Predicting Negotiation Outcomes from Behavioral Data

Prerna Chikersal
pchikers

Jianlin Zhan
jianlinz

Chaitanya Modak
cmodak

Abstract

Our nonverbal behaviors often affect or sometimes even dictate the outcome of everyday conversations. This project attempts to quantify the effect of some indicators of nonverbsl behavior that are gleaned from audio or video data on a negotiation task. We extract relevant features from the data streams and use them as inputs to a variety of classifiers that classify the negotiation score into three categories, based on standard deviation. The goal of our project is to compare various machine learning algorithms to figure out what works best for multimodal data.

1 Introduction

Negotiation is a complex interaction in which two or more parties confer with one another to arrive at the settlement of some matter like resolving a conflict, or to share common resources. The parties involved in the negotiation often have non-identical preferences and goals that they try to reach. Sometimes the parties simply try to change a situation to their favor by haggling over price. In other cases, there can be a more complex trade-off between issues. Being a good negotiator is not a skill that all humans naturally have; therefore, this line of research can potentially be used to help humans become better negotiators.

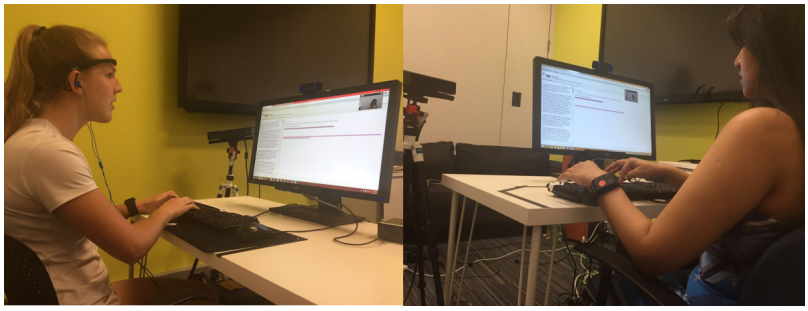

Figure 1: Experiment Setup

This project is based on an ongoing project at the Connected Experiences Lab in HCII. Our goal is to study the physiological and behavioral basis of collective intelligence and negotiation. The study involves two participants working with each other on Skype, on a series of Collective Intelligence (CI) tasks. The tasks have been proposed by Prof. Anita Woolley at Heinz college and we also include a commonly used Negotiation task in our analysis. Our dataset consists of multiple streams such as EEG (4 channels), eye tracking (gaze, pupillary dilation, fixation), skeleton tracking via Kinect, Video, Audio, Electrodermal activity, and Heart Rate variability. Most of the CI tasks and the Negotiation task are automatically scored according to different fixed and recommended scales. The CI tasks that cannot be automatically scored such as the brainstorming task, are being scored

by psychologists working with the lab. For the purpose of the 10-701 project, weve reduced the scope of the project to using features extracted from Video such as Facial Action Units (FACS), gaze direction and head pose, and Audio characteristics like tone, pitch, prosody, as well as characteristics indicative of collaboration like turn taking behaviors. The data is collected from both participants on an individual and collective level to extract features to predict negotiation outcomes in the score on a scale between -8400 to +13200.

2 Related Work

We explored various behavioral representations for synchrony such as DTW as implemented by Hernandez et al. [11]. We calculate DTW, pearsons correlation, and differences in mean and standard deviation over the occurrences in time of happy and sad expressions throughout the video. Our dataset is multimodal that is our features come from multiple sources (i.e. audio and video), and its not necessarily that the same type of kernel will fit all types of features the best. Multiple Kernel Learning is a method commonly used for classification of multimodal data [cite papers here]. It allows us to specify multiple kernels and learn different weights for them for different types of data. Nouri et al., (2014) [10] try to predict the goal and outcome in the using multi-modal features such as sentiments of the turns. Nouri et al. (2015) [9] use models based on initiative-features. They used the mean and standard deviation of acoustic features such as the amount of silence and speaking time. They divide each exchange into four categories: R' (directly relates to prior utterance), 'F' (fulfills a pending discourse obligation), 'L ' (imposes a discourse obligation,) and 'N' (provides new material that is optional). Park et al. [7] consider negotiation as an ongoing process in which participants constantly adapt themselves to each other. Assessing both short-term and long-term behaviors provides a deeper understanding of the current state of negotiation on which to base predictions of future actions. Short-term behavior consists of mutual behavior descriptors are designed to model recent momentum of the negotiation. The short-term mutual behaviors are designed to quickly adapt while long-term behaviours vary more slowly. The approach they follow, and one that we have repliocated to some degree is to divide the features into visual and acoustic streams. Visual features can include smiles, leaning postures, head gaze, eye gaze etc. and acoustic features are the voice quality to indicate breathiness or tenseness of the voice (Values closer to zero are considered as more tense), the base frequency of the speech signal, loudness and intensity of the voice, Energy slope and Spectral Stationarity. Other linguistic features include [8] the number of words spoken by each speaker in each turn of the dialogue, the number of turns taken during the negotiation, the number of times words corresponding to the negotiation items are spoken, sentiment and subjectivity scores calculated for words and turns and the whole dialogue.

