Novel features for capturing cooccurrence behavior in dyadic collaborative problem solving tasks

Vikram Ramanarayanan
Educational Testing Service R&D
90 New Montgomery St, #1500
San Francisco, CA
vramanarayanan@ets.org

Saad Khan
Educational Testing Service R&D
600 Rosedale Road
Princeton, NJ
skhan002@ets.org

ABSTRACT

Temporal patterns of verbal and non-verbal behavior in human interactions reveal our attitudes toward each other and can have an outsized impact on outcome of activities such as negotiations, coordination and team work. However, not much work has been done towards automatically analyzing the temporal aspects of such behavior during collaborative tasks. In this paper, we take an initial step toward this by proposing a novel feature that captures convergent or divergent behavior between dyads involved in a collaborative problem solving task. This feature, dubbed histograms of cooccurrence, capture how often different prototypical behavioral states exhibited by one person co-occur with those exhibited by his/her partner over different temporal lags. We show that not only does such a feature bring out the differences between dyads and non-dyads, but is also interpretable in that it tells us which behavioral states are most likely to occur in true dyads as opposed to nominal or artificial dyads.

Keywords

collaborative problem solving, multimodal analytics, machine learning, entrainment, engagement, mirroring, educational data mining

1. INTRODUCTION

Research shows that complex interactive activities such as team work and collaboration are more effective when participants are not only engaged in the task but also exhibit behaviors that facilitate interaction [17]. Successful collaboration is often manifested in what is known as "entrainment" or convergence between the participants of such collaboration. In spoken face-to-face communication this may include synchronization in speaking rate or intonation patterns [12] of the collaborators as well as non-linguistic aspects such as participants mirroring each other's gestures and other behavioural patterns [8, 5]. The participants of text-based collaboration often converge in their choice of words or communication style [7]). The degree of entrainment has been found to be positively correlated with the overall success of collaboration as well as predictive of the polarity of participants' attitudes (see [9] for review). Social psychologists have further postulated that behavioral mirroring presents an evolutionary component associated with the development of social bonds and empathy through cooperation [8]. In the educational context, entrainment between collaborators or between student and the tutoring system has been shown to be correlated with learning gain and improvement in students' performance (cf. for example [18]). In addition, studies have explored its impact on interpersonal skills, coordinated activity, negotiations, and how individuals influence the behaviors of others [2] [5]. Other studies in the literature have also demonstrated the importance of understanding affect and gaze dynamics during such collaborative interactions in learning environments [16, 1, 4].

Recent research has explored the impact of behavioral synchronicity of cognitive and non-cognitive behavior in interactive collaborative activity [3]. In particular, Luna Bazaldua et al. demonstrated a statistically significant synchronicity of cognitive and non-cognitive behavior between dyads engaged in online collaborative activity [3]. However, in their study participants were not able to see each other and only interacted over a text-based chat interface. This is an important point to note since the ability to converse face-to-face can significantly impact the nature of the dyadic interaction. Therefore, in this paper we focus on behavioral patterns of emotional expressions between dyads during face-to-face conversation through a video conferencing system. Our hypothesis is that dyads engaged in face-to-face collaborative activity demonstrate a significantly different pattern of behavior as opposed to nominal dyads who are artificially paired up with each other. Notation-wise, we use the term nominal dyad or artificial dyad interchangeably to mean two subjects whose data are analyzed as if they were interacting dyadically, but were actually not.

Explicitly modeling temporal information in such dyadic interaction data is important because each person's emotional state or behavior need not stay constant over the course of the interaction – they could get fatigued over time, or be more nervous at the very beginning (resulting in repetitive, cyclic fidgeting behavior), but gradually settle into a comfort zone later, as they get more familiar with the task and each other. For similar reasons their body language and emotional state can also fluctuate over the time series. However, current feature extraction approaches that aggregate information across time do not explicitly model temporal cooccurrence patterns; consider for instance that one person's emotional state – joy – generally follows his interlocutor's emotional state – say neutral – in a definitive pattern during certain parts of the interaction. Capturing such patterns might help us (i) explicitly understand the predictive power of different features (such as the occurrence of a given pair of emotions) in temporal context (such as how often did the emotional state of one person in the dyad oc-

