Weighting Model Based on Group Dynamics to Measure Convergence in Multi-party Dialogue

Zahra Rahimi
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA, 15260
zar10@pitt.edu

Diane Litman
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA, 15260
litman@cs.pitt.edu

Abstract

This paper proposes a new weighting method for extending a dyad-level measure of convergence to multi-party dialogues by considering group dynamics instead of simply averaging. Experiments indicate the usefulness of the proposed weighted measure and also show that in general a proper weighting of the dyad-level measures performs better than non-weighted averaging in multiple tasks.

1 Introduction

Entrainment is the tendency of speakers to begin behaving like one another in conversation. The development of methods for automatically quantifying entrainment in text and speech data is an active research area, as entrainment has been shown to correlate with outcomes such as success measures and social variables for a variety of phenomena, e.g., acoustic-prosodic, lexical, and syntactic (Nenkova et al., 2008; Reitter and Moore, 2007; Mitchell et al., 2012; Levitan et al., 2012; Lee et al., 2011; Stoyanchev and Stent, 2009; Lopes et al., 2013; Lubold and Pon-Barry, 2014; Moon et al., 2014; Sinha and Cassell, 2015; Lubold et al., 2015). One of the main measures of entrainment is convergence which is the main focus of this paper. Within a conversation, convergence measures the amount of increase in similarity of speakers over time in terms of linguistic features (Levitan and Hirschberg, 2011).

While most research has focused on quantifying the amount of entrainment between speaker pairs (i.e., dyads), recent studies have started to develop measures for quantifying entrainment between larger groups of speakers (Friedberg et al., 2012; Danescu-Niculescu-Mizil et al., 2012; Gonzales et al., 2010; Doyle and Frank, 2016; Litman et al., 2016; Rahimi et al., 2017a). To date, mainly simple methods such as unweighted averaging have been used to move from dyads to groups (Gonzales et al., 2010; Danescu-Niculescu-Mizil et al., 2012; Litman et al., 2016).

However, because multi-party interactions are more complicated than dyad-level interactions, it is not clear that the contribution of all group members should be weighted equally. For example, to account for participation differences, Friedberg et al. proposed a weighting method based on the number of uttered words of each dyad (Friedberg et al., 2012), although this did not yield performance improvements compared to simple averaging. Rahimi et al. (Rahimi et al., 2017b) provided examples of group-specific behaviors that were not properly quantified using simple averaging. While this case study nicely identified potential problems with prior measures, their observations were only based on a few example dialogues and no solutions were proposed.

In this paper, we propose a new weighting method to normalize the contribution of speakers based on group dynamics. We explore the effect of our method, participation weighting, and simple averaging when calculating group convergence from dyads. We conclude that our proposed weighted convergence measure performs significantly better on multiple benchmark prediction and regression tasks that have been used to evaluate convergence in prior studies (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012; Rahimi et al., 2017a; Doyle et al., 2016; Lee et al., 2011).

2 Convergence for Multi-Party Dialogue

The convergence measure that we extend in this paper is adopted from prior work. Originally, convergence between dyads (Levitan and Hirschberg, 2011) was measured by calculating the difference

Proceedings of the SIGDIAL 2018 Conference, pages 385–390,
Melbourne, Australia, 12-14 July 2018.

between the dissimilarity of speakers in two non-overlapping time intervals. If the dissimilarity in the second interval was less than in the first, the pair was said to be converging.

Extending this work, multi-party convergence (Litman et al., 2016) was measured using Non-Weighted (NW) averaging of each pairs' convergence, as shown in Equations 1 and 2:

$$GroupDiff_t = \frac{\sum_{\forall i \neq j \in group}(|f_{i,t} - f_{j,t}|)}{|group| * (|group| - 1)} \quad (1)$$

$$Conv_{NW} = GroupDiff_{t_1} - GroupDiff_{t_2} \quad (2)$$

$GroupDiff_t$ corresponds to average group differences calculated for linguistic feature f in time interval t for all pairs (i,j). The convergence is the difference between $GroupDiff$s in two intervals.

In the next subsections, we introduce two weighted variations of these formulas: a baseline based on participation ratios (Friedberg et al., 2012), and a method based on group dynamics.

2.1 Weighting Based on Participation

The idea behind this approach is that the weights for speakers that may have talked very little should be reduced. In prior work on multi-party lexical entrainment (Friedberg et al., 2012), speaker participation was measured by number of uttered words; the participation ratios of speaker pairs were then used as the weights.

