INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Privacy-Preserving Speech Analytics for Automatic Assessment of Student Collaboration

Nikoletta Bassiou[1], *Andreas Tsiartas*[1], *Jennifer Smith*[1], *Harry Bratt*[1], *Colleen Richey*[1], *Elizabeth Shriberg*[1], *Cynthia D'Angelo*[2], *Nonye Alozie*[2]

[1]SRI International Speech Technology and Research (STAR) Laboratory
[2]SRI International Center for Technology in Learning (CTL)

{nikoletta.basiou, andreas.tsiartas, jennifer.smith, harry.bratt, colleen.richey, elizabeth.shriberg, cynthia.dangelo, maggie.alozie} @sri.com

Abstract

This work investigates whether nonlexical information from speech can automatically predict the quality of small-group collaborations. Audio was collected from students as they collaborated in groups of three to solve math problems. Experts in education annotated 30-second time windows by hand for collaboration quality. Speech activity features (computed at the group level) and spectral, temporal and prosodic features (extracted at the speaker level) were explored. After the latter were transformed from the speaker level to the group level, features were fused. Results using support vector machines and random forests show that feature fusion yields best classification performance. The corresponding unweighted average F_1 measure on a 4-class prediction task ranges between 40% and 50%, significantly higher than chance (12%). Speech activity features alone are strong predictors of collaboration quality, achieving an F_1 measure between 35% and 43%. Speaker-based acoustic features alone achieve lower classification performance, but offer value in fusion. These findings illustrate that the approach under study offers promise for future monitoring of group dynamics, and should be attractive for many collaboration activity settings in which privacy is desired.

Index Terms: speech analytics, speech activity detection, spectral, temporal and prosodic features, machine learning, student collaboration, collaborative learning, classroom education

1. Introduction

This study is part of a new multi-year project that aims to build privacy-preserving speech-based analytics for the automatic assessment of multi-student collaboration in a school setting. Collaboration is an important 21st-century skill that students must be able to master as they progress through school and beyond [1]. Research has shown that students need feedback in the school environment to develop collaboration skills. Many do not come to class with experience in how to engage with their peers in collaborative activities and how best to work together productively in groups [2].

Teacher assessment of group collaboration is a challenge in today's classrooms, since class size typically makes it infeasible for a single teacher to monitor a large number of small groups simultaneously [3]. The ultimate goal of the project is to produce knowledge about the feasibility of speech analytics and the creation of adaptive software that could help teachers by identifying groups that need feedback in real time, as well as by helping teachers to better target their interventions.

Information from speech is a key knowledge source for the effort, since collaborative learning in classrooms usually takes place through natural language. Although there are many approaches (e.g., keystroke data, written responses) for gathering diagnostic information about collaborative learning, most collaborative learning involves peer discourse. Automated analysis of peer discourse in collaborative learning has been successful [4, 5, 6], but most past work has focused on non-spoken modalities, e.g. using chat rooms [4].

Though speech data is uniquely central and authentic to peer discourse, the field does not yet have key knowledge of automatically analyzed speech in small group collaboration. Some exploratory work has successfully developed speech analytics for a situation in which one student is asked to answer a question while on camera [7]. Other researchers have taken a different approach that tries to apply speech analytics to very specific and sophisticated aspects of collaborative learning, such as idea co-construction [8] and transactive contributions [9].

This project focuses on simpler behaviors in collaborative situations. To preserve privacy, which is a key issue when working with children, no words and no video signals are used. The setup uses non-lexical features only, is lightweight and requires only basic equipment (microphones). Furthermore, there is no dependency on automatic word recognition, which is a current challenge in the context of the classroom setting.

In a first exploration using a subset of the new corpus [10, 11], we found that features that capture when each participant speaks, as well as how each participant speaks, are good predictors of collaboration quality. In this study, we analyze the full collected data set, and explore a wider range of group speech activity features and prosodic, spectral and temporal features. We also investigate how to fuse features that are taken from the group with those taken from individual talkers, and we explore a range of classifiers for the prediction task.