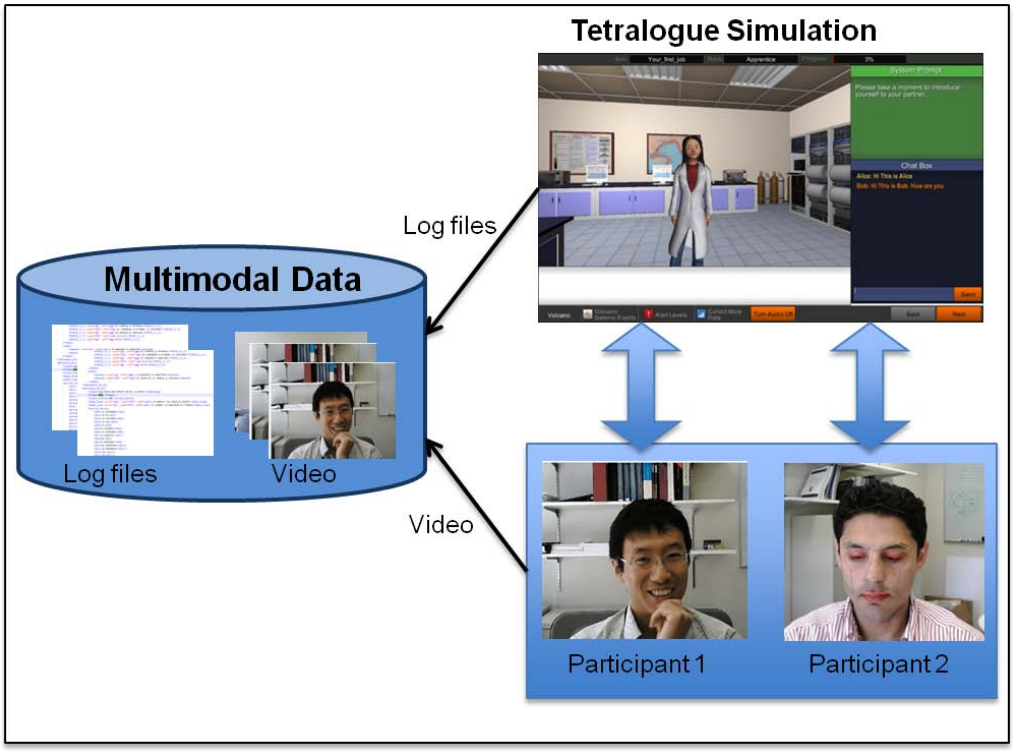

Figure 1: Schematic illustration of the Tetralogue collaborative problem solving simulation. The platform allows for the capture of multimodal data streams including video, audio, text and action log files while participants engage in a collaborative activity.

cur given the previous occurrence of another emotional state of the other person in the dyad), thus allowing us to (ii) obtain features that are more interpretable on visual inspection. We would like to take an initial stab at bridging this gap in this paper. Specifically, we propose to adapt a feature based on histograms of cooccurrences [19, 20, 15] that was developed earlier for analyzing a single time-series (say, from one person), and extend it to the case of dyads. The feature models how different "template" emotional states of one person in a dyad co-occur within different time lags of a "template" emotional states of the other person in the dyad over time. Such a feature explicitly takes into account the temporal evolution of emotional states in different interaction contexts. This feature has been previously shown to perform well for automated multimodal presentation scoring [14], on phone classification tasks [15] as well as for unsupervised pattern discovery [19, 20].

2. DATA

2.1 The Tetralogue CPS Platform

We used an online collaborative research environment developed in-house – the Tetralogue [11, 21, 3]. This platform includes both traditional assessment components, such as a set of multiple-choice items on general science topics, a simulation based assessment, a personality test, and a set of background questionnaire. The simulation task is on geology topics. The simulation-based task was developed as a task for individual test takers who will interact with two avatars and as a collaborative task that requires the collaboration among two human participants and two avatars in order to solve geology problems. The participants, who may be in different locations, interact through an online chat box and system help requests (selecting to view educational videos on the subject matter). The main avatar, Dr. Garcia, introduces information on volcanoes, facilitates the simulation, and requires the participants to answer a set of individual and group questions and tasks. A second avatar, Art, takes the role of another student who shows his own answers to the questions posed by Dr. Garcia, in order to contrast his information with that produced by the dyad.