Since our work focuses on acoustic-prosodic entrainment, we measure speaker participation by amount of speaking time. The Participation Ratio (PR) of each speaker in a given temporal interval is their total speech time divided by the duration of the interval including silences. Speech and silence periods are automatically annotated using Praat (Boersma and Heuven, 2002). The Participation-based Weighted (PW) average of convergence for all pairs p in a group is then computed as follows:

$$Conv_{PW} = \frac{\sum_{\forall p \in group}(Conv_p * PR_p)}{Num_p \sum_{\forall p \in group} PR_p} \quad (3)$$

Num_p indicates number of pairs, and Participation Ratio for a pair, PR_p, for the two intervals is the sum of PRs for both speakers and in both intervals. Finally, convergence for pair $p = (i, j)$ and for two disjoint intervals t_1 and t_2 is calculated as in Equation 4:

$$Conv_{p=(i,j)} = (|f_{i,t_1} - f_{j,t_1}| - |f_{i,t_2} - f_{j,t_2}|) \quad (4)$$

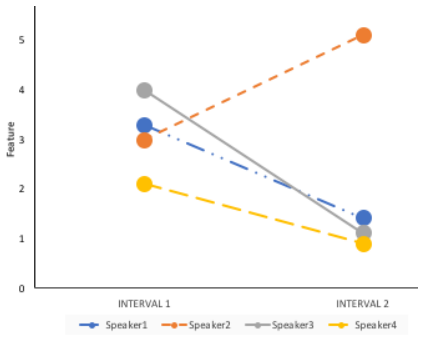

Figure 1: A group in which all speakers except Speaker2 are converging to each other.

2.2 Weighting Based on Group Dynamics

Although participation-based weighting decreases the contribution of less active speakers when calculating group convergence, it does not take group convergence dynamics into account. Rahimi et al. (Rahimi et al., 2017b) argue that it might instead be better to decrease the contribution of speakers whose convergence behaviors differ from the rest of the group (e.g., $Speaker2$ in Figure 1). To tackle this issue, we use weighting to decrease the contribution of outlier speakers. In particular, we propose that the weight for a speaker should be the percentage of individuals who have the same convergence behavior as the speaker.

Equation 5 defines our proposed Group Dynamic-Based Weighted (GDW) convergence measure:

$$Conv_{GDW} = \sum_{g \in G} \frac{|g|}{|N|} * \frac{\sum_{i \in g} \sum_{j \neq i \in N} Conv_{ij}}{|Num_{pair}|} \quad (5)$$

G is a set including three categories: $G = \{Converging, Diverging, MixedBehavior\}$, g is a set of all individuals who belong to a category in G, $|N|$ is the number of all speakers in the group, and $|Num_{pair}|$ is the number of pairs.

Consider the example in Figure 1. There are 12 pairs (6 unique pairs since convergence is a symmetric measure). Each speaker is in three unique pairs with the other three members of the group.

If all conversational pairs that a speaker is involved in have positive convergence values, the speaker is converging to the group and has the *Converging* category. If all involved pairs have negative value, the speaker is diverging from the group. Else, the speaker has a mixed-behavior.

The weight for each category is the number of speakers who have corresponding behavior normalized by the group size. For example, in a group

where all members diverge from each other, the weights will be: $converging = 0$, $diverging = 1$, and $mixedBehavior = 0$. For the group in Figure 1, weights are: $converging = 0$, $diverging = 1/4$, and $mixedBehavior = 3/4$. So, the group convergence for this example is as follows, where $C(i)$ is shortened for sum of pair convergences for speaker i:

$$Conv_{GDW} = 0 * 0 + \frac{1}{4} * C(2) + \frac{3}{4} * [C(1) + C(3) + C(4)] \quad (6)$$

3 Data

To evaluate the utility of weighting based on group dynamics, we measure acoustic-prosodic convergence in the Teams Corpus (Litman et al., 2016). The corpus includes audio files for 62 teams of 3 or 4 individuals playing a cooperative board game in two sessions. First games (Game1) take significantly longer than second games (Game2) (27.1 vs. 18.4 minutes, $p < .001$) and are in chronological order. The teams are disjoint in participants. We break each game into four equal intervals[1] (including silences) and choose the first and last intervals to compute convergence for eight acoustic-prosodic features: maximum (max), mean, and standard deviation (SD) of pitch; max, mean, and SD of intensity; local jitter[2]; and local shimmer[3]. The features are extracted from each of the first and last intervals for each speaker in each team.

Individually taken self-reported pre- and post-game surveys are available for both sessions, including: (1) favorable social outcome measures (perceptions of cohesion, satisfaction, potency/efficacy and perceptions of shared cognition), and (2) conflict measures (task, process, and relationship conflicts). Since favorable measures have high correlations, we z-scored each separate outcome and averaged these scores to make a single omnibus favorable group perception scale and then averaged them for each team to create a team-level **Favorable** measure. Since process conflict was the only conflict measure that could be split at the median without making arbitrary choices [4], we z-scored the process conflict and averaged it in the groups to construct a team-level **Process Conflict** measure. **Favorable** and **Process Conflict** will be used to evaluate the quality of the different convergence measures from Section 2.

