results obtained from the videos of actual dialogues and counterparts in which the ME measurements of one dialogue partners are either randomized (without replacement) or reversed (i.e., the value for the first frame becoming the last, etc.). We find that randomization shows expected differences from actual videos, and reversals have properties that require additional scrutiny but are distinguishable from each other. We see this as validating the use of ME in multimodal dialogue analysis.

We describe past research that provides the method adapt to general dialogue (§II). We then describe a multi-modal corpus of interactions used in this research (§III), and the methods used to analyse the dialogue data (§IV-A-§IV-D). Then, we present results (§V) and discuss our interpretation of these results (§VI). The paper concludes (§VII) with observations about the relevance of this work to multi-media computing.

II. Background

During human dialogue interaction, interlocutors use diverse channels to communicate, including language, gestures, head and body movements, and they often coordinate these multimodal actions. The study of behavioral synchronisation, i.e., the similarity and interlocking of behavioural patterns among speakers that interact with each other, has been seen as one of the ways to approach the dynamics of the conversation [12]. Individuals repeat aspects of each other's communicative behaviour to show their involvement and understanding, to pursue their communicative goals, contribute to the efficient exchange of information, and overall facilitate harmonious interactions [13], [14]. This interpersonal coordination includes alignments that take place in daily face-to-face interaction and happen naturally, even perhaps without consciousness of it [15], and coordination is referred to as a characteristic of every interpersonal interaction [16].

The interlocutor behaviours may grow to be in tune with each other at different levels, including linguistic alignment [17], gestural and lexical alignment [18], facial expressions synchrony [19] and body movement coordination [12], [20].

Some of these behaviours may be discrete events (i.e., gaze, laughter), while others are a series of events, i.e., a series of body movements. The units of these behaviour alignments are cross-participant paired behaviour that indicates the behaviours match or are the same in one or more dimensions [18]. To clarify the relation between any pair of behaviours, [18] has characterized five keys of alignment dimensions, out of which we explore two: time and similarity. These five dimensions can occur between paired behaviour of co-speakers or paired behaviours of the speaker alone. Time indicates defining a specific temporal lag between aligned pairs of behaviours. The temporal lag between paired behaviours may vary from zero, which shows simultaneity between paired behaviours, to a delay in time units, such as a millisecond, second, minute, or hour. Also, (dis)similarity refers to paired behaviour being more or less similar in form. Because of highlighting temporal aspects such as rhythm, the meshing of nonverbal behaviours, and simultaneous movement as quantitative characteristics, [21] classifies this type of non-verbal synchrony as movement synchrony, considering the exact time and coordinate movement between interlocutors.

Out of the various methods that capture interpersonal coordination, here we work with Motion Energy Analysis (MEA) [9], a video-based tracking method in the context of nonverbal synchrony research. MEA has been extensively deployed in studies that quantify non-verbal synchrony in the context of psychotherapy (see [20], [22]–[25], among others). It has also been used in a number of works researching interpersonal synchronisation, including the contribution of rhythmic features of movement to bonding between individuals [26], measuring coordination in the workplace in mentor-mentee dyads [27] and examining gender differences in synchrony in dyadic unstructured conversations with same-gender strangers [28]. In addition to its reported effectiveness in measuring body movement coordination, MEA has been considered a preferable analysis method for videos of conversations with a fixed camera and set lighting conditions (as is the case with our data), compared to pose estimation methods [29].

Therefore, we opt for using this technique to analyse our data as we consider it an appropriate method to profile our dialogues in terms of body movement synchronisation and to enable the interpretation of their dynamics.

III. Dataset

Our study uses the MULTISIMO dataset [30], a multi-modal corpus consisting of 18 videos of three-party dialogue sessions in English.[1] In each session, two interlocutors collaborate to answer three questions provided by a moderator, and then rank answers in terms of popularity, following the responses of external groups. A number of 36 interlocutors were randomly paired with each other for each session, and a moderator helped them to find answers. In all sessions, one of the interlocutors is seated on the left side of the moderator, and the other interlocutor on the right. Each video includes one person on the screen; the setup of the camera, lights, and background is fixed, and only the interlocutor moves. The duration of all videos is around 4 hours. For this study we used synchronized high-quality zoomed front views of each interlocutor.