2.2 Data collection

Twenty-six subjects participated in this study and were paired in dyads using random selection. At the time of the study, the participants were graduate students or recently graduated from different universities across the United States and Europe attended the Summer Internship Program at Educational Testing Service. Information about the study was provided to each participant individually and consent forms were obtained from them. In addition, each participant filled a brief checklist after the session in order to obtain their opinions about the experiment.

As mentioned earlier, the dyads were able to interact with each other over a videoconferencing interface. The video from each session was captured resulting in thirteen pairs of time aligned video recordings. Each dyad reviewed and responded to the same educational material and academic questions. The duration of dyad sessions varies from 36 minutes to 68 minutes, with an average length of about 52 minutes. This resulted in a total of 1,356 minutes of video data, which formed our core evidentiary dataset and was analyzed utilizing automated facial expression analysis as described in the following section.

2.3 Video Processing

Facial expression analysis of the video data was performed using the FACET SDK, a commercial version of the Computer Expression Recognition Toolbox (CERT) [10]. This tool recognizes fine-grained facial features, or facial action units (AUs), described in the

Facial Action Coding System [6]. FACET detects human faces in a video frame, locates and tracks facial features and uses support vector machine based classifiers to output frame-by-frame detection probabilities of a set of facial expressions: anger, joy, contempt, surprise, etc.

3. HISTOGRAMS OF COOCCURRENCE (HOC) FEATURES

Recall that one of the goals of this work is to capture the temporal evolution of the emotional states (as captured by the FACET features) of each person in *true* dyads and understand how these emotional states converge or diverge relative to *nominal* or *artificial* dyads. The motivation is that explicitly examining and modeling the evolution of each of these time series will result in richer features as opposed to time-aggregated features. With this in mind, we elucidate below a general methodology to compute such a feature called histograms of cooccurrences (or HoC) that can be applied to any multivariate time-series – in this paper we compute this feature for the Emotion data stream. The advantage of this feature vector over conventional time-aggregated one is that it explicitly encapsulates information regarding temporal co-occurrence patterns; so, for example, it would model how often a certain prototypical emotional state (such as joy) follows a second prototypical emotional state (say, neutral) in a definitive pattern during different parts of the collaborative task.

So that being said, the idea behind the histogram of cooccurrence (HoC) feature is to count the number of times different prototypical behavioral/emotional states of one person in the dyad (represented by A in Figure 2) co-occur with those of the other person in the dyad (B in Figure 2) *at different time lags* over the course of the time series. As to what these prototypical behavioral/emotional states are – while this is an interesting research question in itself, for the purposes of this paper we use cluster centroids derived from simple K-means clustering on the space of emotional states (or FACET features) as prototypical states. Note that we performed this clustering on the FACET features obtained from *all* speakers in the dataset [1]. We experimented with different cluster sizes $(8, 16, 32)$ and found that 16 clusters did a good job of capturing the different clusters in the data (increasing the number of clusters resulted in repeated cluster centroids while reducing it missed out on some centroids).

Once we perform this clustering, we can replace each frame of each input time series data matrix corresponding to both speakers in the dyad (**A** and **B**) with the best matching cluster label (corresponding to the cluster to which it belongs). This way, the data matrix is now represented by a single row vector of cluster labels, $\mathbf{A}_{quant}$ and $\mathbf{B}_{quant}$. A HoC-representation of *lag* τ is then defined as a vector where each entry corresponds to the number of times all pairs of cluster labels are observed τ frames apart. In other words, we construct a vector of lag-τ co-occurrences where each entry (m,n) signifies the number of times that an entry in the first time-series **A** is encoded into a cluster label m at time t, while an entry in the second time-series **B** is encoded into cluster label n at time $t+\tau$ (in the row vectors $\mathbf{A}_{quant}$ and $\mathbf{B}_{quant}$, respectively) [15, 14]. By summing across the columns as shown in Figure 2, each interval can be represented by a single column vector where the elements express the sum of all C^2 possible lag-τ co-occurrences (where C is the number of clusters; in our case, 16). We can repeat the procedure for different values of τ, and stack the results into one "supervector". Note however, that the dimensionality of the HoC feature increases by a factor of C^2 for each lag value τ that we want to consider. In our case, we decide to choose four lag values of 0 to 3 frames (corresponding to 0-3s) in order to capture behavioral synchronicity or asynchronicity within a 3 second window (see for instance [13]).