2. Data Collection

Collaborative math activities included 12 separate math problems. Participating students, organized in groups of three, had to work together and talk to each other to coordinate their three answers to the problems. 141 middle school students (67 in sixth grade, 40 in seventh grade, and 34 in eighth grade) from six different schools participated in the study. The gender breakdown was evenly split across the students.

The data was collected during 86 collaborative sessions, each lasting about 15-20 minutes. Most students participated in 2 sessions with different group configurations. In each ses-

http://dx.doi.org/10.21437/Interspeech.2016-1569

sion, each group was recorded by video, and audio recordings were collected using individual noise-cancelling microphones worn by each student. These audio recordings were divided into segments that corresponded to the time that the group spent on a particular math problem (items). The items were further divided into 30-second windows. Depending on the length of the item, some windows were less than 30 seconds. Windows less than 5 seconds long were discarded. In total, there were 866 items and 2942 windows.

Data at the levels of item and window were annotated by a team of five education researchers. In order to ensure reliability on the annotations, all annotators were trained on the coding scheme and went through a calibration process. The average of the Cohen's kappa score [12] for each pair of judges across four sessions was 0.612 after training, which is an acceptable level of reliability for this type of annotation task [13]. During the annotation process, additional calibration instances were selected to prevent significant drift on the application of the codes. All disagreements were discussed by the annotators and a final code was assigned. The annotators had to assign one of four collaboration quality codes (Q codes). The Q codes represented the degree to which the three students of the group were collectively engaging in good collaboration. It should be noted that the codes depend on whether and how much each student was intellectually engaged in the group problem solving, and not on simply the duration of each student's speech. More successful collaboration occurs when students engage each others' thinking [14]. In other words, the collaboration quality codes differentiated between simple engagement (whether or not students were talking and paying attention) and intellectual engagement (whether or not the students were engaged in actively solving the problem at hand). The annotators made their decisions based on both the audio and the video recordings. The Q codes are defined as follows:

- Good Collaboration ("Good"): All three students are working together and intellectually contributing to problem solving.
- Out in the Cold ("Cold"): Two students are working together, but the third is either not contributing or is being ignored.
- Follow the Leader ("Follow"): One student is taking the intellectual lead on solving the problem and is not bringing in others.
- Not Collaborating ("Not"): No students are actively contributing to solving the problem; each is either off-task, or working independently.

The distribution of the Q codes assigned at the window level is 0.34 for the "Good Collaboration" class, 0.27 for the "Out in the cold" class, 0.21 for the "Follow the leader" class and 0.18 for the "Not Collaborating" class.

3. Features

3.1. Speech activity features

During the data collection and experimental setup, students were recorded by individual noise-cancelling microphones. As a result, a separate audio-channel was collected for each student in the group. Also, students were allowed to speak freely resulting in audio recordings that exhibit overlapping speech from the three students. To overcome this problem, a Speech Activity Detection (SAD) system was used to identify the speech regions and exclude the silent and noisy regions. This SAD system, which was based on a speech variability threshold optimized on a small set of four samples [15, 16], was run independently on each of the 3 student channels. The thresholded output on each audio channel was used to identify the student-specific speech signal and eliminate the noise, silence or cross-talk regions.

The features derived from SAD output capture information about the amount, duration, and location of speech regions, much like the features used in studies of dominance in multiparty meetings [17, 18]. However, the features we extracted differ. In detail, several duration-related statistics were created using the SAD output. These features are the total duration of speech for each student ("Total Duration 1", "Total Duration 2", and "Total Duration 3"), the duration in which each student was the only speaker ("Solo Duration 1", "Solo Duration 2", and "Solo Duration 3"), the duration of overlapping speech from each pair of students ("Overlap Duration 1-2", "Overlap Duration 1-3", and "Overlap Duration 2-3"), the duration of overlapping speech among the three students ("All Duration"), and the duration in which all students were silent ("No Duration").

