SMM23, Workshop on Speech, Music and Mind 2023
18 August 2023, Trinity College Dublin, Ireland

Teamwork Quality Prediction Using Speech-Based Features

Martin Meza[1], *Lara Gauder*[1,2], *Lautaro Estienne*[1,2], *Ricardo Barchi*[1], *Agustín Gravano*[3], *Pablo Riera*[1,2], *Luciana Ferrer*[1]

[1] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[2]Departamento de Computación, FCEN, Universidad de Buenos Aires (UBA), Argentina
[3]Laboratorio de Inteligencia Artificial, Universidad Torcuato Di Tella, Argentina

mmeza@dc.uba.ar, mgauder@dc.uba.ar, lestienne@dc.uba.ar, gbarchi@dc.uba.ar, agravano@utdt.edu, priera@dc.uba.ar, lferrer@dc.uba.ar

Abstract

This paper describes a novel protocol for annotating teamwork quality and related variables, based only on the speech signal. Our protocol was designed to annotate a Spanish version of the Objects Games corpus, a publicly available corpus that contains dialogues of people playing a collaborative computer game. The corpus was annotated by 4 raters, who achieved an Intraclass Correlation Coefficient of 0.64 for the main teamwork quality metric. Using the resulting annotations, we developed a system for automatic prediction of the average teamwork quality across raters using features extracted from the conversations, reaching a coefficient of determination, R^2 of 0.56. This result suggests that automatic prediction of teamwork quality from the speech signal of the teammates is a feasible task.
Index Terms: teamwork quality prediction, data annotation

1. Introduction

Teamwork is considered a crucial factor for success in many contexts, including software development, healthcare and medical management, and sports [1, 2, 3, 4, 5]. A widely accepted definition of a team is two or more individuals with specified roles interacting adaptively, interdependently, and dynamically toward a common and valued goal [6]. Further, in [7], teamwork is defined as a set of interrelated thoughts, actions, and feelings between team members that are needed to function as a team and that combine to facilitate coordinated, adaptive performance and task objectives resulting in value-added outcomes.

The study of teams and effective teamwork has been addressed from many different domains over the years. This leads to multiple definitions of factors that affect teamwork including team composition [8], interpersonal and self-management skills [9], and amount of effort required by the task [10]. Also, studies argue that teamwork is a dynamic component that should be measured and studied at the team level and is inextricably tied to context [11]. Therefore, teamwork involves several core components and supporting coordinating mechanisms [7].

Measuring the quality of teamwork from team members' interactions could allow for early interventions when the quality of the teamwork is poor. To this end, multiple teaming metrics have been proposed, such as sociometric badges, physiological metrics, external observer-based metrics [12], and communication-based metrics. In particular, communication-based metrics are an effective way to assess teamwork quality [13, 14] for being holistic measures of team behavior as opposed to an aggregate of individual behavior of team members.

In this paper, we focus on the study of teamwork using measures based on the oral communication between team members while solving a specific task. To our knowledge, no speech corpus with teamwork annotations was available to perform our study. While the dataset in [15] contains teamwork-related labels for speech data, unfortunately, the quality of the audio samples is poor, making automatic processing very challenging. For this reason we developed a protocol for annotating a publicly available dataset. The selected corpus was a portion of the UBA Games and Directions Corpus [16], where teams of two subjects play a collaborative game in which they have to interact to locate an object in the correct place. Raters were asked to report several metrics related to the teamwork between players, from more general to more specific concepts, based only on the speech recorded by the subjects during each game. The implemented annotation protocol and the chosen annotated variables were designed to be as general as possible to allow the protocol to be useful for annotating other databases in the future.

In the following sections, we describe the data annotation process as well as some baseline approaches for automatically predicting the resulting teamwork ratings using the audio recordings. We analyze the resulting annotations and show that moderate to high levels of agreement across raters were achieved, which validates the proposed protocol. Further, we show that the automatic prediction of teamwork quality reaches R^2 values of over 0.5 using simple features extracted from the conversation. To our knowledge, no prior work had attempted the automatic prediction of teamwork quality based on the speech of the team members. We hope this work will serve as a baseline for future work on this topic. The teamwork annotations and scripts needed to produce the results in this paper are freely available for research purposes upon request.