[1] All features obtained from the FACET SDK are normalized and baselined (speaker-specifically) to a $[-5,5]$ scale. Therefore we assume that features obtained from different speakers are comparable due to this preprocessing.

Table 1: Means and standard deviations of distances between HoC features computed between dyads and nominal dyads of each speaker.

Speaker	(True) Dyad		Nominal Dyad	
	Mean	Std	Mean	Std
1	0.25	0.50	1.66	0.41
2	0.25	0.50	1.60	0.50
3	0.29	0.38	1.30	0.44
4	0.29	0.38	1.48	0.41
5	0.28	0.42	1.24	0.53
6	0.28	0.41	1.49	0.49
7	0.26	0.42	1.35	0.45
8	0.26	0.42	1.45	0.47
9	0.16	0.58	1.4	0.62
10	0.16	0.58	1.72	0.5
11	0.29	0.46	1.5	0.52
12	0.28	0.46	1.48	0.52
13	0.18	0.54	1.48	0.55
14	0.18	0.54	1.41	0.57
15	0.21	0.53	1.39	0.6
16	0.21	0.52	1.57	0.53
17	0.17	0.58	1.7	0.4
18	0.17	0.58	1.68	0.57
19	0.16	0.62	1.97	0.44
20	0.16	0.62	1.4	0.68
21	0.24	0.36	1.48	0.37
22	0.25	0.37	1.31	0.31
23	0.23	0.4	1.44	0.44
24	0.22	0.4	1.2	0.45
25	0.25	0.48	1.62	0.47
26	0.25	0.48	1.5	0.52

4. ANALYSES AND OBSERVATIONS

In order to observe how well HoC features capture dyadic behavior, we randomly extracted 100 time-intervals (each 10 seconds long) from the post-processed and synchronized feature streams for all 26 subjects. We then computed HoC features for each of these intervals for each subject, respectively. Now recall that in this pool of subjects, each subject has one true dyad with whom they completed the Tetralogue task collaboratively. We hypothesize that the HoC features computed for true dyads will be significantly different as compared to the HoC features computed between artificial or nominal dyads (who did not actually engage in a dyadic interaction). While it should not be too surprising to have this hypothesis indeed be true, explicitly testing this also allows us to ensure that our proposed features capture something meaningful. So the goal of this particular exercise is not necessarily to show that true dyads exhibit different characteristics as compared to artificial dyads, but to *discover meaningful features that help us capture these differences*.

In order to test the above hypothesis, we performed the following steps for *each* speaker: (i) if a candidate dyad for that speaker was