From these SAD-derived statistics, only "All Duration" and "No Duration" could be used directly as group-level features, since they characterize the whole group. The remaining sets of features (three each for Total, Solo and Overlap Durations) reflect the SAD activity for individual speakers or speaker pairs. In order to obtain group-level features for these sets, each of the three statistics in each set was converted to proportions $p(x)$ by dividing them by their sum. Then, the distribution of each set was estimated by means of the Shannon Entropy [19]:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (1)$$

In our case, there are 3 speaker-level measurements per set. Thus, the entropy values range between 0 and $\log_2 3 \approx 1.585$. The minimum value indicates a window during which only one of the students (or overlapping pairs) speaks, while the maximum value indicates a window during which all three students (or overlapping pairs) are speaking equally.

Since only minimum and maximum entropy values have a clear interpretation in this context, we created another type of group-level feature to capture the relationship between speaker durations: ratio statistics. "Ratio 1" is computed by dividing the second most talkative student (or pair) by the most talkative student (or pair). "Ratio 2" is computed by dividing the least talkative student (or pair) by the most talkative student (or pair). These ratios can be interpreted as the relative duration of the second most and least talkative students (or pairs) relative to the most talkative student.

3.2. Spectral, temporal and prosodic features

In our data collection setup, students were allowed to speak freely and one of our core goals has been to capture a diverse set of speech features, such as spectral, temporal, prosodic and tonal. We aimed to use a such diverse set of features to have a holistic view of which major categories of speech features are indicators of good collaboration. In addition, we extracted all speech features for each student independently. The frame shift of the features is 30ms and the window varies from $20 - 40$ms. The frame level features include the Mel frequency cepstral coefficients (MFCC) [20]. The MFCC represent the cepstral information of the signal. Other spectral and energy-based features which were computed include the energy of 4 frequency bands as features, the time-domain energy, and the statistics of the spectrum (mean, variance, kurtosis and skewness). Noise-robust features include the RASTA features, which filter invari-

ant and rapidly variant noise types. Furthermore, we included features which capture the harmonic content of the signal, such as harmonicity and voicing. For the last two features, we used zero crossing rate and chroma, which measure the dominant formants and tonality of the speech signal. To extract the features, we used a variety of open source [21, 22] and SRI-owned tools. Finally, after the frame features were extracted, the features were averaged at the segment level. All features were speaker-normalized by subtracting the speaker mean over the session. Unlike the speech activity features described in Section 3.1, these speaker-based extracted features are "blind" to the prosodic activity and speech characteristics of the other participants. This approach provides a real-time processing advantage, but the performance is expected to be suboptimal since the features from each individual speaker contain no information about the behaviour of the other speakers.

3.3. Feature fusion

Speaker-based features were also combined with speech activity features by means of early fusion. To achieve this, the spectral, temporal and prosodic features had to be transformed from the speaker level to the group level. To this end, three different approaches that map these features to the group level were proposed:

- Entropy-based mapping: The distribution of each speaker-level feature was combined by means of the Shannon entropy [19], as described in Section 3.1.
- SAD-ordered based mapping: The features extracted from each speaker of the group were stacked into a single feature vector by taking into account the duration of speech of each speaker in the group. That is, the speakers of each group were sorted based on their speech duration within each window, and their corresponding features were then stacked based on this ordering. In this sense, the features at the group level are comprised of the feature values for the most talkative speaker within the window, followed by the feature values for the second most talkative speaker, followed by the feature values for the least talkative speaker within the same window.
- MinMax-ordered based mapping: This approach is similar to the previous one in the sense that the speaker-level features for speakers of the same group are stacked, but the stack ordering is determined by the raw feature values. That is, the features at the group level are comprised of the maximum feature value, followed by the second maximum features values within the window, followed by the minimum feature values within the same window.

These transformation approaches attempt to capture a variety of dynamics of speech characteristics within the group. For example, regarding loudness, Entropy-based mapping can differentiate between a group in which one student is much louder than the rest and a group of equally loud students. Similarly, SAD-ordered mapping and MinMax-ordered mapping capture the loudness of the least-talkative student and the quietest student, respectively.

4. Classification

The dataset was partitioned into a development set and a held-out set. Special care was exercised to prevent speaker overlap between these two sets. 70% of the data were used in the development set, and 30% of the data were used in the held-out set. The development set was used for tuning the parameters and training the classifiers, while the held-out set was used for the assessment of the classification performance.