2. Data

For this work, we used the first batch of 14 dialogues from the UBA Games and Directions Corpus [17] – a version of the Objects Games Corpus [18]. This subset of the corpus consists of a compilation of 14 dialogues between native speakers of Spanish playing the Objects Game, in which both players see the same set of 5 to 7 objects on their laptop screens, except for one: the target. The Describer's target appears in a random location among the other objects, while the Follower's target appears at the bottom of their screen. The Describer must explain the position of their target so the Follower can move their target to the same spot. Once they agree on a location match, they are awarded 1 to 100 points depending on how close the selected location was to the correct one. The subject's goal was to achieve a high score, with a financial bonus depending on the number of points they secured. Each player used a separate laptop which was not visible to the other player, divided by a curtain to ensure all communication was verbal. Subjects alternated in the describer and follower roles. This game setup has been shown to elicit natural task-oriented dialogue [18, 17].

10.21437/SMM.2023-1

A total of 14 subjects participated in the corpus, each one playing two sessions with a different randomly-selected teammate. Each session consisted of 14 instances of the game which we will call *tasks*, resulting in a total of 196 audio samples. The subjects' ages ranged from 19 to 59 years ($M = 28.6$, $SD = 12.7$). The 14 sessions were composed of 5 male-male pairs, 5 female-female pairs and 4 male-female pairs. The corpus contains approximately 386 minutes of speech, and the average task duration is 110 seconds.

3. Annotation protocol

For the annotation process, the 196 audio samples corresponding to the 14 individual tasks from the 14 sessions were randomly grouped in blocks of 14 with the restriction that only one task per team appeared in each block. This was done to encourage raters to evaluate each task independently of the other tasks in a session, considering the time-varying nature of teamwork quality [11]. A different random order was used for each rater. The blocks had an average audio duration of 25 minutes. Raters were instructed to annotate at most 2 blocks per day, one in the morning and one in the afternoon. The average time it took raters to annotate each block was 54 minutes.

The annotation process was carried out remotely, through a web page. The raters were required to have a computer with a stable internet connection, and over-ear headphones. They were asked to do the task in a silent room and take precautions to reduce possible sources of noise (close their windows, turn off home appliances as TV, fans, etc.). Next, they were asked to log in with their credentials and go through an audio setup stage: they had to listen to two sample monologues, one from a male and one from a female speaker, and were asked to adjust their volume to a comfortable level. Next, we presented a brief description of the Objects Game, to provide some context about the audio recordings that they would be evaluating, and a guided tour of the form they had to fill in for each audio sample.

After the guided tour, raters were asked to read the definitions of all the variables, dividing them into three groups: general teamwork quality, teamwork components and mechanisms, and social behavior. They were able to return to these definitions at any time during the annotation process. The complete definitions shown to the raters both in Spanish and in English can be requested by email to the authors. In Section 3.1, we give summarized definitions of all the variables annotated.

After reading the definitions, raters were presented with a first block of hand-picked tasks. The recordings were selected to represent the variety of conversations in the corpus. The annotations for this block were then discarded, and its recordings were repeated in other blocks toward the end of the sequence of blocks. This way, the first block serves as a calibration process for the raters to adjust their use of the 5-level Likert scale.

Five raters were recruited for this task. All of them were native speakers of Spanish. Three of them were considered experts as they were Psychology researchers, while the other 2 were considered non-experts raters. A deadline of 30 days was given to annotate the audio files of all 196 tasks. One expert annotator did not finish the task by the deadline; therefore their annotations were discarded.

3.1. Annotation variables

For general teamwork quality, raters had to answer the question "How good was the teamwork in this recording?" in a 5-level Likert scale, and report the level of confidence in their response. This question was meant to provide a general assessment of teamwork quality as perceived by the raters. Next, raters were asked to indicate at least one teamwork component or mechanism they had taken into consideration to respond to the first question, indicating negative or positive influence on a scale from -2 to 2. These components are taken from [7] and they are briefly defined as follows. **Team leadership** (**TL**): Ability to direct and coordinate the activities of team members, assess team performance, assign tasks, develop team knowledge/skills/abilities, motivate team members, plan/organize, and establish a positive atmosphere. **Mutual performance monitoring** (**MPM**): Ability to develop common understandings of the team environment and apply strategies to accurately monitor team performance. **Backup behavior** (**BB**): Ability to anticipate team members' needs through accurate knowledge of their responsibilities. This includes the ability to shift workload among members to achieve balance during high periods of workload or pressure. **Adaptability** (**Ad**): Ability to adjust strategies based on information gathered from the environment through the use of backup behavior and reallocation of intrateam resources. **Team orientation** (**TO**): Propensity to take other's behavior into account during group interaction and the belief in the importance of team goals over individual members' goals. **Shared mental models** (**SMM**): Organizing knowledge structure of the relationships among the task the team is engaged in and how the team members will interact. **Mutual trust** (**MT**): Shared belief that team members will perform their roles and protect the interests of their teammates. **Closed-loop communication** (**CLC**): Exchange of information between sender and receiver irrespective of the medium.

