Feature Selection and Boosting Methods for Prediction of Cognitive Load from Acoustic Data

Jing Su

School of Computer Science and Statistics, Trinity College Dublin

sujing@tcd.ie

Abstract

An analysis of acoustic features for a ternary cognitive load classification task and an application of a classification boosting method to the same task are presented. The analysis is based on a data set that encompasses a rich array of acoustic features as well as electroglottographic (EGG) data gathered for the COMputational PARalinguistic ChallengE (ComParE 2014). Supervised and unsupervised methods for identifying constitutive features of the data set are investigated with the ultimate goal of improving prediction. Our experiments show that the different tasks used to elicit the speech for this challenge affect the acoustic features differently in terms of their predictive power and that different feature selection methods might be necessary across these sub-tasks. The sizes of the training sets are also an important factor, as evidenced by the fact that the use of boosting combined with feature selection was enough to bring the unweighted recall scores for the Stroop tasks well above a strong support vector machine baseline.

Index Terms: Paralinguistic information, cognitive load modelling, feature selection, classification.

1. Introduction

Non-verbal and paralinguistic characteristics of speech have received increasing attention from the research community. It is now commonly accepted that non-verbal sounds form an important part of human communication [1], and that non-verbal features may help identify important structural aspects of speech interaction in real-life [2] as well as laboratory settings [3] for use in, for instance, browsers for multiparty interaction data [4, 5]. A more recent trend in the use of paralinguistic features is their analysis for predicting levels of cognitive and physical workload. Determination of workload levels is relevant in fields such as ergonomics, where it could help improve human computer interaction [6]. While most research in this field has been based on neurophysiological measuring, which involves specialised and intrusive equipment, the use of voice features for assessment of cognitive and physical load levels seems promising, for instance, for speech-based human computer interaction.

This year's COMputational PARalinguistic ChallengE (ComParE) was set up to provide a focus for research on approaches to the identification of cognitive and physical load levels through acoustic features [7]. The challenge provides two separate datasets: one for cognitive load (CLSE) and another for physical load (MBC). The work reported in this paper focuses on the former, which contains both speech and electroglottographic data. This challenge consists of inferring the correct cognitive load level out of three possible levels. We investigate unsupervised and supervised approaches to this challenge.

A brief inspection of the CLSE dataset reveals a ratio of approximately 3.5 features (instance descriptor attributes) to instances. The training set, for instance, has 6,376 features for 1,838 instances. A limited number of samples distribute sparsely in the high dimensional space, which makes it hard to separate the three levels of cognitive load. Therefore, our first concern was to reduce the high dimensionality of the feature set. Principle Component Analysis (PCA, [8]) is a popular approach to dimensionality reduction by feature extraction through reconstruction of the original feature space. The first principle component represents the most prominent axis of variance in the feature space, while the other principle components indicate high variance in orthogonal directions with the first component. Thus, the principal components encompass the majority of feature variance within a much smaller space of orthogonal dimensions onto which both training and test data can be mapped. In addition to feature extraction by PCA, we also tested a supervised feature selection method which selected features according to their individual correlation with the cognitive load label and inter-correlation with other attributes.

For classification, we investigated classifier ensembles throght the boosting technique, which can be regarded as another way of dealing with the imbalance between data representation in a high dimensional space and scarcity of training data. Ensemble classification [8] is a supervised learning scheme which combines the predictions of multiple classifiers. Boosting is a typical ensemble classifier. Boosting was introduced by [9] to produce one accurate prediction by combining moderately inaccurate predictions from a group of weak learners. The most popular Boosting algorithm is AdaBoost. Boosting effects guided changes of the training data to direct further classifiers toward more "difficult cases" [10]. AdaBoost implements this idea in two steps. The first step is to run a base learner repeatedly for a number of times and to maintain a distribution (a set of weights) over the training set. The second step is to update the weights in each round. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set [11]. The final prediction is the categorised weighted sum of the predictions from base learners.