Classification was performed by employing two different types of classifiers: support vector machines (SVMs) [23] and random forests [24]. In our earlier work on a subset of the data and features [11], tree-based classifiers yielded good classification results. In this experimental setup, we also added SVMs that are known to give good results in many complex classification tasks. For SVMs, a Radial Basis Function (RBF) kernel was used with three different values for the kernel parameter, $\gamma = 0.1, 0.01, 0.001$. For random forests, experiments with $10, 20, 50, 100, 500$ and 1000 estimators were performed using the information gain as a measure of quality for each split. Additionally, automatic feature ranking and selection was performed by means of a Recursive Feature Elimination (RFE) procedure. Initially, the estimators were trained on the full set of features. At each iteration step, a number of features were removed until a pre-selected number of features was reached. Based on the higher unweighted F_1 measure estimated across a 10-fold cross-validation scheme on the development set, the best classifier with its optimal parameters and the optimal number of features were selected. As before, during cross-validation, folds were created so that no speakers were present in both train and test set partitions.

5. Results and Discussion

Group-level features based on speech activity were comprised of 2942 datapoints with 20 dimensions each. Spectral, temporal and prosodic features were extracted at the speaker level. Since there were 3 speakers per group, there were 3 times as many datapoints (8826 datapoints) with 138 feature dimensions each. After applying the transformation to the group level, the resulting features consisted of 2942 datapoints with 138 dimensions each when the entropy-based fusion was used, and with 414 dimensions each when the other two fusion methods were used.

Initially, a set of classification experiments was conducted using a first subset of 5 extracted speech activity features. These features are: "All Duration", "No Duration", and the entropy statistic for "Total Duration", "Solo Duration" and "Overlap Duration" features. Then, all the 20 speech-activity features were included in a second set of experiments. In the rest of the paper, we refer to these two sets of experiments as Experiment I and Experiment II, respectively.

Classification performance was evaluated by estimating both the accuracy and the F_1 measure in the held-out set. These results are presented in Tables 1 and Tables 2 for Experiments I and II, respectively. Results are shown at both the class level (Q codes) and across classes by means of unweighted averages that account for the performance of each class equally. Best performance across each line is shown in bold for F_1 and accuracy. SAD features alone are better predictors than the temporal, spectral and prosodic features alone. In terms of overall F_1, SAD features outperform the speaker-based extracted features by 7.8% in Experiment I and by 16.1% in Experiment II. Similarly, SAD features achieve a higher overall accuracy than the speaker-based extracted features by 10.4% and by 16% in Experiments I and II, respectively. This was expected because classification using the speaker-based features uses information from individual speakers to predict group-level labels. However, these features also seem promising, since they show the ability to predict the "follow" and "not" classes. In Experiment I, the accuracy of the "follow" class is higher by 8.7%

Table 1: *Per-class and overall unweighted (UW) F_1 and accuracy values when only spectral, temporal and prosodic (S/T/P) features are used; when only SAD features are used; and when the fused features are used (Experiment I).*

	S/T/P (speaker-level)		SAD (group-level)		Entropy-based fusion		SAD-ordered fusion		MinMax-ordered fusion	
Q Code	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy
Good	40.3%	43.2%	52.5%	**73.4%**	**65.0%**	67.0%	64.9%	70.2%	53.8%	64.5%
Cold	27.8%	27.3%	46.5%	48.0%	**51.0%**	58.4%	45.8%	56.4%	49.2%	**58.8%**
Follow	21.8%	22.5%	21.0%	13.8%	**31.1%**	**26.5%**	24.6%	18.9%	26.5%	19.0%
Not	19.2%	16.5%	21.3%	15.8%	31.6%	**27.9%**	**47.5%**	41.9%	27.0%	20.9%
UW average	27.3%	27.4%	35.1%	37.8%	44.7%	45.0%	**45.7%**	**46.9%**	39.1%	40.8%

Table 2: *Per-class and overall unweighted (UW) F_1 and accuracy values when only spectral, temporal and prosodic (S/T/P) features are used; when only SAD features are used; and when the fused features are used (Experiment II).*