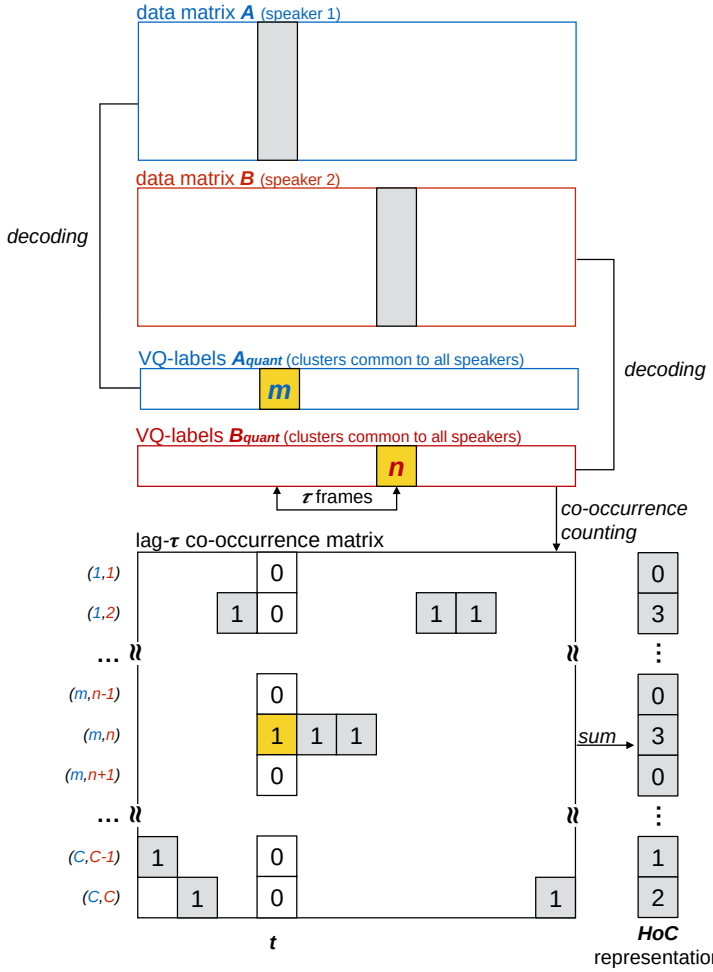

Figure 2: Schematic depiction of the computation of histograms of cooccurrences (HoC) (adapted from [14]). For a chosen lag value, τ, and a time step t, if we find labels m corresponding to the first person (A) and n corresponding to the second person (B) occurring τ time steps apart (marked in gold), we mark the entry of the lag-τ cooccurrence matrix corresponding to row (m,n) and the t^{th} column with a 1 (corresponding entry also marked in gold). Note that the indices corresponding to the first person in the dyad is marked in blue while those corresponding to the second in red. We sum across the columns of this matrix (across time) to obtain the lag-τ HoC representation.

his true dyad, then we computed distances between each of the (100) HoC features computed for that true dyad *only*; (ii) however, if this was not the case, then we computed distances between HoC features computed on that speaker and each of the other 24 candidate dyads in the pool of speakers. We then applied a Wilcoxon rank-sum test[2] to test the hypothesis that the medians of the distance distributions computed in cases (i) and (ii) described above were equal ($\alpha = 0.95$).

We found that the distances computed between HoC features extracted from true dyads were significantly lower ($p \approx 0$) than those of distances between HoC features computed on artificial dyads (see Table 1 for means and standard deviations of these populations computed for each speaker). This finding suggests that (i) not only do true dyads engaged in a collaborative interaction exhibit specific characteristic patterns of emotional state cooccurences that clearly sets them apart from artificial dyads, but (ii) such HoC features allow us to capture these differences in an effective manner.

Figures 3 and 4 gives us some more insight into why these features perform well. Figure 3 depicts the 16 cluster centroids computed on (and therefore common to) all speakers. Notice that each column of Figure 3 represents one cluster centroid, comprising different relative activation of different emotions – for instance, cluster 2 represents an emotional state with a higher activation of joy and positive emotion, while cluster 6 represents a more neutral emotional state, encompassing an equal (and approximately zero) activation of all emotions. Recall that these emotion clusters are common to *all* speakers. Figure 4 shows feature distributions of HoC features computed on one particular speaker and his/her actual dyadic partner, and those computed on that same speaker and an artificial dyadic partner. We observe that the feature distributions of the former are more peaky, with specific certain clusters of emotions

[2]We used the results of a non-parametric Wilcoxon rank-sum test instead of a parametric counterpart such as a t-test, as the data failed a Kolmogorov-Smirnov test of parametricity. Note however that applying a t-test gave similar results as in the case of its non-parametric counterpart.