Experiments showed that the cognitive load prediction task is better handled with supervised feature selection and different classification schemes. Contrary to our expectations, PCA feature extraction proved quite ineffective. However, with supervised feature selection a boosting global model achieved unweighted average recall (UAR) scores 20.5% and 18% higher than a published baseline based on a tuned support vector ma-

Technical Report - Dept of Computer Science, Trinity College Dublin. This study was funded by the Science Foundation Ireland through the CNGL grant number 07/CE/1142.

chine (SVM) classifier [7], in the Stroop time pressure and dual task, respectively. Similar per-task models were not quite as successful, but still yielded an improvement of 12% in the Stroop dual task data.

2. The Dataset

The "Cognitive Load with Speech and EGG (CLSE)" dataset [12, 7] was designed to support the investigation of acoustic features and evaluation of algorithms for the determination of a speaker's cognitive load and working memory during speech. THe CLSE database comprises recordings of 20 male and 6 female native Australian English speakers. These recordings encompass four types of experimental tasks, namely: *reading span Sentence*, *reading span Letter*, *Stroop time pressure* and *Stroop dual task*. These tasks define four partitions of the CLSE dataset. In each case, the data instances are classified objectively ("objective_load_level") into three distinct cognitive load levels: low (L1), medium (L2) and high (L3) levels.

The "span" tasks are used to measure the working memory capacity of a subject [12], in which participants are required to remember concepts or objects in the presence of distractors [13, 7]. The reading span task is based on the protocol described by Unsworth et Al. [12, 14]. It required the participants to read a series of (between two to five) possibly illogical short sentences, indicate whether the sentence read was true or false, and then remember a single letter presented briefly between sentences. This setup allowed the gatherer of the dataset to label memory load levels objectively as: L1, for data from the first sentence, L2, for data from the second sentence, and L3, for data from the third, fourth, and fifth sentences (for which no further distinctions were made). Data instances from letter reading (*readingspanLetter*), which contain single letter utterances, were considered insufficient for reliable categorisation and therefore did not form part of the ComParE 2014 challenge [7].

The Stroop tasks (*Stroop time pressure* and *Stroop dual task*), named after JR Stroop's seminal experiments [15], aim to induce increased cognitive load through presentation of conflicting stimuli to the participant. In this case, the stimuli are word and colour. The participant is asked to name the font colour of words corresponding to different colour names. Data instances produced in conditions where both the colour and the word that named the colour were the same were labelled as L1 (low cognitive load). Where the font colours and the colour names differed, data were labelled L2 or L3 (medium or high level of cognitive load). The high level was defined in terms of the time pressure on the subject (i.e. the colour had to be named in a short period of time, namely .8s) or in terms of task complexity (i.e. participants were required to perform a tone-counting task in addition to naming the font colour). These distinctions characterise the Stroop time pressure and Stroop dual task subsets of the CLSE dataset, respectively. These subsets each contain three utterances for each of three cognitive load levels per speaker.

For a more comprehensive description of the CLSE dataset and other details on the ComParE 2014 challenge the reader is referred to the paper Schuller et Al. [7].

3. Exploratory Analysis

In this section we take initial steps in analysing the CLSE dataset and illustrate the statistics of key factors. Table 1 shows that the validation and the test set contain roughly same number of instances, while the training set contains about 50% more data. Among the four types of tasks employed in data collection, the two *span* tasks occupy the majority of the dataset while the two *Stroop* tasks comprise only about 10% of each dataset. Considering that the dataset has 6,374 attributes in total, one can readily see that the *Stroop* sets are affected more severely by the curse of dimensionality.

Table 1: Summary of instance quantities in each type of task

	Training	Validation	Test
reading span letter	815	499	576
reading span sentence	825	525	600
stroop time pressure	99	63	72
stroop dual task	99	63	72
Total	1838	1150	1320

3.1. Data Cleansing

A fair portion of features in the training set have very low variance. This includes, for instance, all quadratic regression coefficients of level 1, and a number of other prosodic features. Some low level descriptors of spectral features also suffer from this problem. The root mean square signal frame energy feature (pcm_RMSenergy_sma_lpgain) is a case in point, with mean 1.98e-05 and variance 9.55e-10 in the training set. Such features are nearly constant and bring little discriminatory power to the classification model. We therefore removed all features with standard deviation less than 0.01. In Total 252 features (3.95% of all features) were removed from the training set, as a preprocessing step for all modelling experiments in this report. This yielded the extra benefit of reducing the training time.