	S/T/P (speaker-level)		SAD (group-level)		Entropy-based fusion		SAD-ordered fusion		MinMax-ordered fusion	
Q Code	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy	F_1	Accuracy
Good	40.3%	43.2%	**71.2%**	**85.0%**	70.5%	73.3%	64.6%	72.9%	66.4%	73.3%
Cold	27.8%	27.3%	**50.7%**	44.6%	48.5%	54.7%	48.0%	**61.3%**	48.9%	48.6%
Follow	21.8%	22.5%	38.2%	36.0%	**41.3%**	**37.5%**	33.2%	23.9%	29.2%	29.7%
Not	19.2%	16.5%	13.5%	8.0%	**40.2%**	**31.5%**	28.6%	21.2%	25.8%	19.0%
UW average	27.3%	27.4%	43.4%	43.4%	**50.1%**	**49.3%**	43.6%	44.8%	42.6%	42.7%

when the speaker-level S/T/P features are used instead of the SAD features, while in Experiment II, the S/T/P features yield a higher accuracy by 8.5% for the "not" class compared to the SAD features. This promising performance of the speaker-level features is further verified by the fusion results. In detail, when the speaker-based features are combined with the SAD features in Experiment I, the latter's performance in terms of unweighted F_1 is improved by 9.6% for the entropy-based fusion, by 10.6% for the SAD-ordered fusion, and by 4% for the MinMax-ordered fusion. For Experiment II, there is a gain of 6.7% in F_1 when the Entropy-based fusion is used, while the other two fusion methods do not seem to contribute towards improving the initial SAD features performance. This implies that when the SAD features are not powerful predictors on their own (as in the case of Experiment I), they can be enhanced by speaker-based features. This is important, since the extraction of speaker-based features is straightforward and independent of the group information. It is also worth noting that all the results are well above chance performance when a "brute force" method is used that assigns all samples to the label with the most frequent class (i.e., "good"). The unweighted F_1 in this case is 12.2%.

The best results for the SAD features alone were derived when SVMs were used for classification. For Experiment I, the SVM kernel parameter is $\gamma = 0.01$ and the optimal number of selected features is 4 out 5. For Experiment II, $\gamma = 0.001$ and 14 out of 20 features are kept. In the case of spectral, temporal and prosodic features, the best results are obtained with random forests employing 100 estimators. The optimal number of features in this case is 48 out of 138. The best results for the fusion methods are derived with random forests employing 1000 estimators. When fusion with the SAD subset of Experiment I is applied, the optimal number of features is 12 out of 143 for the entropy-based approach, 29 out of 419 for the SAD-ordered method and 19 out of 419 for the MinMax-ordered approach. The corresponding optimal numbers of features for Experiment II are: 35 out of 158 for the entropy-based approach, 69 out of 434 for the SAD-ordered approach, and 26 out of 434 for the MinMax-ordered based approach. It is also worth mentioning that in all fusion approaches the top ranked 4 features in Experiment I are SAD features. In Experiment II, 17 SAD features are included in the optimal features, and most of them are ranked higher. This observation further supports our expectations, which were also validated by the classification results. That is, the SAD features alone are better collaboration predictors than the prosodic features alone, since they are directly extracted on the group level in contrast to the speaker-level S/T/P features which are agnostic to the group information. The complimentary power of the S/T/P features was also validated.

6. Conclusions

We studied the automatic prediction of collaboration quality among students by exploiting group-based durational statistics and speech analytics. Speech activity features were estimated on the group level and spectral, temporal and prosodic features were extracted at the speaker level. The combination of the two types of features was also investigated by employing three different approaches for mapping the speaker-level features to the group level. Results reveal that both speech activity features and speaker-based features are good predictors of collaboration quality, while their combination by means of fusion can considerably improve their collaboration prediction performance. Results demonstrate that privacy-preserving automatic speech features offer promise for future applications that can monitor multiple groups simultaneously for collaboration quality. Future work will focus on examining whether the proposed approach can detect specific features of participation such as turn-taking, crosstalk, emotion and off-task behaviors. We will also work with more fine-grained annotations for collaboration prediction. To this end, a wider range of features and modeling approaches will be investigated. Also, prediction using lexical features will be explored. Finally, the utility of the automatic feedback for teachers in the classroom will be investigated.