IV. Method

We compute ME for each interlocutor in each dialogue in MULTISIMO (cf. §IV-A). Then, we compute correlations between paired interlocutors using the rolling window method (cf. §IV-B) over the whole video as a measure of alignment. Each window has a length of 0.3 seconds (9 frames). Next, we label each windowed correlation based on the significance of their p-values and the sign of their correlation values (cf. §IV-D). We imagine that in actual data, communication dynamics will sometimes yield positive correlations (when interlocutors act together), sometimes negative correlations (when the activity of one is matched by decreasing activity of the other, such as in turn-taking), and sometimes no correlation. We apply three labels: "Open", "Neg", and "Pos". We form bi-grams over the successive labeled windows, studying bi-gram distributions to gain insights into the sequence of alignment and non-alignment in actual and synthetic data.

This approach is applied to various data types. We compare actual ME values with manipulated data, where ME values of one participant in each dialogue are reversed (the last value becoming first, etc.) or shuffled. We expect the bi-gram distribution of labeled window correlations to have significant interaction with different data types (actual ME vs. reversal ME). Furthermore, we examine the correlation of paired interlocutors' ME throughout each dialogue. We expect their correlation to become non-significant in the random and reversal data. Separately, another method is used for ME extraction to examine ME extraction method used in this study. ME of interlocutors are extracted using Single Shot MultiBox Detector (SSD) [31], OpenCV [32], and frame difference algorithm (FDA). Then, the results are compared to ME extracted by the MEA system [9], [10].

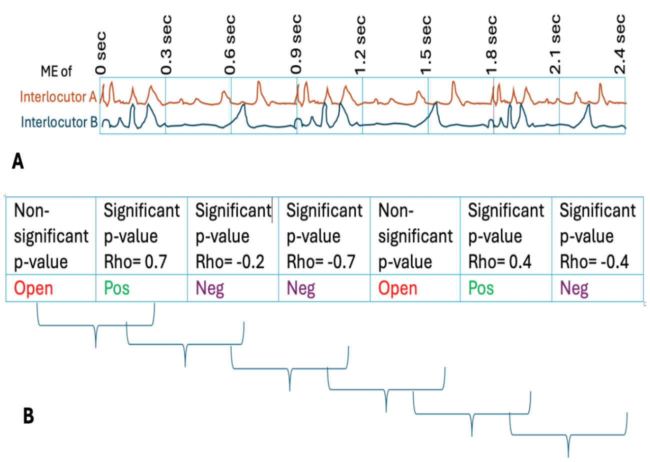

Fig. 1. Section A shows rolling window on ME of paired interlocutors. Section B demonstrates labeled windowed correlation and defining bi-grams of these windows

[1] See multisimo.eu, last verified May 2024.

A. Motion Energy Computing

A video consists of a series of still images called frames. A frame includes many pixels; the pixels' colours change when transitioning to the next frame in case a change happens, e.g., an object movement. The difference between the colours of pixels in two successive frames is called ME [9]. ME is quantified by the number and amount of pixel changes summed between each frame and its prior frame. The first frame has an ME of zero since no prior frame exists. This method is one of the frame-difference algorithms that compute changes, although they do not capture the direction of movement [9]. The colours of pixels may be affected by background colour, masking, light sources, noise, etc. Hence, to apply frame-difference methods, some conditions should be controlled; the camera, background, and light should be stable since light or background changes can change pixel colour and can thus be mistakenly considered as a movement. Thus, these conditions should be constant in the input video. Also, objects should have clear boundaries and not be masked, as frame-difference algorithms recognize pixel changes and not objects. In frame-difference algorithms, the area of target objects defined to calculate their movement is called the region of interest (ROI).

In the MULTISIMO data that we use for our research, the camera position, light, and environment conditions are constant, and the dialogue interlocutors may move while they are speaking or listening. We opted to use the MEA tool [9], [10], [20], a tool that computes ME of video frames using a frame-difference algorithm and deploying a user interface. At first, after opening the video within the tool, the whole body of an interlocutor is defined as the ROI on the MEA system. Then, the system computes the motion of each frame and provides the related measurements as an output series in a text file. The MEA tool has a user-friendly interface, allowing users to conveniently define the ROI using a mouse, with no additional coding or other time-consuming processes required.