Finally, raters had to answer three yes-no questions related to social behavior: *Does the conversation flow naturally? (**FN**) Are the participants having trouble understanding each other? (**HTU**) Is the conversation awkward? (**CA**)*. These same questions had previously been annotated for the Columbia Games corpus [19]. For each task, raters were also able to leave comments about the audio sample or their annotations.

4. Analysis of teamwork annotations

The Intraclass Correlation Coefficient (ICC) [20] between the four raters for the teamwork quality rating was 0.64. Values between 0.5 and 0.75 are thought to be indicative of moderate inter-rater agreement [20]. A permutation test with 1000 permutations resulted in ICC values below 0.26 ($M = 0.01$, $SD = 0.11$) indicating that the ICC of 0.64 is statistically significantly better than random agreement in this dataset.

Figure 1 shows the ICC, the average teamwork quality rating, and the average reported confidence in the rating, as a function of the task duration. To obtain each point in these lines, a shifting window of 14 samples sorted by increasing task duration was used. The figure shows that the agreement falls dramatically for a task duration above 100 seconds. We hypothesize that this happens because long samples may include changes in teamwork quality, as teamwork quality is a dynamic characteristic. Hence, if one annotator focuses their attention on the beginning of the task while another focuses on the end, their ratings might differ. These results suggest that, for long audio samples, teamwork quality annotations should be done by splitting the audio into shorter windows for more accurate annotations. Given this finding, in order to avoid using samples for which annotations are not reliable, only tasks shorter than 100 seconds were used for the experiments, keeping 111 of the 196 original tasks. The annotator agreement for these samples

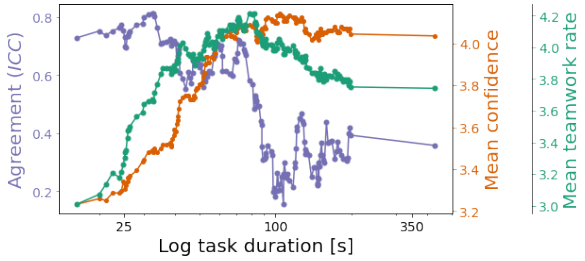

Figure 1: *Annotator agreement, average rating and average confidence as a function of task duration.*

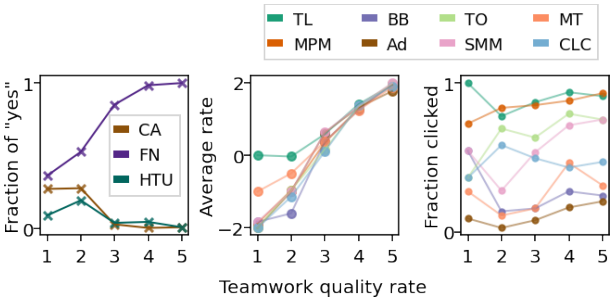

Figure 2: *Annotation variables vs. teamwork quality rating*

increases to 0.79 (compared to 0.64 over all samples). The remaining tasks will need to be split into shorter samples and re-annotated before they can be used for experimentation.

Another interesting observation from Figure 1 is that shorter tasks have the lowest average confidence (average confidence of around 3 compared to the peak confidence of over 4 at longer tasks). Based on comments left by the raters, these shorter tasks were difficult to annotate due to the small amount of information exchanged between the players. In most of these short tasks, the describer gave a few indications, and the follower did not answer or answered with a backchannel. This limited interactions resulted in the raters having low confidence in their ratings. Nevertheless, despite the low confidence, raters tended to agree that these short tasks corresponded to a low quality teamwork, as seen in Figure 1.

Finally, Figure 2 briefly describes the annotations, as a function of teamwork quality rating. The left pane shows the fraction of affirmative answers to the three social-behavior questions, where we see reasonable trends of FN increasing and HTU and CA decreasing as the annotated teamwork quality increases. The right pane shows the fraction of times each dimension was selected for each rating. We can see that the variables selected the most were TL and MPM. Other components, like Ad, BB and MT, were selected less frequently, perhaps for not being too relevant for these specific short-duration game-oriented dialogues. Lastly, the middle pane shows the average ratings for the selected variables. We see a clear trend with values increasing with teamwork quality. Notably, the worse average TL values are not negative but close to 0 indicating that raters related neutral TL with bad teamwork.