7. Acknowledgements

We gratefully acknowledge the contributions and support of Diana Jang, Erik Kellner, Tiffany Leones, Tina Stanford, Jeremy Fritts and Jeremy Roschelle. This material is based upon work supported by the National Science Foundation under Grant No. DRL-1432606. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. References

[1] N. R. Council, *Assessing 21st century skills: Summary of a workshop*. Washington, D.C.: The National Academies Press, 2011.

[2] G. Ladd, B. Kochenderfer-Ladd, K. Visconti, I. Ettekal, C. Sechler, and K. Cortes, "Grade-school childrens social collaborative skills: Links with partner preference and achievement," *Am. Educ. Res. J.*, vol. 51, no. 1, pp. 152–183, 2013.

[3] E. Cohen, "Restructuring the classroom: Conditions for productive small groups," *Rev. of Educ. Res.*, vol. 64, no. 1, pp. 1–35, 1994.

[4] G. Erkens and J. Janssen, "Automatic coding of dialogue acts in collaboration protocols," *Int. J. Comput. Collab. Learn.*, vol. 3, no. 4, pp. 447–470, 2008.

[5] B. McLaren, O. Scheuer, M. D. Laat, R. Hever, R. D. Groot, and C. Rosé, "Using machine learning techniques to analyze and support mediation of student e-discussions," *Frontiers in Artificial Intelligence and Applications*, vol. 158, no. 1, pp. 331–338, 2007.

[6] C. Rosé, Y. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer, "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," *Int. J. Comput. Collab. Learn.*, vol. 3, no. 3, pp. 237–271, 2008.

[7] M. Worsley and P. Blikstein, "Towards the development of multimodal action based assessment," in *LAK 13 Proc. Third Int. Conf. Learn. Anal. Knowl.*, 2013, pp. 94–101.

[8] G. Gweon, P. Agrawal, M. Udani, B. Raj, and C. Rosé, "The automatic assessment of knowledge integration processes in project teams," in *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL 2011 Conference Proceedings - Long Papers*, vol. 1, 2011, pp. 462–469.

[9] G. Gweon, M. Jain, J. McDonough, B. Raj, and C. Rosé, "Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation," *Int. J. Comput. Collab. Learn.*, vol. 8, pp. 245–265, 2013.

[10] C. Richey, C. D'Angelo, N. Alozie, H. Bratt, and E. Shriberg, "The SRI speech-based collaborative learning corpus," in *INTERSPEECH 2016 – 17^{th} Annual Conference of the International Speech Communication Association, Proceedings*, San Francisco, California, USA, September 8-12, 2016.

[11] J. Smith, H. Bratt, C. Richey, N. Bassiou, E. Shriberg, A. Tsiartas, C. D'Angelo, and N. Alozie, "Spoken interaction modeling for automatic assessment of collaborative learning," in *Speech Prosody 2016, Proceedings*, Boston, USA, 2016, pp. 277–281.

[12] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 36–46, 1960.

[13] S. Stemler, "An overview of content analysis," *Practical Assessment, Research & Evaluation*, vol. 7, no. 17, pp. 137–146, 2001.

[14] D. Kuhn, "Thinking together and alone," *Educ. Res.*, vol. 44, no. 1, pp. 46–53, 2015.

[15] A. Tsiartas, T. Chaspari, N. Katsamanis, P. Ghosh, M. Li, M. V. Segbroeck, A. Potamianos, and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *INTERSPEECH 2013 – 14^{th} Annual Conference of the International Speech Communication Association, Proceedings*, Lyon, France, August 25-29, 2013, pp. 718–722.

[16] G. Kumar, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 19, no. 3, pp. 600–613, 2011.

[17] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 19, no. 4, pp. 847–860, 2011.

[18] R. Rienks and D. Heylen, "Dominance detection in meetings using easily obtainable features," *Machine Learning for Multimodal Interaction*, vol. 3869, no. 76-86, 2006.

[19] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*, Florence, Italy, October 2010, pp. 1459–1462.

[22] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[23] B. Schölkopf, C. Burges, J. Christopher, and A. Smola, *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press, 1999.

[24] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.