Analysing Speech Patterns to Predict Agreement in Multi-Party interaction.

Speech Technology Project Group 19.
Target Grade: A

RHODRI MEREDITH
GUSTAV ENGELMANN

Speech Technology
Date: March 12, 2025

Supervisors: Jens Edlund, Ghazaleh Esfandiari
Examiner: Jens Edlund

Swedish title: Analysera talmönster för att förutsäga överenskommelse i flerpartsinteraktion.
Swedish subtitle: Talteknologi Projektgrupp 19

Contents

Abstract

This project explores the feasibility of predicting inter-speaker agreement in three-party conversations using speech pattern data from the MEET Corpus. Based on audio recordings of thirty participants working on object-ranking tasks under three different conditions (face-to-face, Zoom, hybrid), we extracted turn-related speech pattern features and quantified individual agreement levels by comparing personal rankings with group decisions. We then applied predictive modeling techniques, including linear models, classification models, and k-means clustering, to assess whether our selection speech patterns could effectively predict agreement. Our results indicate that the selected speech features did not manage to predict agreement for our corpus, suggesting that alternative approaches such as incorporating lexical or acoustic features may be necessary for more accurate agreement modeling.

Chapter 1

Introduction

1.1 Problem

In this project, we used speech pattern data from three-party meetings and investigated the feasibility of making model predictions about the level of inter-speaker agreement. We used audio recordings of 30 participants completing an object-ranking task in groups of three for three different conditions (face-to-face, on Zoom, hybrid) and extracted information about their speech patterns. We then quantified each participant's level of agreement with the group decision by comparing their own ranking with that of the group. Finally, we attempted to model this agreement level based on the speech patterns to see if it was possible to predict.

1.2 Motivation

With the results from this project, we intended to inform future research into inter-speaker agreement in group settings. Our project investigated the extent to which speech patterns could play a role in automatic agreement detection, data from which could prove to be invaluable when analysing the most significant contributing factors.

In doing this, we had also hoped to determine the features of speech patterns that are associated with agreement between participants. This information was used to inform our models and can be used as reference for models produced by future work.

Chapter 2

Background

2.1 Dataset

The data used in this project was collected and published by Ghazaleh Esfandiari-Baiat and Jens Edlund for the MEET Corpus [1][2] with the aim of studying the effect of different conditions in work meetings on joint decision making. It consists of audiovisual recordings of three-person group discussions, involving 30 participants between the ages of 23 and 48 of which 13 were female and 17 male. For three different meeting conditions, each group was asked to collaboratively rank a list of objects based on their importance in one of three different survival scenarios. The meeting conditions were as follows: (1) Collocated – all participants were in the same room; (2) Remote – all participants joined via Zoom; and (3) Hybrid – two participants were in the same room, while one joined via Zoom.

The discussions lasted between 10-18 minutes, amounting to 380 minutes in total for all recordings, and resulted in the group reaching an agreement on a final ranking of the 10 to 15 items, depending on the task, from most to least important. For on-site meetings, conversations were recorded using identical microphones and a 360-degree camera (Meeting Owl Pro). For online meetings, Zoom was used to record the session, along with each participant's personal computer camera and microphone. In all conditions, individual audio recordings were also captured in a similar manner.

The ranking tasks included the NASA moon survival task [3], the Desert survival task [4], and the Camping survival task [5]; both the order of these tasks as well as the conditions in which they were completed were randomised for each group. Participants were asked to provide their personal rankings both before and after the group discussion, with five minutes allocated for each response.

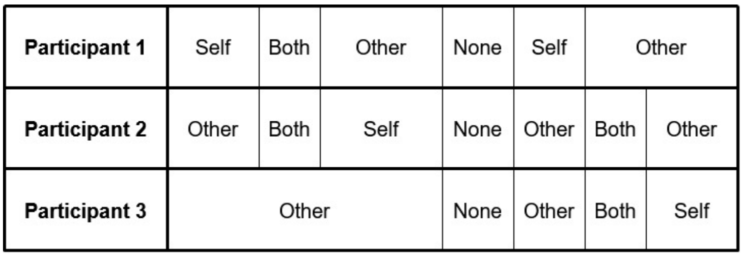
Our analysis will focus exclusively on the audio recordings, more specifically, the individual participant's recordings of each group for all of the conditions. The MEET corpus data have already been manually annotated for each participant's turns, focus, events, breathing, silence, and laughter/smile. A more detailed description of the annotations can be found in Ghazaleh and Edlund (2024) [1]. The relevant annotation for our project is that at the turn level.