4. Predicting Objective Load Labels

In this section we propose supervised learning models to predict the objective load level class. A training containing with 1,838 instances described by 6374 features challenges most classifiers since the data points are sparse with respect to dimensionality. The sparsity is more severe for models trained on subsets that contain only instances of a particular task (per task models). Two dimensionality reduction methods are tested in this section, along with an ensemble classifier AdaBoost.M1 and different base learners.

4.1. PCA experiments

PCA, as an unsupervised learning method, is assumed to reduce feature dimensions and keep the majority of data variation at the same time. We use scaled and centered all features so that they had unit variance before analysis. In the training set with four types of objective load tasks, the first 8 principal components (PC) explain over 95% of cumulative variance. We took 20 PCs and reencoded training and validation set into this new space. The cleaned features are projected onto the 20 PCs, and used for training (the transformed training set has 1838 instances with 20 features). When testing with the validation set, features need to be projected to the 20 PCs before the prediction step.

Here a global model is trained with 1838 instances altogether, and used to predict on each instance in the validation set. UAR scores were collected for each task. Contrary to our expectation, both a the naive Bayes classifier and the AdaBoost classifier failed to produce satisfactory results. We found that the UAR scores were far below baseline with the SVM global model of [7]. We speculate that the reason of this low performance on the PCA-reduced sets is the lack of an effective method for normalising the data per speaker on the training and test set. In the absence of such normalisation, PCA may be dominated by a few predominant features which can easily lead this method to overfit.

4.2. Feature Selection and Global Model

Faced with the failuse of an unsupervised method of dimensionality reduction, we attempted a spervised approach. The CfsSubsetEval feature filter provided by the Weka package [16] was employed. It selects attributes by individual correlation with the class variable and inter-correlation with other attributes. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred [17]. We compare global model prediction UAR with and without CfsSubsetEval pre-filtering in Table 2.

About classifier, we prefer Boosting with decision tree base learner instead of decision stump. The latter is a single node tree and classifies an instance by one feature. Although the feature is chosen by entropy, decision stump is too simple as a base learner in load level corpus. On the other hand, a decision tree with branching factor M=2 (minimum number of instances per leaf) by default naturally incorporates more attributes in base learner and helps the ensemble classifier.

Table 2: The effect of feature selection with AdaBoost classifier on validation set. UAR scores are from the global model, and AdaBoost is trained 30 iterations with decision tree base classifier. FS indicates feature selection with the CfsSubsetEval filter

	FS = No	FS = Yes	baseline
reading span sentence	48.50%	55.39%	61.3%
stroop time pressure	57.14%	65.08%	54.0%
stroop dual task	49.21%	52.38%	44.4%

Table 2 shows the efficacy of feature selection combined with an AdaBoost.M1 with Decision Tree base learner. Without feature selection, AdaBoostM1 beats SVM baseline slightly with Stroop tasks, but is 13% lower than baseline with reading task. This observation shows the power of ensemble classification in this dataset when there is a proper base learner. When feature selection is in use, the global model achieves higher accuracy for each task. In Stroop time pressure task, the best UAR is 65.08%, an improvement of 11 points over the baseline. In the Stroop dual task, the best UAR is 52.38%, an 8-point improvement over the baseline. However, reading span sentence task is still 6% lower than baseline. In the next section we investigate per task models, where classifiers are trained on relatively more uniform training sets.

4.3. Per Task Model

In the preceding section, we predicted objective load level with a global model which trains a single model on all available instances and predicts on a validation set of each task. In this section we apply an alternative approach, training one model with data from one task and predicting on a validation set of the corresponding task. This is called a per task model [7]. A comprehensive training set contains objective load level instances from four tasks, part of which could be redundant for predicting on one task. Since the SVM baseline shows significantly better UAR scores with Stroop tasks, Per Task models are expected to outperform the global model in our experiments.