4 Experiments and Discussion

Our experimental evaluations use two tasks that have been used for convergence measure evaluations in previous studies (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012; Rahimi et al., 2017a; Doyle et al., 2016; Lee et al., 2011).

Predicting Social Outcomes: Our first task examines how the NW, PW, and GDW measures of acoustic-prosodic convergence (independent variables) relate to the social outcome measures (dependent variables) from Section 3. This is similar to prior studies which have evaluated convergence in terms of predicting outcomes (Doyle et al., 2016; Lee et al., 2011; Rahimi et al., 2017a). We hypothesize that the group-dynamic weighted convergence measure will outperform the non-weighted and participation-based measures.

First, we train a hierarchical multiple regression with each of the three groups of convergence measures, added once in the first level and the other time in the second, to measure if the second level predictors significantly improve the explanation of variance. We only keep predictors with significant coefficients when presenting the models.[5]

For **Process Conflict,** the results show that all NW, PW, and GDW predictor groups are as good as each other; no matter which group is entered in the first level, the predictors in the second level do not significantly improve model fit.

For **Favorable**, neither PW nor NW in the second level significantly improves performance. However, Table 1 shows that adding the GDW measures at the second level significantly improves a model with only NW features at the first level. The amount of variance explained in Model 2 is significantly above and beyond Model 1, $\Delta R^2 = 0.048$, $\Delta F(2, 119) = 3.179$, $p = 0.045$. The reverse order, GDW at first level and NW at the second level, shows that the improvement at the second level is not significant, $\Delta R^2 = 0.031$, $\Delta F(2, 119) = 2.068$, $p = 0.131$. These results indicate that the proposed weighted (GDW) convergence (for intensity max and SD) are the best

[1] Any method of breaking the games to compare two disjoint intervals can be used.

[2] The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

[3] The average absolute difference between consecutive periods, divided by the average amplitude.

[4] The median split is required for our classification tasks.

[5] To control for the effect of first versus second dialogue (game) for each group, we also included an independent variable for game. However, the coefficient was never significant.

	Independent Vars	M1 (β)	M2 (β)
	Intensity_max (NW)	0.248*	-0.164
	Intensity_SD(NW)	-0.055	-0.479+
	Intensity_max(GDW)		0.430+
	Intensity_SD(GDW)		0.457+
R^2		0.063	0.110
F		4.034*	3.678*

Table 1: Hierarchical regression results with intensity max and SD convergence as independent, and **Favorable** as dependent, variables. The NW measures are added in the first level and GDW measures in the second level. Significant / trending results if p-value is < 0.05 (*) or < 0.1 (+).

	Favorable	Process Conflict
Majority	50	53
NW	50	66.93
PW	53.23	**67.74+(GDW)**
GDW	**62.90****	62.90
GDW+PW	58.87	66.13

Table 2: LOOCV prediction accuracies of binary favorable social outcome and process conflict variables. (**) indicates GWD model significantly outperforms both PW and NW models. (+) indicates PW improvement over GDW is trending.

predictors of the favorable social outcome compared with the other two measures of convergence.

Next, we reduce the task from regression to a binary classification by splitting the two social outcome variables at the median. We perform Leave-One-Out Cross-Validations (LOOCV) using a logistic regression (L2) algorithm and all eight acoustic-prosodic features to predict binary outcomes. The results in Table 2 show that the GWD model significantly[6] outperforms both PW and NW models to predict the favorable social outcome. In the prediction of process conflict, the PW model outperforms both NW and GDW models and its improvement over GDW is trending.

In sum, the results in both tables support our hypothesis for the favorable social outcome, where the proposed GDW convergence measure is a better predictor of the outcome. For process conflict, we do not see any significant difference.

Predicting Real Dialogues: The existence of entrainment should not be incidental. To evaluate this criteria, we use permuted versus real conversations as in (De Looze et al., 2014; Lee et al., 2011; Jain et al., 2012). We hypothesize that GDW will be the best convergence measure for distinguishing real versus permuted dialogues.

[6]Corrected paired t-test was performed to address instance dependency from both games (Nadeau and Bengio, 2000).

	All	Game1	Game2
Majority	50	50	50
NW	54.43	60.48	49.19
PW	53.62	58.06	51.61
GDW	54.03	**67.74*+**	48.39

Table 3: Accuracies using the linear SVM models and LOOCV to predict real conversations. (+) indicates GWD outperforms NW with p = 0.06 , (*) indicates GWD outperforms PW with p = 0.004.