2.2 Subjective Interaction Models

The starting point for feature extraction was with subjective interaction models. It was created from the annotated objective interaction model that was part of the raw data we had obtained. This had time-series annotations for each speaker, with a 0 when the speaker was silent and a 1 when they were speaking. For our purposes, we decided that a subjective interaction model (SIM) would provide a better description of the speech patterns that we hoped would predict agreement. A SIM has the following time-series encoding for each speaker:

- None (0): When nobody in the group is speaking
- Other (1): When anyone other than the speaker is speaking
- Self (2): When the speaker is speaking alone
- Both (3): When the speaker and any other participant is speaking simultaneously

Such an encoding would allow us to analyse whether the participant in question was speaking alone, speaking over someone or listening to others speak by only needing to look at the model for that participant. An example of the SIM plot is shown in figure 2.1.



<table>
<tr><td><b>Participant 1</b></td><td>Self</td><td>Both</td><td>Other</td><td>None</td><td>Self</td><td colspan="2">Other</td></tr>
<tr><td><b>Participant 2</b></td><td>Other</td><td>Both</td><td>Self</td><td>None</td><td>Other</td><td>Both</td><td>Other</td></tr>
<tr><td><b>Participant 3</b></td><td colspan="3">Other</td><td>None</td><td>Other</td><td>Both</td><td>Self</td></tr>
</table>

Figure 2.1: Subjective Interaction Model

2.3 Background Literature

The background literature for this project includes the proceedings of the ACL (Association for Computational Linguistics) workshop [2] that details the goal and creation of the MEET corpus [1]. The three ranking tasks employed in the collection of the corpus data are mentioned and discussed in [3],[4],[5] respectively. The agreement scores we tried to predict are either directly employing Kendall's measure of rank correlation [6] or using classifications derived from it. Previous work on agreement prediction uses a combination of different sets of characteristics, combining lexical, prosodic, and structural measures [7], [8], whereas our approach attempts to predict agreement solely from structural measures using the OIM and SIM.

Chapter 3

Method

3.1 Data processing

The data was processed from csv files in Python. All scripts used for the project are submitted along with this report. Processing the data consisted of two main areas: creating the SIMs and extracting the speech pattern features, and generating the agreement scores.

3.1.1 Processing speech data

Firstly, we took the objective interaction model that gives binary values for each speaker showing at what points in time they were speaking, and converted them into subjective interaction models for each speaker using a Python script. Each of the 90 SIMs as outlined in section 2.2 was saved in csv files. These were then used to extract the relevant speech pattern features which we would implement in our machine learning models.

Speech pattern features

From the SIM we extracted and quantified certain features for analysis:

1. **Turn duration:** The total amount of time a participant was speaking during the meeting. Calculated by summing up all time frames of **self** and **both** for every participant and dividing by the total amount of time frames of the respective sessions. This results in a percentage of how many turns each speaker occupies of all possible turns and was meant to give us an idea of how active participants were during the sessions.

2. **Overlaps:** The number of times the speaker speaks simultaneously with another speaker. Calculated similarly to turn duration by extracting the amount of turns annotated with **both** followed or preceded by **self** and dividing by the amounts of total turns. The resulting percentage was meant to inform on how much each speaker tended to speak at the same time as the other participants indicating the amount of speech interrupted by others.

3. **Back-channels to speaker:** The number of times other participants interject the speaker with back-channels (e.g. "hmm", "yeah", "okay"). Calculated by first establishing whether a speaker has the floor by setting a floor-taking-threshold and then counting the amount of other speakers talking during that time, and conforming to the set back-channel-duration. Several back-channels during a five-second period are counted as one. Both back-channel times and the full back-channel-to count were extracted for every participant.

4. **Back-channels by speaker** Using the back-channel times, we established the amounts that every speaker was back-channeling others. For both back-channeling features the raw numbers were extracted for the amounts. Back-channeling was intended to give us a measure of the feedback participants gave and received during the sessions.