The split training sets are filtered in the same way as previous experiments. Features with standard deviation less than 0.01 are dropped off. The CfsSubsetEval filter selects 93, 74 and 51 features by sequence for each task, although there are 6374 features in total. Then AdaBoost.M1 is employed as a classifier for the corresponding per task models. The number of training iterations is set to 20 for each base learner. Since the Decision Tree (DT) base learner works well for the Global model, it is used again. Moreover, we also use a Decision Stump (DS) base learner for comparison.

We expect AdaBoost.M1 with a DT base learner to predict better than with DS, since the former has a more complex tree structure and could learn more subtle rules. However, but the outcome is the opposite of what we expected. DS beats DT in both stroop tasks (Table 3). We discuss this result observation in the following section.

Table 3: The effect of feature selection with AdaBoost classifier on validation set. UAR scores are from Per Task model, and AdaBoost is trained 20 iterations with each base learner

	Ada+DT	Ada+DS	baseline
reading span sentence	54.98%	48.86%	61.2%
stroop time pressure	68.25%	73.02%	74.6%
stroop dual task	66.67%	71.43%	63.5%

4.4. Over-fitting and Under-fitting

Decision Stump, as the simplest tree structure, outperforms Decision Tree (branching factor 2) in AdaBoost for both Stroop tasks. This observation comes from per task model prediction on the validation set and seems quite surprising. In order to test its validity, we further analyse the Stroop Dual Task model prediction within the training set. Figure 1 shows the performance of both DS and DT base learners under different numbers of AdaBoost iterations. It is clear that AdaBoost with the DT base learner reaches 100% UAR in the training set regardless of the number of training steps (10 to 100 iterations). At the same time, its prediction accuracy on the validation set oscillates between 61.90% and 68.25%. When we run more iterations for DT, there is no clear trend of increase or decrease in UAR on the validation set. This suggests over-fitting. In this situation, accuracy on the validation set depends on randomness of the decision boundary in the hypothesis space, and the boundary margin is already too narrow.

On the other hand, DS as a simpler model improves with more training steps. Its UAR score improves in both training set and validation set when iteration increases from 10 to 20. The accuracy on the training set is far below 100%, but cannot be improved when iteration is over 20. DS reaches its upper bound of prediction power. We have seen that DS and DT both exhibit their best results on the Stroop Dual Task model, and there is no need to explore a more complex model structure. The fact that DS outperforms DT as an AdaBoost base learner is therefore to be expected. The sub-tasks with the smallest numbers of instances (Stroop dual, and Stroop time pressure) tend to favour simpler models that are less prone to overfitting.

However, DT outperforms DS as a base learner for AdaBoost.M1 in the Reading Span Sentence task (Table 3). DS training UAR keeps stably below 50% when training iterations

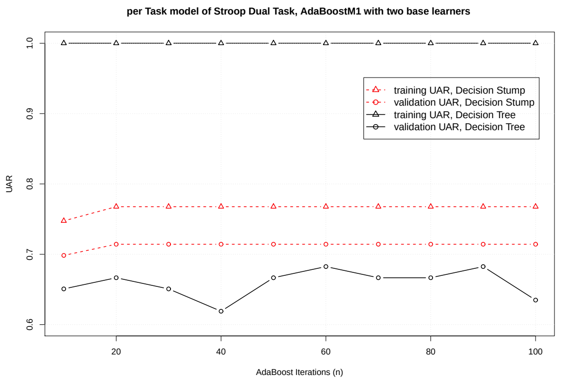

Figure 1: Per task model of Stroop Dual Task, AdaBoost.M1 with Decision Stump and Decision Tree base learners