Furthermore, to validate ME extracted by the MEA system, we use an additional approach to compute ME. The approach involves the application of SSD [31], which is designed for object detection in videos and images. Using OpenCV [32] and SSD, an interlocutor is detected as a person in the video. Then, a box is drawn around the interlocutor as ROI. Finally, ME of pixels in ROI is calculated through the FDA, and a series of ME per frame is generated for each interlocutor.

B. Rolling Window

Dynamics of conversation, in general, suggest that the relationship between interlocutors that are leading, following, or are aligned with each other, may change during the course of the dialogue. In human-human interaction analysis, the method of video windowing is employed to compute the correlation between interlocutors within conversation segments. In this sense, rolling a window on ME series can capture the dynamics in the conversation, which may change as the dialogue proceeds. A window is a vector derived from sequential measurements taken from a time series. We use the windowing method to determine if any correlation (alignment) exists between ME of interlocutors across dialogue segments. This alignment at lag-zero indicates interlocutors' simultaneous movements.[2]

In this research, a window involves a series of ME values. To apply the method, first, a window with a constant length of 0.3 seconds is defined. Then, the window is rolled on ME of paired interlocutors from second 0 of the conversation to 0.3 seconds later, thus covering time series from seconds 0-0.3. Then, Spearman correlation between ME of the interlocutors covered by the window is computed. Next, the window transitions to the next point, maintaining a constant length. The start point of the subsequent window is set to the endpoint of the current window, and the process iterates until the rolling window reaches the end of the conversation.

Windowing presents various parameters: window length and window step size or overlap. Yet, given that these parameters may vary across dialogues due to factors like context or interlocutors, they should be dynamically set based on the specific data under investigation. In our study, the window length was set to 0.3 seconds. To avoid repetition, we opted not to use overlaps, ensuring no shared time between the window and the subsequent window. The reason for the choice of the 0.3 seconds window length is to take an intermediate view between 30 and 3000 milliseconds, considered as minimal duration of perception and cognitive integration of temporal information [33]. Additionally, we employ Spearman-rank correlation, which is a non-parametric method as the ME measurements are non-normally distributed for this data.

C. Data Construction

We use a few data constructions. The first type is the Actual data, ME values from the recorded videos, without modification. The second type is Reversed data, where the ME frames of the right-side interlocutors are temporally reversed. This approach allows us to create synthetic conversation data with lengths matching the actual data while also having time-directional influences. The third type of data is yet another synthetic version, Shuffled: the ME of the right-side interlocutor is shuffled ten times, and the mean of ten shuffled ME measurements is calculated as ME of the frame. Shuffling breaks temporal dependencies, thereby serving as a null model.

D. Successive window n-grams

Applying the windowing method results in a series of windows with p-value and correlation coefficients. We apply discrete labels to each window based on the significance and the direction of its correlation to determine the distribution of bi-grams of these labels. This causes observing and analyzing patterns in the sequence of window states rather than just considering individual windows in isolation. Therefore, we can explore changes in the correlation between interlocutors over time. The bi-grams enable a view of the data related to successive windows and not just the frames within a window. This can provide insights into the dynamics and transitions of

[2] In a larger body of work not reported here, we examine distinct lags on the series of ME to estimate from times when interlocutors correlate which interlocutor may be thought to "lead".

TABLE I

SPEARMAN CORRELATION RESULTS BETWEEN ENTIRE ME MEASUREMENTS FOR DIFFERENT DATA AND METHODS, AND THEIR COMPARISON. IN THE CATEGORY OF THE MEA, COLUMNS ACTUAL-RHO, SHUFFLED-RHO, AND REVERSED-RHO SHOW CORRELATIONS OF ME EXTRACTED BY THE MEA SYSTEM. IN THE CATEGORY OF SSD-FDA, COLUMN ACTUAL-RHO INDICATES THE SAME, BUT ME IS EXTRACTED BY SSD-FDA. THE LAST TWO COLUMNS SHOW THE CORRELATION BETWEEN ME OF THE LEFT SIDE INTERLOCUTOR EXTRACTED BY THE MEA AND SSD-FDA; THE SAME FOR THE RIGHT SIDE INTERLOCUTOR. NS SHOWS THE P-VALUE IS NOT SIGNIFICANT, SO THE CORRELATION IS NOT REPORTED.