5. **Skew:** A measure of the temporal distribution of the turns of the speaker during the meeting. Calculated by using the objective interactive model consisting only of ones and zeros, for not speaking and speaking, to obtain the Fisher-Pearson-Coefficient giving us a value between -1 and and 1 for each participant where values close to 1 translate to a skew of the speaker being active towards the beginning of the session and -1 to a skew towards the end of the session.

6. **Decision time:** Estimating the fraction of decisions that were verbally agreed upon by the speaker. We used the decision-time annotations in the elan file of the original recording for the ranking of each item and, where annotations existed, checked whether there was back-channeling happening in a 5 second period around the decision time and increased the decision score for both the person speaking during the decision time and the participants back-channeling. Then we calculated the percentage of decisions made of the total for each participant. This feature was supposed to approximate agreement with the individual rankings of the

items, although we are making the assumption that a back-channel while the decision is being made always means agreement.

3.1.2 Agreement scores

We then used the information on the rankings that were compiled in order to quantify the level of agreement that was reached by the three-party group. The data we had available to us were: individual rankings made before the group task (PRE), individual rankings made after the group task (POST) and the ranking that was decided upon by the group (CONSENSUS). To quantify the difference between any two rankings, we calculated the normalised Kendall Tau distance [6] between them. This takes each pair of items within one ranking, compares the order of that pair in both rankings and calculates the fraction of pairs that were in a different order in the two rankings. We deemed this method to be a good choice for our purpose since it is primarily based on pairwise comparisons of items in the list which corresponds to how humans typically operate when completing rankings tasks. The normalised Kendall Tau distance produces a result on a scale of 0 to 1, where 0 means the rankings were identical and 1 means the rankings were in the exact opposite order.
Since we had three rankings, there were three comparison scores that we could be made: Comparison of PRE and CONSENSUS rankings (referred to as Δ_1), comparison of PRE and POST rankings (referred to as Δ_2), and comparison of CONSENSUS and POST rankings (referred to as Δ_3). We decided that Δ_3 provided the most obvious choice for quantifying agreement within the group and therefore intended for this to be used in the Linear Model, but we calculated all 3 Δs to see if they could be useful in classification models or unsupervised models.

3.2 Models

Once all the data was processed and we had our features and Δs we began training our supervised and unsupervised learning models. All features and Δs were standardised before performing calculations.

3.2.1 Supervised learning methods

Linear Regression

Our first approach to a predictive model was linear regression analysis. Since it was not obvious which speech pattern features, if any, would have predictive power on the agreement scores, we decided to implement cross-validated LASSO regression with L1 regularisation, since this would also perform a subset selection and allow us to see the most predictive features. We had 90 samples of participants' speech data (10 groups of three participants in three conditions); 70 samples were used in training and 20 in testing. We reduced the L1 penalty (λ) until the model coefficients became non-zero and measured the loss from the training and test sets to evaluate the model.

Classification

Our second approach was to try to classify the participants based on their Δ scores and then use machine learning classification techniques to attempt to predict the class of test participants using the speech pattern features. We tried many different methods of classifying the participants. One method was to assign them a class of DOMINANT, PERCEPTIVE or INDEPENDENT. Definitions of these rankings are illustrated in table 3.1. The few participants whose lowest Δ was Δ_1 were included in the INDEPENDENT class.

Class	**Description**	**Delta definition**
DOMINANT	Those who mostly maintained their original rankings throughout and also imparted a large amount of their original rankings on to the group consensus	All three Δ scores below 0.4
PERCEPTIVE	Those who largely changed their minds on the rankings due to the group discussion	Lowest Δ score was Δ_3
INDEPENDENT	Those who maintained their original ranking and went along with the group consensus despite not agreeing with it much	Lowest Δ score was Δ_2

Table 3.1: Classification of Participants Based on Ranking Behavior

After having sorted the participants into these groups, a Random Forest classification model was run in order to test if it could be used to learn the classifications based on the speech pattern features. Another reason we chose this method was because it does not assume linearity of variables, a property which we had already tested. 72 of the 90 samples were used for training, and the rest were used for testing.