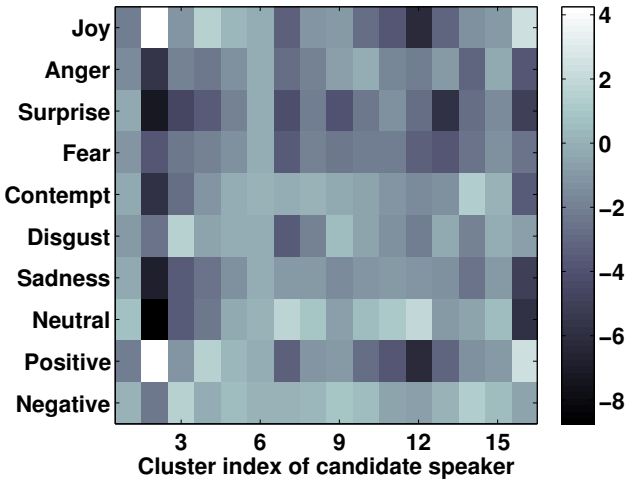

Figure 3: Schematic illustrations of the emotion feature clusters computed for *all* speakers. Each column represents an emotional cluster centroid, which is a particular distribution of emotional state activations. There are 10 dimensions that describe an emotional state, represented by different rows. The colors represent the odds, in logarithmic (base 10) scale, of a target expression being present (typically range: $[-5,+5]$).

co-occurring more often than others. However, in the case of the latter, this distribution is more flat and uniformly distributed. Note that while specific results shown in Figure 4 are particular to the chosen speaker, we observe the aforementioned trends are in general for all speakers. In other words, true dyads display specific patterns of behavioral cooccurrence and synchronicity that are not observed in artificial dyads, and such a HoC feature is helpful in understanding and bringing out these differences.

5. CONCLUSIONS AND OUTLOOK

This paper has made an initial attempt at proposing a novel feature to capture behavioral synchronicity between dyads involved in a CPS task. This feature, dubbed histograms of cooccurrence, captures how often different prototypical behavioral states exhibited by one person co-occur with those exhibited by his/her partner over different temporal lags. We have shown that not only does this feature bring out the differences between dyads and non-dyads, but is also interpretable in that it tells us which behavioral states are most likely to occur in dyads as opposed to non-dyads. In the future, we plan to analyze these features further in order to understand specific aspects of behavioral entrainment and convergence in collaborative interactions, such as a more in-depth analysis of mirroring phenomena. In addition, we aim to extend this analysis beyond just emotion features to data/feature streams derived from multiple channels including video, audio, and text.

6. ACKNOWLEDGMENTS

We would like to thank Ketly Jean-Pierre and Jiangang Hao for help in capturing the data used in this study, and Yuchi Huang and David Suendermann-Oeft for insightful technical discussions.

7. REFERENCES

[1] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learnersâĂŹ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010.

[2] S. Barsade. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, 47:644–âĂŞ675, 2002.

[3] D. L. Bazaldua, S. Khan, A. von Davier, J. Hao, L. Liu, and Z. Wang. On convergence of cognitive and non-cognitive behavior in collaborative activity. In *The 8th International Conference on Educational Data Mining (EDM 2015)*.

[4] D. Belenky, M. Ringenberg, J. Olsen, V. Aleven, and N. Rummel. Using dual eye-tracking to evaluate studentsâĂŹ collaboration with an intelligent tutoring system for elementary-level fractions. In *Proceedings of COGSCI*, pages 176–181, 2014.

[5] S. Bilakhia, S. Petridis, and M. Pantic. Audiovisual detection of behavioural mimicry. In *IEEE Humaine Association*