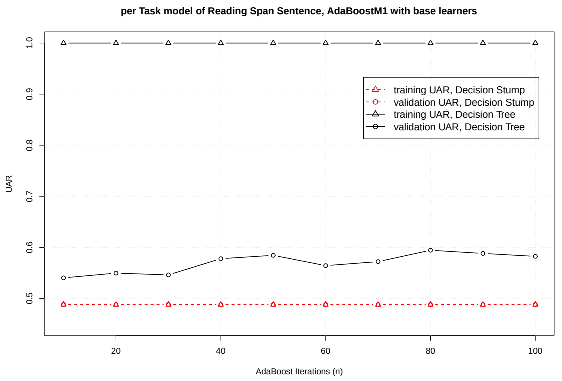

Figure 2: Per task model of Reading Span Sentence, AdaBoost.M1 with Decision Stump and Decision Tree base learners

increases from 10 to 100 (Figure 2). This is a sign of underfitting. DS as a base learner cannot represent the variances in a Reading task with 825 instances (Table 1). As in the previous per task Stroop models, the DT based classifier's training UAR is 100% when iteration equals 10, indicating that it does not suffer from the same problem. Unlike the previous case, however, in the reading task model, the UAR of DT on the validation set has a roughly increasing trend with more iterations. Prediction power is increasing with a more complex model, so here there is no indication of over-fitting. More iterations or more complex DT base learners could induce better UAR on the validation set.

5. Discussion

In this paper we proposed solutions for classifying three levels of objective load, with evidence of 6374 acoustic features and EEG features. In contrast to the rich feature set, there are only 1838 instances spanning four different tasks. Since a moderately tuned SVM classifier only achieves a 44.4% baseline on a Stroop task, our results serve to emphasise the importance of data cleansing and dimensionality reduction in this modelling challenge.

In data cleansing, we dropped 252 features with standard deviation less than 0.01. These features are nearly constant, offering little value for discriminating among the three class levels while adding to the computational load. Experiments show that boosting models work well without these features, and the training time is reduced significantly. However, the number of features remaining after this pre-processing step is still very large, and dimension reduction is needed.

We found that dimensionality reduction by feature extraction through principle component analysis harms boostingm as well as other models such as naive Bayes. The reason of PCA low performance may be attributed to the huge differences of mean values among the features and the lack of an effective unsupervised way of normalising these values on a per speaker basis. On the other hand, the (supervised) CfsSubsetEval filter proved to be an effective feature selection method. The features with high correlation with class variable and low intercorrelation with other features were favoured. Multicollinearity is thus alleviated in this large feature set and finally 52 out of 6374 features are selected for a boosting global model. The reduced feature set does improve accuracy and improves on the SVM baseline for the Stroop data (Table 2).

The outcome of feature selection is encouraging, but we still need to improve model accuracy by controlling the complexity of a supervised learning model. The boosting model combines the predictions from multiple classifiers and is generally more accurate than a single classifier. The training iterations act as a controller of model complexity. In the first round, a base classifier is built. In the next round, the weight of the $n+1$ base learner is D_{n+1}, which is higher on instances that learner n has error on. The final decision is a collective vote by weighted N base learners. When boosting has no error on the training set, the generalisation power of base learner is enough for the current input. When validation accuracy keeps increasing with training accuracy stable at 100%, it is necessary to try to model with more iterations, thereby increasing the risk of over-fitting. However, when training accuracy keeps stable at low values as the number of iterations increases, there is little point in preceding. Such base learner is not complex enough to represent feature variances adequately.

Through the trials reported in Section 4, we see that certain objective load tasks favour certain classifier settings, especially with per task models. In order to show this clearly, we summarise the best UAR scores in Table 4.

Table 4: Summary of the best UAR scores for the validation set

UAR[%]	Per Task model		Global model	
Task	Valida.	baseline	Valida.	baseline
reading span sentence	(1) 59.45	61.2	(4) 55.39	61.3
stroop time pressure	(2) 73.02	74.6	(4) 65.08	54.0
stroop dual task	(3) 71.43	63.5	(4) 52.38	44.4

(1) AdaBoost.M1 with DT base learner (M=2, iterations I=80)
(2) AdaBoost.M1 with Decision Stump base learner, I=10
(3) AdaBoost.M1 with Decision Stump base learner, I=20
(4) AdaBoost.M1 with Decision Tree base learner (M=2, I=30)

6. Conclusion

We presented an initial exploration of feature selection and modelling trade-offs to be taken into account when approaching the challenge of categorising a speaker's cognitive load state based on acoustic features. This task to is relevant to practical applications, such as "meeting browsing" [18], and successful prediction can help add structure to multiparty communication [19], possibly in conjunction with other non-verbal features [20].