3.2.2 Unsupervised learning methods

K-means clustering

Our final approach at producing a model with the data was with unsupervised learning techniques. Our intention was to use machine learning techniques to find a pattern in the speech pattern features, and then find a connection between that pattern and the agreement scores. We decided to look at all 3 Δ scores for this method. We used k-means clustering to cluster the speech data and then performed one-way analysis of variance (ANOVA) in order to assess a connection between the clusters and Δ values.

Chapter 4

Results

4.0.1 Data processing

The Python code that we wrote to extract the speech pattern features and agreement scores worked successfully. Table 4.3 shows 4 of the features from participants of one meeting along with their agreement scores (Δ_3). All values were z-score standardised before applying them to our models. All Δs were shown to be normally distributed as evidenced in figure 4.4.

	Turn length	Speaker BC	Skew	Decision score	Δ_3
Participant A	0.393	3	0.166	0.444	0.244
Participant B	0.176	9	0.267	0.444	0.178
Participant C	0.21	6	0.269	0.111	0.333

Table 4.1: Example dataset from the participants of one meeting with a few selected features and their agreement scores. BC refers to back-channels.

4.1 Supervised Models

4.1.1 Linear Model

The LASSO regression model we implemented had very poor predictive powers. This is demonstrated in figure 4.1. In this plot, the mean squared error of the training and test loss is 1 when the regularisation term is high. This is to be expected since the regularisation term is shrinking the coefficients to 0, so the model will simply be equal to the mean of the standardised agreement scores. When this shrinkage was reduced by lowering the penalty term, the coefficient values became non-zero but this was accompanied by a sharp increase in the mean squared error in the test data. This is highly indicative of over-fitting on the training data and shows that none of the features have any predictive power on the agreement scores.

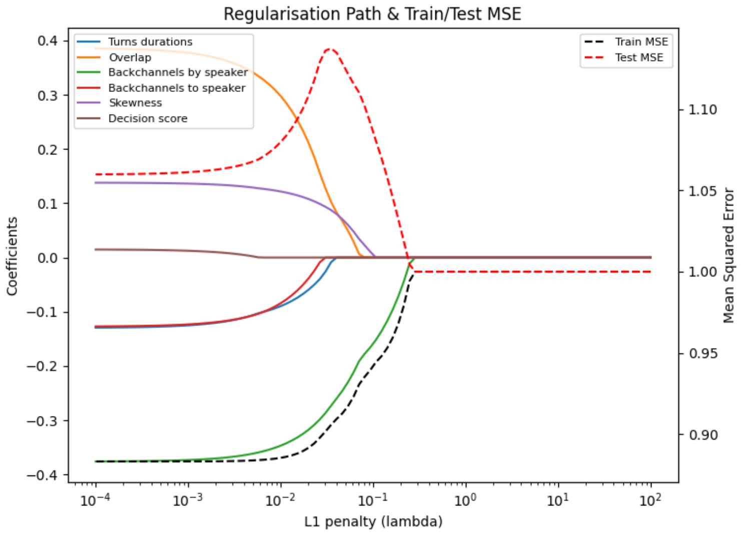


Figure 4.1: Regularisation path for LASSO regression along with train and test loss for the model.

4.1.2 Classification Models

The report from the Random Forest classification is shown in table 4.2. The model produced an accuracy of 39%, which is the amount which would be expected if the model had guessed DOMINANT for all participants. Therefore the model does not have any predictive power on the categories.

4.2 Unsupervised Model

4.2.1 K-means Clustering

We first ran an elbow test to determine an ideal number of clusters. The results proved fairly inconclusive as figure 4.2 shows a minimal "elbow" shape the curve.

Class	Precision	Recall	F1-score	Support
DOM	0.33	0.14	0.20	7
IND	0.33	0.40	0.36	5
PERC	0.44	0.67	0.53	6
Accuracy	0.39 (18 samples)			
Macro Avg	0.37	0.40	0.37	18
Weighted Avg	0.37	0.39	0.36	18