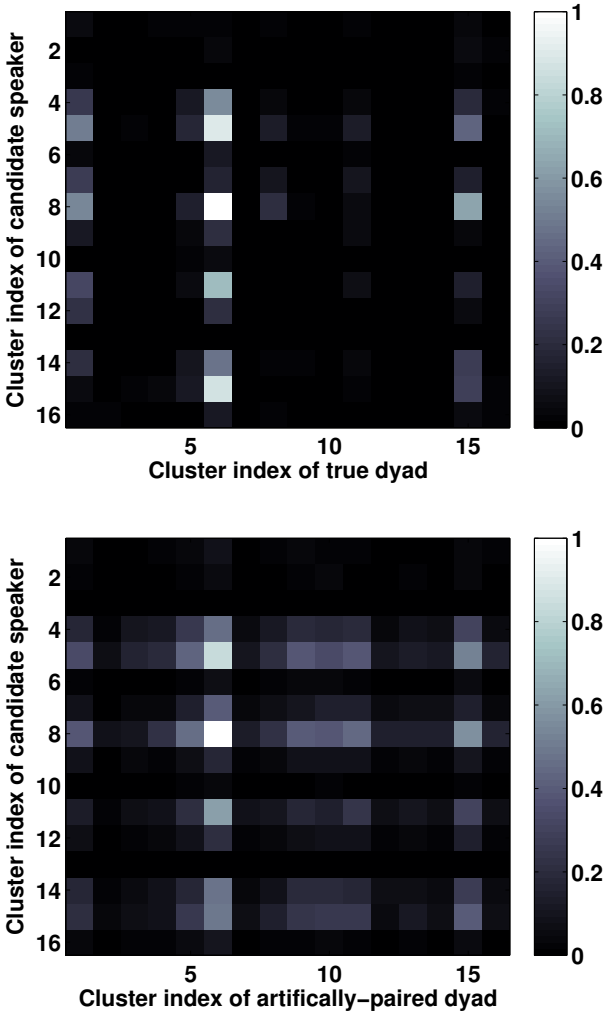

Figure 4: Average HoC feature distributions (across lags) for the true and nominal dyad, respectively, of one particular speaker in the database. The color in the $(m,n)^{th}$ square represents the average normalized activation (between 0 and 1) of cluster m of the speaker represented along the y-axis co-occuring with cluster n of the speaker represented along the x-axis.

Conference on Affective Computing and Intelligent Interaction, Chicago, 2013.

[6] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[7] R. Jucks, B.-M. Becker, and R. Bromme. Lexical Entrainment in Written Discourse: Is Experts' Word Use Adapted to the Addressee? *Discourse Processes*, 45(6):497–518, Nov. 2008.

[8] J. L. Lakin, V. E. Jefferis, C. M. Cheng, and T. L. Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3):145–162, 2003.

[9] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova. Acoustic-prosodic entrainment and social behavior. *Proceedings of NAACL/HLT 2012*, pages 11–19, 2012.

[10] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[11] L. Liu, J. Hao, A. A. von Davier, P. Kyllonen, and D. Zapata-Rivera. A tough nut to crack: Measuring collaborative problem solving. *Handbook of Research on Technology Tools for Real-World Skill Development*, page 344, 2015.

[12] J. S. Pardo. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382, 2006.

[13] A. Paxton, D. H. Abney, C. T. Kello, and R. Dale. Network analysis of multimodal, multiscale coordination in dyadic problem solving. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 2735–2740, 2013.

[14] V. Ramanarayanan, C. W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 23–30. ACM, 2015.

[15] V. Ramanarayanan, M. Van Segbroeck, and S. Narayanan. Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Computer Speech and Language*, 36:330–346, 2016.

[16] B. Schneider and R. Pea. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, 8(4):375–397, 2013.

[17] A. A. Tawfik, L. Sanchez, and D. Saparova. The effects of case libraries in supporting collaborative problem-solving in an online learning environment. *Technology, Knowledge and Learning*, 19(3):337–358, 2014.

[18] J. Thomason, H. V. Nguyen, and D. Litman. Prosodic entraiment and Tutoring Dialogue Success. In H. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *Artificial Intelligence in Education, AIED 2013*, pages 750–753. Springer, 2013.

[19] H. Van hamme. HAC-models: a novel approach to continuous speech recognition. In *Interspeech*, 2008.

[20] M. Van Segbroeck and H. Van hamme. Unsupervised learning of time–frequency patches as a noise-robust representation of speech. *Speech Communication*, 51(11):1124–1138, 2009.

[21] D. Zapata-Rivera, T. Jackson, L. Liu, M. Bertling, M. Vezzu, and I. R. Katz. Assessing science inquiry skills using trialogues. In *Intelligent Tutoring Systems*, pages 625–626. Springer, 2014.