For each of the 124 game sessions, we construct artificially permuted versions of the real dialogues as follows. For each speaker, we randomly permute the silence and speech intervals extracted by Praat. Next, we measure convergence for all the groups with permuted audios. We perform a leave-one-out cross-validation experiment to predict real conversations using the convergence measures. We examined several classification algorithms including logistic regression; linear SVM was the only one that showed significant results.

The "All" results in Table 3 show that none of the models significantly outperform the majority baseline. To diagnose the issue, we perform the prediction on each game separately. The proposed GDW model significantly outperforms other models for Game 1. However, for Game 2, none of the results are significantly different. One reason might be that convergence occurs quickly during Game 1, and there is not much convergence occurring at Game 2. Thus, there is no significant difference between permuted and not permuted convergence for any of the features during Game 2.

5 Conclusion

In this paper, we introduced a new weighted convergence measure for multi-party entrainment which utilizes group convergence dynamics to weight pair convergences. Experimental results show that the proposed weighted measure is more predictive for two evaluation tasks used in prior entrainment studies: predicting favorable social outcomes and predicting real versus permuted conversations. In future work we plan to apply the proposed weighted convergence measure to features other than acoustic-prosodic, e.g., lexical.

Acknowledgments

This work is supported by NSF 1420784,1420377. We thank Susannah Paletz for her feedback.

References

Paul Boersma and Vincent van Heuven. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

Celine De Looze, Stefan Scherer, Brian Vaughan, and Nick Cambpell. 2014. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34.

Gabriel Doyle and Michael C Frank. 2016. Investigating the sources of linguistic alignment in conversation. In *ACL (1)*.

Gabriel Doyle, Dan Yurovsky, and Michael C Frank. 2016. A robust framework for estimating linguistic alignment in twitter conversations. In *Proceedings of the 25th international conference on world wide web*, pages 637–648. International World Wide Web Conferences Steering Committee.

Heather Friedberg, Diane Litman, and Susannah B. F. Paletz. 2012. Lexical entrainment and success in student engineering groups. In *Proceedings Fourth IEEE Workshop on Spoken Language Technology (SLT)*, Miami, Florida.

Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37:3–19.

Mahaveer Jain, John W. McDonough, Gahgene Gweon, Bhiksha Raj, and Carolyn Penstein Ros. 2012. An unsupervised dynamic bayesian network approach to measuring speech style accommodation. In *EACL*, pages 787–797.

Chi-Chun Lee, Athanasios Katsamanis, Matthew P. Black, Brian R. Baucom, Panayiotis G. Georgiou, and Shrikanth Narayanan. 2011. An analysis of pca-based vocal entrainment measures in married couples' affective spoken interactions. In *INTERSPEECH*, pages 3101–3104.

Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech*.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2013. Automated two-way entrainment to improve spoken dialog system performance. In *ICASSP*, pages 8372–8376.

Nichola Lubold and Heather Pon-Barry. 2014. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 5–12. ACM.

Nichola Lubold, Heather Pon-Barry, and Erin Walker. 2015. Naturalness and rapport in a pitch adaptive learning companion. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 103–110. IEEE.

Christopher Michael Mitchell, Kristy Elizabeth Boyer, and James C. Lester. 2012. From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue. In *SIGDIAL Conference*, pages 94–98.

Seungwhan Moon, Saloni Potdar, and Lara Martin. 2014. Identifying student leaders from mooc discussion forums through language influence. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 15–20.

Claude Nadeau and Yoshua Bengio. 2000. Inference for the generalization error. In *Advances in neural information processing systems*, pages 307–313.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 169–172.

Zahra Rahimi, Anish Kumar, Diane Litman, Susannah Paletz, and Mingzhi Yu. 2017a. Entrainment in multi-party spoken dialogues at multiple linguistic levels. *Proc. Interspeech 2017*, pages 1696–1700.

Zahra Rahimi, Diane Litman, and Susannah Paletz. 2017b. Acoustic-prosodic entrainment in multi-party spoken dialogues: Does simple averaging extend existing pair measures properly? In *International Workshop On Spoken Dialogue Systems Technology*.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Meeting of the Association of Computational Linguistics*, pages 808–815.

Tanmay Sinha and Justine Cassell. 2015. Fine-grained analyses of interpersonal processes and their effect on learning. In *Artificial Intelligence in Education: 17th International Conference*, pages 781–785.

Svetlana Stoyanchev and Amanda Stent. 2009. Lexical and syntactic priming and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 189–192, Stroudsburg, PA, USA. Association for Computational Linguistics.