Table 4.2: Classification report of the Random Forest model

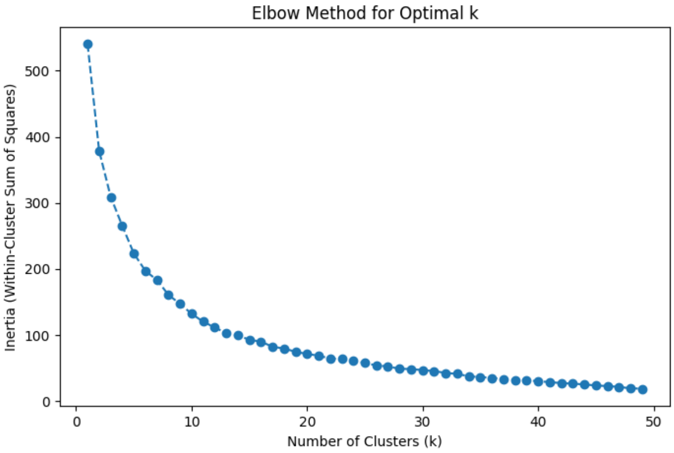


Figure 4.2: Elbow method for finding optimal number of clusters

Nevertheless, we decided to continue with 10 clusters based on the slight levelling-off of the elbow curve at this point. After running the k-means clustering, we assessed whether the resulting clusters had any relation to any of the Δ values, as a last attempt to uncover a pattern in our results so far. The boxplots shown in figure 4.3 show these clusters plotted against each Δ. They show no pattern among the clusters in relation to Δ values and demonstrate that there is no link between them.

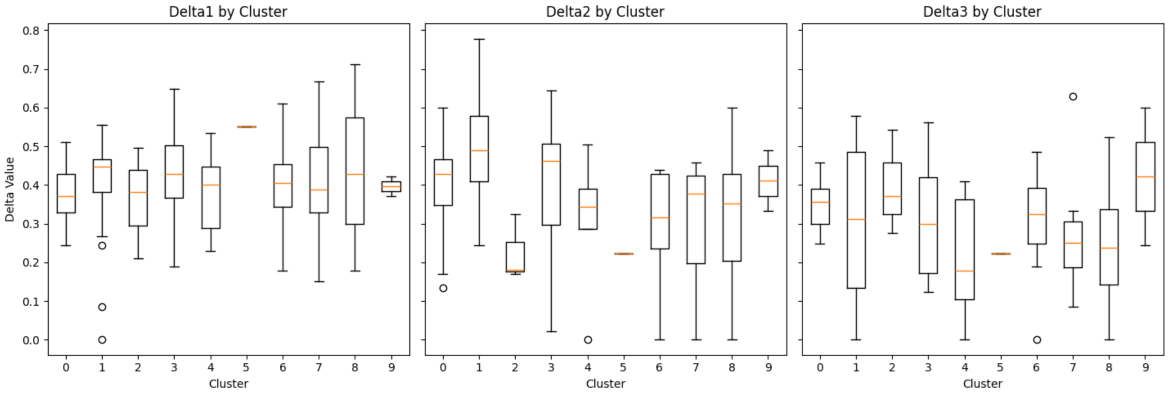


Figure 4.3: Box plots of clusters from k-means clustering against Δ values. Red line shows the mean of the cluster. Outliers are shown as circles outside the box and whiskers.

To quantify our findings, we confirmed that the Δs were normally distributed (figure 4.4) and performed a one-way ANOVA, the results of which are shown in table 4.3. The F-statistics and p-values from the one-way ANOVA were not sufficient to suggest any connection between the clustering and the Δ values. Given that the elbow method showed that the clustering itself was already tenuous, this was enough for us to conclude that there was no connection to be found between the speech pattern features and agreement score.

	F-statistic	p-value
Δ_1	0.357	0.952
Δ_2	0.719	0.690
Δ_3	0.739	0.672

Table 4.3: Table showing the results from the one-way ANOVA, assessing whether the k-means clustering model can be used to predict any of the Δ scores.