	MEA			**SSD-FDA**	**MEA and SSD-FDA**	
Session	**Actual-Rho**	**Shuffled-Rho**	**Reversed-Rho**	**Actual-Rho**	**Left Side Interlocutor**	**Right Side Interlocutor**
S02	0.3037	NS	-0.10188	0.3949	0.82206	0.8727
S03	0.17849	NS	-0.0308	0.15896	0.90086	0.87728
S04	0.214	NS	NS	0.16176	0.9323	0.8215
S05	0.061345	NS	-0.0468	0.03132	0.87164	0.87514
S07	0.21809	NS	-0.04477	0.1676	0.91719	0.8823
S08	0.23644	NS	NS	0.29423	0.87365	0.91788
S09	0.23988	NS	NS	0.25365	0.89866	0.86
S10	0.19234	NS	0.04526	0.121	0.72412	0.92637
S11	0.12786	NS	NS	0.093	0.912395	0.8
S13	0.25729	NS	NS	0.42964	0.805188	0.9129
S14	0.52032	NS	-0.07944	0.43	0.88859	0.94836
S17	0.35154	NS	NS	0.26998	0.9144	0.92249
S18	0.05019	NS	NS	0.050817	0.90552	0.86967
S19	0.38266	NS	NS	0.40138	0.818179	0.92707
S20	0.25607	NS	0.0597	0.1952	0.9368	0.944527
S21	0.34278	NS	-0.0798	0.42477	0.7661	0.91409
S22	0.030847	NS	-0.1035	-0.0059	0.90835	0.90999
S23	0.28411	NS	NS	0.22425	0.88053	0.9268

states within the data, which might be particularly useful in understanding temporal relationships and dependencies.

In the process of applying discrete labels, if the p-value of a window is non-significant ($p \geq 0.05$), it is labeled as "Open," if the p-value of the window is significant and its correlation is positive ($\rho \geq 0, p < 0.05$), it is labeled as "Pos," and if the p-value is significant and its correlation is negative ($\rho < 0, p < 0.05$), it is labeled as "Neg." Then, the bi-grams of these discrete labels are formed. We apply this approach to our different data constructions, i.e., Actual data and manipulated data (the correlation details derived from Reversed and Shuffled data). Finally, we have a series of bi-grams of labeled windows. Next, the bi-gram distributions for each of these orderings are compared using a contingency table test. Comparing distributions across different data manipulations, such as Reversed and Shuffled, helps to assess the robustness and characteristics of the observed patterns. This can help distinguish actual patterns from those arising due to random fluctuations or inherent data structure.

V. EXPERIMENTS

A. *Meaningfulness of ME correlations*

ME extraction methods described in §IV-A are employed to extract ME of each interlocutor to explore the correlation between pairs of interlocutors during the entire dialogue for different data constructions. Paired interlocutors' ME extracted by the MEA system are considered, and then the correlation between them is computed. This process is applied to Actual, Reversed, and Shuffled data constructions. In addition, to validate the MEA system and assess its outputs, its performance is compared against the SSD-FDA method (cf. §IV-A). For this purpose, the correlation between paired interlocutors' ME extracted by SSD-FDA is computed. In another experiment, we examine the ME of the left-side interlocutor, as extracted by the MEA system, as well as by the SSD-FDA. We then calculate the correlation between the ME extracted by MEA and the ME extracted by SSD-FDA for the same left-side interlocutor. The same process is followed for the right-side interlocutor.

Table I shows the results for different methods and data. In the case where ME is extracted by the MEA, the column Actual-Rho illustrates that the interlocutors' correlations within Actual data are significant, yet some are weak. Column Shuffled-Rho shows none of the correlations being significant (NS refers to "not significant"). In the column of Reversed-Rho (reversed data), some dialogues have significant but very weak correlations, some have negative correlations, and some have non-significant correlations. In the case in which ME is extracted by SSD-FDA, the Actual-Rho column shows that all correlations are significant, yet some are weak. In the category of MEA and SSD-FDA, the columns for Left Side and Right Side Interlocutors indicate a significant and strong correlation between ME extracted by the two methods.