This is, however, a complex challenge and as the results reported here demonstrate, there is ample room for further exploration. In the near future we plan to investigate unsupervised ways of normalising the features per speaker as well as explore models that can take advantage of global data in per task modelling.

7. References

[1] N. Campbell, "On the use of nonverbal speech sounds in human communication," in *Verbal and nonverbal communication behaviours*, ser. Lecture Notes in Computer Science, A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro, Eds. Springer, 2007, pp. 117–128.

[2] S. Luz, "Locating case discussion segments in recorded medical team meetings," in *Proceedings of the ACM Multimedia Workshop on Searching Spontaneous Conversational Speech (SSCS'09)*. Beijing, China: ACM Press, Oct. 2009, pp. 21–30.

[3] S. Luz and J. Su, "The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings," in *Proceedings of INTERSPEECH 2010*. Chiba, Japan: ISCA, 2010, pp. 1369–1372.

[4] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings: The AMI and AMIDA projects," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*. Kyoto, Japan: IEEE, Dec. 2007, pp. 238–247.

[5] D. M. Roy and S. Luz, "Audio meeting history tool: Interactive graphical user-support for virtual audio meetings," in *Proceedings of the ESCA workshop: Accessing information in spoken audio*. Cambridge University, Apr. 1999, pp. 107–110.

[6] A. Gevins and M. E. Smith, "Neurophysiological measures of cognitive workload during human-computer interaction," *Theoretical Issues in Ergonomics Science*, vol. 4, no. 1-2, pp. 113–131, Jan. 2003.

[7] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH'14)*. ISCA, 2014.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting,," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.

[10] L. I. Kuncheva, M. Skurichina, and R. Duin, "An experimental study on diversity for bagging and boosting with linear classifiers," *Information Fusion*, vol. 3, pp. 245–258, 2002.

[11] Y. Freund and R. Schapire, "A short introduction to boosting," *J. Japan. Soc. for Artif. Intel.*, vol. 14, no. 5, pp. 771–780, 1999.

[12] T. F. Yap, "Speech production under cognitive load: Effects and classification," Ph.D. dissertation, The University of New South Wales, 2012.

[13] A. R. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle, "Working memory span tasks: A methodological review and users guide," *Psychonomic bulletin & review*, vol. 12, no. 5, pp. 769–786, 2005. [Online]. Available: http://link.springer.com/article/10.3758/BF03196772

[14] N. Unsworth, R. P. Heitz, J. C. Schrock, and R. W. Engle, "An automated version of the operation span task," *Behavior research methods*, vol. 37, no. 3, pp. 498–505, 2005. [Online]. Available: http://link.springer.com/article/10.3758/BF03192720

[15] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of experimental psychology*, vol. 18, no. 6, pp. 643–662, 1935.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[17] M. A. Hall, *Correlation-based Feature Subset Selection for Machine Learning*. The University of Waikato, Hamilton, New Zealand, 1999.

[18] S. Luz and D. M. Roy, "Meeting browser: A system for visualising and accessing audio in multicast meetings," in *Proceedings of the International Workshop on Multimedia Signal Processing*. IEEE Signal Processing Society, Sep. 1999, pp. 489–494.

[19] T. P. Moran, L. Palen, S. Harrison, P. Chiu, D. Kimber, S. Minneman, W. van Melle, and P. Zellweger, ""I'll get that off the audio": A case study of salvaging multimedia meeting records," in *Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems*, vol. 1, 1997, pp. 202–209.

[20] S. Luz, "The non-verbal structure of patient case discussions in multidisciplinary medical team meetings," *ACM Transactions on Information Systems*, vol. 30, no. 3, p. article 17, 2012.