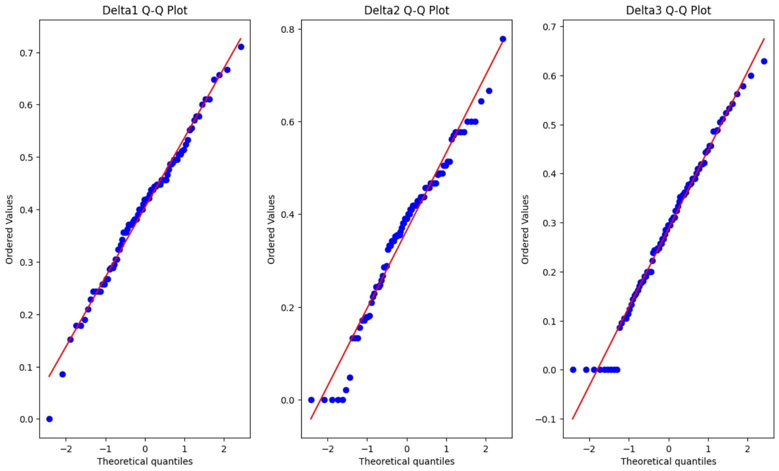


Figure 4.4: Q-Q plots for ∆s

Chapter 5

Discussion

5.1 Conclusions

Considering the results of the linear model, the classification models and the k-means clustering, the conclusion we can draw is that the features we selected cannot be used in the way we attempted to help predict inter-speaker agreement. This shows that incorporating interaction models in the form we did in this project into an agreement prediction model would not help its accuracy; it would most likely be better to use a different approach or combine SIM features with additional lexical or acoustic features when creating such a model.

5.2 Project Limitations

5.2.1 Time and group size

We initially started this project as a group of three, but after finalizing the final project bid, a team member withdrew from the master's program and, consequently, from the project. After consulting with the course coordinator, Jens Edlund, we adjusted our strategy as a team of two, which led to a more focused selection of features and models. Additionally, we faced a delay in receiving the MEET data, which arrived on March 3rd, eleven days after the scheduled project start on February 20th. As a result, our execution was more rushed than planned.

5.2.2 Data

Our initial hypothesis of being able to predict agreement rather accurately was made based on limited knowledge of the MEET corpus. Upon a closer examination and analysis of the data it became apparent that the features we used to predict agreement in our models resulted in a significantly lower prediction accuracy than expected. Given the availability of the data before the formulation of our hypothesis and selection of features used in the modeling, a more informed hypothesis and careful selection of features could have been developed. Variability in participants' conversational patterns across the different conditions introduced inconsistencies in the data used for training our models that potentially affected their performance.

5.2.3 Choice of features

The evaluation of our models' results showed that the features we extracted from the SIM were not suitable for accurately predicting the agreement scores. A similar approach based on the SIM would have to introduce other speech pattern features and focus on conversations recorded in similar conditions to improve agreement prediction. A combination of these speech pattern features with intonation, emphasis, and pitch change analysis during crucial moments of the conversation, for example back-channels and decision times, or a qualitative analysis of positive and negative speech markers could enhance the accuracy of agreement prediction.

References

[1] G. Esfandiari-Baiat and J. Edlund, "The MEET corpus: Collocated, distant and hybrid three-party meetings with a ranking task," in *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, H. Bunt, N. Ide, K. Lee, V. Petukhova, J. Pustejovsky, and L. Romary, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 1–7. [Online]. Available: https://aclanthology.org/2024.isa-1.1/

[2] H. Bunt, N. Ide, K. Lee, V. Petukhova, J. Pustejovsky, and L. Romary, Eds., *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*. Torino, Italia: ELRA and ICCL, May 2024. [Online]. Available: https://aclanthology.org/2024.isa-1.0/

[3] J. Hall and W. H. Watson, "The effects of a normative intervention on group decision-making performance." *Human Relations*, vol. 23, no. 4, pp. 299–317, 1970. doi: 10.1177/001872677002300404. [Online]. Available: https://doi.org/10.1177/001872677002300404

[4] J. C. Lafferty and A. W. Pond, *The Desert Survival Situation: Manual: a Group Decision Making Experience for Examining and Increasing Individual and Team Effectiveness*. Human Synergistics, 1974.

[5] A. P. Hare, "A study of interaction and consensus in different sized groups," *American Sociological Review*, vol. 17, no. 3, pp. 261–267, 1952. [Online]. Available: http://www.jstor.org/stable/2088071

[6] M. G. KENDALL, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 06 1938. doi: 10.1093/biomet/30.1-2.81. [Online]. Available: https://doi.org/10.1093/biomet/30.1-2.81

[7] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of bayesian

networks to model pragmatic dependencies," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004, pp. 669–676.

[8] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation," in *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009, pp. 7–14.