B. *Bi-gram label distributions*

By applying the method detailed in §IV-D, we get a series of bi-grams of labeled windows, which are derived from successive window correlation values – a series of bi-grams of successive correlation situations at different time spans. Here, we report on the interaction of comparison and inspection of residuals for bi-grams labels. A contingency table of occurrence of bi-grams labels is constructed per dialogue for Actual vs. Reversed data and also for Actual vs. Shuffled data. The contingency table of Actual vs. Reversed data includes counts of the bi-grams distribution of successive window correlation label driven from Actual data (actual ME measurements) and Reversed (reversed ME of the right-side player). The contingency table of Actual vs. Shuffled includes counts of the bi-grams distribution driven from Actual data and Shuffled (randomized ME of right-side player). Then, Chi-squared test is applied to each comparison table to assess statistical significance interaction between bi-grams of Actual data and constructed data. The interaction was significant for nine dialogues out of 18 on Actual vs. Reversed. Regarding Actual vs. Shuffled, the interaction is significant for all dialogues.

We aggregated contingency tables from all dialogues to get the count of all bi-grams categories and came up with only one table to investigate overall. One table for Actual vs. Reversed, and one table for Actual vs. Shuffled. Applying Chi-squared test on the comparison table of Actual vs. Reversed shows significant interaction between bi-grams categories of Actual data and Reversed data ($\chi^2 = 188.9, df = 8, p < 2.2e-16$). Analysis of the residuals of interaction reveals observations of "Open Open" (successive windows without significant correlation) are more than would be expected for Reversed if there were no interactions between the "Open Open" bi-gram category (null hypothesis were true, and variables were independent). It is the opposite for Actual; observations of "Open Open" are less than would be expected if there were no interactions between the "Open Open" category. Similarly, observations of the bi-grams of "Open Pos," "Pos Open" and "Pos Pos" are in significant abundance for Actual data than would be expected with no interaction between each of these bi-grams; and in a significant dearth for the Reversed data than would be expected with no interaction for these bi-grams categories. Furthermore, Chi-squared test shows Actual vs. Shuffled interaction is also significant ($\chi^2 = 3419.7, df = 8, p < 2.2e-16$). Inspecting residuals reveals that all bi-grams categories involving at least one significant correlation (i.e., any combination that includes the labels "Pos" or "Neg") are in significant abundance rather than would be expected no interaction in these categories of Actual data; dearth rather than would be expected to have no interaction between these bi-gram categories of Shuffled data. The case of successive windows without significant correlation in either ("Open Open") is in significant dearth for Actual and significant abundance for Shuffled. Table II shows the residuals.

TABLE II
RESIDUALS OF ACTUAL VS. REVERSED COMPARISON AND ALSO, ACTUAL VS. SHUFFLED COMPARISON. BONFERRONI ADJUSTMENT IS APPLIED BY DIVIDING $\alpha = 0.05$ BY COUNTS OF CATEGORIES WHICH IS 18, $\alpha' = 0.0027$ AND THE CRITICAL VALUE FOR $N(0,1)_{\alpha'} = 2.99$.

	Residual		**Residual**	
Bi-grams	**Actual**	**Reversed**	**Actual**	**Shuffled**
Neg Neg	-0.66	0.66	14.41	-14.41
Neg Open	-1.02	1.02	16.15	-16.15
Neg Pos	2.40	-2.40	14.44	-14.44
Open Neg	0.71	-0.71	16.23	-16.23
Open Open	-9.43	9.43	-55.98	55.98
Open Pos	6.25	-6.25	26.31	-26.31
Pos Neg	1.66	-1.66	14.08	-14.08
Pos Open	6.63	-6.63	26.38	-26.38
Pos Pos	9.01	-9.01	21.78	-21.78

VI. DISCUSSION

The results reported in §V-A demonstrate significant correlations of ME of paired interlocutors throughout the dialogue for the Actual, and non-significant for the manipulated ME data (see Table I). Moreover, the results reported in §V-B indicate that, in the analysis of bi-grams of correlation labels for successive windows, there are clear and significant differences between Actual and Reversal (ME measurements for one party reversed) and Shuffled (ME for one party shuffled) data. Actual ME data shows significantly fewer instances of successive windows without any significant correlation ("Open Open") than would be expected if there were no interaction between the bi-gram counts for "Open Open;" it is the opposite for constructed data in which "Open Open" is more in Reversed and Shuffled data (see Table II): this means that extended periods of interlocutor non-alignment are less likely in actual dialogue than synthetic counterparts. The significant abundance of observations of "Pos Pos" in actual data (relative to what one expect if the actual vs. treated data made no difference) indicates that successive windows of interlocutor alignment are more likely in actual data than either synthesized sort. Similarly, "Neg Neg" (which encodes successive windows of interlocutors being in complementary alignment) is also more likely in actual data than shuffled, but the difference with respect to the reversal is not significant. More generally, window sequences involving at least one "Pos" are always in greater abundance for actual data than for treatments (and significantly so, except for two cases). These patterns overwhelmingly indicate that in this data, alternations and continuations in positive ME correlation between interlocutors separate actual and synthetic dialogues.