5. Prediction of teamwork ratings

Given the teamwork quality ratings obtained using our proposed protocol we could now explore the use of machine learning techniques for the automatic prediction of these ratings. To this end, we used the average rating across all four raters for each task as regression target for our models. For this initial effort, we used random forest regressors (RFR) as models, implemented using scikit-learn [21]. RFRs were chosen for being robust general models. We explored a variety of input features extracted from the speech of the two participants in the task.

Acoustic features were extracted following the eGeMAPS specification. This set of features was first proposed in [22] and, since then, it has been used successfully for a variety of high-level speech processing tasks, including the detection of mental states [23, 24]. The eGeMAPS features include *frequency, energy, spectral* and *temporal*-related features. They were extracted using the OpenSmile toolkit (v2.4.2) [25]. A second set of features are those related to word and turn counts and speech and silence duration. The speaker turn information was manually annotated in the UBA games corpus by the authors of the corpus [26]. A turn is defined as a maximal sequence of inter-pausal units (IPUs – maximal sequences of words surrounded by silence longer than 50 ms) from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor [16]. Further, each turn is annotated with a turn-type label (e.g., smooth switch, backchannel, or pause interruption). Table 1 includes a description of each feature set.

Features for each task were extracted separately for each speaker role (describer/follower) and then concatenated, listing the describer's features first. Also, a duration-normalized version of the features that contain duration information was computed to analyze the effect of ignoring the information about the total duration of the task. Table 1 shows the types of normalization applied to the features that have a normalized version.

For hyperparameter tuning we used the Columbia Games Corpus [18], which is an English version of the corpus we use for testing, since holding out data from this corpus for tuning was not feasible given the its size. The features from Table 1 were extracted on that corpus and a RFR was trained to predict one of the labels available in that dataset which corresponds to whether the conversation flowed naturally. These ratings were in a 5-level scale. A grid search parameter was performed for the number of trees (100-500 with steps of 10), the maximum depth (2-20 with steps of 1) and the percentage of features considered at each spit (0-100, with steps of 10). For each combination of parameters, the test metric was calculated over 10 models with different seeds and then averaged. The best RFR for that task on that dataset used 300 trees in the ensemble, a maximum depth of 3 and 100% of features considered at each split. All other hyperparameters were left at their default values provided by the scikit-learn method.

For the experiments, 7-fold cross-validation was used to obtain predicted ratings on all samples longer than 100 s. The folds were created by splitting by session. The splits were regenerated 100 times with different seeds. Finally, for each seed, the predictions from all the folds are pooled and 100 bootstrap sets are generated to obtain confidence intervals. For each bootstrap set, the coefficient of determination, R^2, is computed as performance metric [27]. The R^2 for all the bootstrap sets for all seeds are pooled together to obtain confidence intervals.

Figure 3 shows the RFR model R^2 values across the different feature sets. For reference, a baseline system which always outputs the mean rating over the training set is included. This system is the best possible system that does not have access to the input samples. The best performance is obtained by using only the turn_total_cnt feature set. Further, the top feature subsets are those that reflect in some way the duration of the task (which we call "task-duration-aware" sets). This result was expected since Figure 1 showed a strong correlation between the task duration and the average teamwork rating for short dura-

Table 1: *Features extracted over each task recording; μ, σ, P_5 and P_{95} are the mean, standard deviation, and percentiles 5% and 95%, respectively, and are calculated after normalization when required. All features except task_duration were extracted separately for each speaker and then concatenated. The normalization statistics are always obtained over both speakers. The second block of features are extracted using manual annotations. The bottom block corresponds to various subsets of eGeMAPS features.*