These results imply that at the dialogue level, there is a clear and quantifiable difference in ME dynamics between actual dialogue and manipulated datasets. This suggests that the ME values reported are meaningful, a claim that would be refuted if actual and synthetic ME sequences are indistinguishable. ME measurements reflect conversational dynamics, in particular, interlocutor alignment in dialogue. The significant correlations observed in actual dialogues and the insignificant correlations in manipulated datasets support the hypothesis that ME measurements can be used to measure alignment in conversation. In addition, the differences between dynamics transition in actual and manipulated data provided by bi-grams, refer to patterns of ME alignment that are different for actual and manipulated data and can help to distinguish an actual interaction from anartificial one.

The strong relationships between ME extracted by the MEA system and ME extracted by SSD-FDA provides further evidence of reliability in ME measurement. The research supports motion energy as a valid measurement to analyze dialogue by comparing actual data with manipulated data. However, on reversal ME, differences are not as expected to actual ME. One reason may be that ME measurements belong to the dialogue, and some parts still correlate while the time of their occurrence is different. Thus, the correlation is independent of time. This is an aspect that remains to be investigated.

VII. Conclusion

This research studied ME quantified by the MEA system and alignment by applying the windowing method. Using windowing and bi-grams methods on actual and manipulated data, we demonstrated the validity of ME in dialogue analysis, and in particular, that ME correlation is an approximate measure of interlocutors' alignment. Furthermore, the results validate MEA, the ME extraction system, by comparing its results to MEs extracted by the SSD-FDA. As reproducibility and replication are important in science, we see this as a useful contribution. We have used the MEA system (developed by others) in our own work published elsewhere, and aimed here to confirm the validity of the measurements that we and others use to ground inferences about the dynamics of dialogue. In general, we wish to find correspondences between ME measurements for interlocutors and independent assessments of successful communication within the dialogues analyzed.

Acknowledgments

This work was conducted with the financial support of Science Foundation Ireland under Grant No. 18/CRT/6223.

References

[1] P. Baranyi and A. Csapo, "Cognitive infocommunications: Coginfocom," in *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*, 2010, pp. 141–146.

[2] A. Csapo and P. Baranyi, "Coginfocom channels and related definitions revisited," in *IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY)*, 2012, pp. 73–78.

[3] P. Baranyi, A. Csapo, and P. Varlaki, "An overview of research trends in CogInfoCom," in *IEEE 18th International Conference on Intelligent Engineering Systems*, ser. INES, 2014, pp. 181–186.

[4] A. Esposito, G. Cordasco, C. Vogel, and P. Baranyi, "Cognitive infocommunications," *Frontiers in Computer Science*, vol. 5, 2023.

[5] F. Bonin, C. Vogel, and N. Campbell, "Social sequence analysis: Temporal sequences in interactional conversations," in *5th IEEE International Conference onf Cognitive Infocommunications*, e. a. Baranyi, Péter, Ed., 2014, pp. 403–406.

[6] C. Vogel and A. Esposito, "Linguistic and behavior interaction analysis within cognitive infocommunications," in *10th IEEE Conference on Cognitive Infocommunications*, 2019, pp. 47–52.

[7] ——, "Interaction analysis and cognitive infocommunications," *Infocommunications Journal*, vol. 12, no. 1, pp. 2–9, 2020.

[8] Á. B. Csapó, A. Sudár, B. Berki, B. Gergely, B. Berényi, and C. Czabán, "A test space for virtual rotation measurement in maxwhere," in *2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2020, pp. 585–586.

[9] F. T. Ramseyer, "Motion energy analysis (MEA): A primer on the assessment of motion from video." *Journal of counseling psychology*, vol. 67, no. 4, p. 536–549, 2020.

[10] J. R. Kleinbub and F. T. Ramseyer, "rMEA: An R package to assess nonverbal synchronization in motion energy analysis time-series," *Psychotherapy research*, vol. 31, no. 6, pp. 817–830, 2021.

[11] C. Vogel and K. Ahmad, "Agreement and disagreement between major emotion recognition s ystems," *Knowledge Based Systems*, vol. 276, p. 110759, 2023.

[12] R. Dale, G. Bryant, J. Manson, and M. Gervais, "Body synchrony in triadic interaction," *Royal Society Open Science*, vol. 7, p. 200095, 09 2020.

[13] F. J. Bernieri and R. Rosenthal, "Interpersonal coordination: Behavior matching and interactional synchrony." *Fundamentals of Nonverbal Behavior. Studies in Emotion and Social Interaction*, 1991.

[14] I. M. Vicaria and L. Dickens, "Meta-analyses of the intra- and interpersonal outcomes of interpersonal coordination," *Journal of Nonverbal Behavior*, vol. 40, no. 4, pp. 335–361, 2016.

[15] A. Kendon, R. M. Harris, and M. R. Key, *Organization of behavior in face-to-face interaction*. Walter de Gruyter, 2011.

[16] V. L. Manusov, *The sourcebook of nonverbal measures: Going beyond words*. Psychology Press, 2014.

[17] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[18] M. Rasenberg, A. Özyürek, and M. Dingemanse, "Alignment in multimodal interaction: An integrative framework," *Cognitive science*, vol. 44, no. 11, p. e12911, 2020.

[19] D. McIntosh, "Sponteous facial mimicry, liking and emotional contagion," *Polish Psychological Bulletin*, vol. 37, no. 1, p. 31–42, 2006.

[20] F. Ramseyer and W. Tschacher, "Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome." *Journal of consulting and clinical psychology*, vol. 79, no. 3, pp. 284–295, 2011.

[21] ——, "Synchrony in dyadic psychotherapy sessions," *Simultaneity: Temporal structures and observer perspectives*, pp. 329–347, 2008.

[22] D. Schoenherr, J. Paulick, S. Worrack, B. M. Strauss, J. A. Rubel, B. Schwartz, A.-K. Deisenhofer, W. Lutz, U. Stangier, and U. Altmann, "Quantification of nonverbal synchrony using linear time series analysis methods: Lack of convergent validity and evidence for facets of synchrony," *Behavior Research Methods*, vol. 51, no. 1, pp. 361–383, 2019.

[23] D. Dean, A. Samson, R. Newberry, and V. Mittal, "Motion energy analysis reveals altered body movement in youth at risk for psychosis," *Schizophrenia Research*, vol. 200, pp. 35–41, 2018.

[24] J. Paulick, A.-K. Deisenhofer, F. Ramseyer, W. Tschacher, K. Poster, J. Rubel, and W. Lutz, "Nonverbal synchrony: A new approach to better understand psychotherapeutic processes and drop-out," *Journal of Psychotherapy Integration*, vol. 28, 12 2017.

[25] Z. Kupper, F. Ramseyer, H. Hoffmann, and W. Tschacher, "Nonverbal synchrony in social interactions of patients with schizophrenia indicates socio-communicative deficits," *PloS one*, vol. 10, p. e0145882, 12 2015.

[26] K. Fujiwara, M. Kimura, and I. Daibo, "Rhythmic features of movement synchrony for bonding individuals in dyadic interaction," *Journal of Nonverbal Behavior*, vol. 44, no. 1, pp. 173–193, 2020.

[27] A. Vodneva, G. Oreshina, T. Kustova, I. Tkachenko, M. Tcepelevich, and E. Grigorenko, "Interpersonal synchrony in mentor-mentee dyads: An analysis of nonverbal synchrony and trait-empathy," *Journal of Modern Foreign Psychology*, vol. 13, pp. 47–57, 04 2024.

[28] K. Fujiwara, M. Kimura, and I. Daibo, "Gender differences in synchrony: Females in sync during unstructured dyadic conversation," *European Journal of Social Psychology*, vol. 49, no. 5, pp. 1042–1054, 2019.

[29] K. Fujiwara and K. Yokomitsu, "Video-based tracking approach for nonverbal synchrony: A comparison of motion energy analysis and openpose," *Behavior Research Methods*, vol. 53, no. 6, pp. 2700–2711, 2021.

[30] M. Koutsombogera and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, e. a. Calzolari, Nicoletta, Ed. France: European Language Resources Association (ELRA), 2018, pp. 2945–2951.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[32] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[33] T. Smith, "An attentional theory of continuity editing," Ph.D. dissertation, University of Edinburgh, 2005.