Attribute	Description	Normalized version (divided by)
task_dur	Total task duration	-
turn_total_cnt	Total number of turns	Task duration, and speech duration
turn_type_cnt	Number of each turn-taking category	Number of turntakings
word_cnt	Total number of words, μ, σ, P_5 and P_{95} of number of words in each turn	Total number of words, task duration
speech_dur	Total speech duration, μ, σ, P_5 and P_{95} of speech duration in each turn	Total speech duration and task duration
sil_dur	Total silence duration, μ, σ, P_5 and P_{95} of silence duration in each turn	Total silence duration and task duration
turn_dur	μ, σ, P_5 and P_{95} of turns duration	-
frequency	Pitch, jitter, formant 1, 2 and 3 frequency, formant 1 bandwidth	-
energy	Shimmer, Loudness, and harmonic to noise ratio	-
spectral	Spec slopes, formants, harmonic diff, relative energy, spec flux and MFCC 1-4	-
spec_voiced	Alpha Ratio, Hammarberg Index, spec slopes, spec flux and MFCC 1-4 in voiced regions	-
spec_unvoiced	Alpha Ratio, Hammarberg Index, and spec slopes over unvoiced segments	-
temporal	Voiced and unvoiced segments per second and length of segments	-

tion tasks (up to 60 seconds approximately). Interestingly, the turn_total_cnt feature outperforms the task_dur feature, indicating that the most relevant information for the task is not the total duration of the task but, rather, the number of turns, with the total duration being a good –albeit imperfect– proxy for this feature. Note, though, that while task_dur can be extracted automatically from the signal, all the other task-duration-aware sets require manual annotations. Nevertheless, automatic versions of these features could potentially be extracted. We plan to explore this direction in future work.

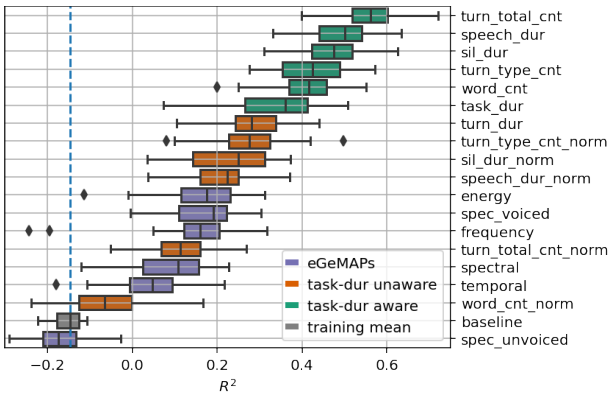

Figure 3: *R^2 for each set of features in Table 1 including the normalized versions. Colors indicate the type of feature.*

Among subsets that are not directly affected by the total task duration (which we call "task-duration unaware"), the best performance is achieved with turn_dur, closely followed by turn_type_cnt_norm. Overall, the best task-duration aware systems reach R^2 values over 0.5, while the best task-duration unaware systems reach R^2 values around 0.3. Clearly, the biggest source of information about teamwork quality for this dataset is given by task duration. Yet, this may not be the case for other types of tasks. Fortunately, we see that, even if total duration information is ignored, other features can be used for predicting teamwork quality with better-than-chance performance.

Finally, in order to study whether combinations of different sets can lead to better performance than the individual sets, we train a model using all the available features and compare its performance with the best single feature and the best two-way combination of sets. We do this considering groups of feature sets: 1) all sets, and 2) only the task-duration unaware sets. Figure 4 shows that, for the first case in which all sets are available, the best performance is achieved with the combination of turn_total_cnt and sil_dur_norm, and for the second condition the best performance is achieved with turn_dur and sil_dur_norm. Finally, for comparison, a system that includes all features in Table 1 is included. Notably, no gains can be observed from the use of more than one feature. Adding features seems to be causing the RFR to overfit to the training data resulting in poorer results in the test data. This is likely due to the very small amount of data available for training these models.

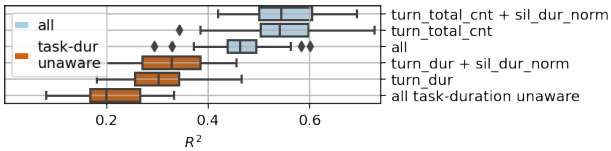

Figure 4: *Best feature sets among task-duration aware and task-duration unaware sets.*

6. Conclusions

We presented a protocol for annotating teamwork quality using a web-based application. The protocol was used to annotate a corpus of game-driven dialogs. We found a moderate to high agreement across raters for the task of rating the teamwork quality, validating our protocol. Automatic detection of the resulting ratings using speech-based features achieved a coefficient of determination of 0.56, using only the number of turns from each of the two speakers in the dialog. For this dataset, the number of interactions between speakers was a strong predictor of teamwork quality, with samples with fewer interactions being detected as having lower teamwork quality. This is possibly a dataset-specific conclusion. When considering only features that do not encode the total number of interactions, a maximum R^2 of 0.30 was achieved using statistics on the distribution of durations of the speaker turns for each speaker, indicating that even with features that are unaware of the total number of turns, better-than-chance results can be obtained for this task. These result are a promising first step in the study of automatic teamwork quality prediction from speech.

7. Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award no. FA9550-18-1-0026.

8. References

[1] A. Richter, J. F. Dawson, and M. A. West, "The effectiveness of teams in organizations: a meta-analysis," *The International Journal of Human Resource Management*, Aug. 2011.

[2] M. Hoegl and H. G. Gemuenden, "Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence," *Organization Science*, Aug. 2001.

[3] T. Dingsøyr and T. Dybå, "Team effectiveness in software development: Human and cooperative aspects in team effectiveness models and priorities for future studies," *Cooperative and Human Aspects of Software Engineering (CHASE)*, Jun. 2012.

[4] H. K. Westli, B. H. Johnsen, J. Eid, I. Rasten, and G. Brattebø, "Teamwork skills, shared mental models, and performance in simulated trauma teams: an independent group design," *Scandinavian Journal of Trauma Resuscitation and Emergency Medicine*, Aug. 2010.

[5] S. I. Sabin and S. D. Alexandru, "Study regarding the importance of developing group cohesion in a volleyball team," *Procedia - Social and Behavioral Sciences*, no. 180, p. 1343–1350, May 2015.

[6] E. Salas, T. L. Dickinson, S. A. Converse, and S. I. Tannenbaum, "Toward an understanding of team performance and training," *R. W. Swezey & E. Salas (Eds.), Teams: Their training and performance*, p. 3–29, 1992.

[7] E. Salas, D. E. Sims, and C. S. Burke, "Is there a "big five" in teamwork?" *Small Group Research*, Oct. 2005.

[8] J. M. Levine and R. L. Moreland, "Progress in small group research," *Annual review of psychology*, no. 41, p. 585–634, Oct. 1990.

[9] M. J. Stevens and M. A. Campion, "The knowledge, skills and ability requirements for teamwork: Implications for human resources management," *Journal of Management*, Oct. 1994.

[10] J. R. Hackman and C. G. Morris, "Group tasks, group interaction process, and group performance effectiveness: A review and proposed intergration," *Journal of Management*, Oct. 1994.

[11] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Durand, "Interactive team cognition," *Cognitive Science: A Multidisciplinary Journal*, Nov. 2012.

[12] E. Salas, R. Grossman, A. M. Hughes, and C. W. Coultas, "Measuring team cohesion: Observations from the science," *Human Factors: The Journal of Human Factors and Ergonomics Society*, no. 41, p. 585–634, May 2015.

[13] P. A. Kiekel, N. Cooke, P. Foltz, and S. Shope, "Automating measurement of team cognition through analysis of communication data," *Usability evaluation and interface design*, p. 1382–1386, Jun. 2001.

[14] P. A. Kiekel, N. J. Cooke, P. W. Foltz, J. Gorman, and M. Martin, "Some promising results of communication-based automatic measures of team cognition," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, p. 298–302, Jun. 2002.

[15] M. Braley and G. Murray, "The group affect and performance (gap) corpus," in *Proceedings of the ICMI 2018 Workshop on Group Interaction Frontiers in Technology (GIFT)*, 2018.

[16] P. Brusco, J. Vidal, S. Benus, and A. Gravano, "A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool," *Speech Communication*, 2020.

[17] A. Gravano, R. Levitan, L. Willson, Štefan Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," *Cooperative and Human Aspects of Software Engineering (CHASE)*, Jun. 2012.

[18] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, Aug. 2011.

[19] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2012.

[20] P. A. Kiekel, N. Cooke, P. Foltz, and S. Shope, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, pp. 155–163, 2016.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[23] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *INTERSPEECH*, Oct. 2020.

[24] R. Li, J. Zhao, J. Hu, S. Guo, and Q. Jin, "Multi-modal fusion for video sentiment analysis," *MuSe'20: Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, Oct. 2020.

[25] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," *Proceedings of the 2013 ACM Multimedia Conference*, p. 835–838, 2013.

[26] A. G. Pablo Brusco, Juan Manuel Pérez, "Cross-linguistic study of the production of turn-taking cues in american english and argentine spanish," *Interspeech*, 2017.

[27] N. R. Draper and H. Smith, *Applied Regression Analysis*. New York: Wiley-Interscience, 1998.