3 Dataset

The data essentially consists of two speakers talking to each other over Skype for a period of 22 minutes. Thus, we have audio and visual streams corresponding to each pair of participants for around 22 minutes. The task to be performed by the two persons simulates a negotiation between a job recruiter and a job candidate. The issues to be debated are: Bonus, Job Assignment, Vacation Time, Starting Date, Moving Expense Coverage, Insurance Coverage, Salary and Location. Each criterion has 5 sub-categories , each of which corresponds to specified scores for the candidate and interviewer respectively. The goal of both participants (candidate and employer) is to maximize their score on the aforementioned parameters. For instance, the candidate will try to get as high a bonus as possible with more vacation time at a the most attractive location. The interviewer on the other hand will try to hire the candidate at the lowest possible cost to the company and will try to get the candidate sign up for a less favoured location. Both participants are only aware of how their own points are calculated and do not know the scoring system for the other person. This prevents them from colluding to artificially generate a high score. At the same time, they are aware of the trade-offs they must make in order to maximize their own score.

For the purpose of this report, we are using data from 42 dyads. Thus, we have 42 scores corresponding to candidate and interviewer each. As mentioned earlier, the negotiation scores can lie between -8400 to 13200. We find that in our data the highest candidate score is 10800 while the lowest score by a candidate is -600. The interviewer scores vary between -2400 and 10100. The data is collected via multiple streams. These are:

- Eye Tracking
- Electrodermal Data
- Skype Video
- Skype Audio

4 Samples, Labels and Feature Extraction

4.1 Samples

As we mentioned in the previous section, we have data from 42 Skype conversations. This corresponds to 84 samples, one sample being one person.

4.2 Labels

The negotiation scores are thresholded and split into three labels: Good (+1), Fair (0) or Poor (-1).

4.3 Pre-processing

CLM-tracker [3] labels the occurrence of the following facial action coding unit (henceforth referred to as AU) [4] for each frame: regression score for AU1, AU2, AU4, AU5, AU6, AU9, AU10, AU12, AU14, AU15, Au17, AU20, AU26 and AU26, and classification output (0 or 1) for AU4, AU12, AU15, AU23, AU28 and AU45. Whenever, the regression score is greater than 0.5, we assume that the AU in question occurs in that frame. Whenever classification output is 1, we assume that the AU in question occurs in that frame.

4.4 Individualistic Features

Individualistic features are features extracted for each individual and are completely independent of any information about the individuals negotiation opponent. As discussed above, we already have a prediction of which AUs occur in each frame. Now, we use existing literature on AUs, which correlates different combinations of AUs to emotions. Whenever AUs 6 and 12 occur together in a frame, we can say that the person expresses happiness in that frame. Whenever AUs 1, 4, and 15 occur in a frame, we can say that the person expresses sadness in that frame. Whenever AUs 1, 2, 5, and 26 occur in a frame, we can say that the person expresses surprise in that frame. Whenever AUs 1, 2, 4, 5, 20, and 26 occur in a frame, we can say that the person expresses fear in that frame.

From this we define the following features:

- Mean of each of the 20 AUs (14 regression, and 6 classification) output in frames: This is given by (Number of frames in which AU_i occurs) divided by (Number of frames in which all $AU_i s$ occur)
- Mean of the occurrence of i^{th} expression in frames(where i can be happiness, sadness, surprise, or fear): This is given by (Number of frames in which person expresses i) divided by the sum of the (Number of frames in which happiness occurs, Number of frames in which sadness occurs, Number of frames in which surprise occurs, Number of frames in which fear occurs)
- Count of each of the 20 AUs: This does not count the number of frames in which each AU occurs. Rather, it counts the occurrence of AU(i) in N consecutive frames, as one occurrence of AU(i). Thus it effectively counts then number of onsets of each AU.
- Count of each expression: This does not count the number of frames in which each expression occurs. Rather, it counts the occurrence of expression(i) where i = happiness, sadness, surprise, and fear, in N consecutive frames, as one occurrence of expression(i)
- Max length of each of the 20 AUs: For each AU(i) it returns the longest number of consecutive frames in which AU(i) occurs.

- Max length of each of the 4 expressions (happiness, sadness, surprise, fear): For each $expression_i$, it returns the longest number of consecutive frames in which $expression_i$ occurs
- Average length of each of the 20 AUs: For each AU_i it returns the average number of consecutive frames in which AU_ioccurs.
- Average length of each of the 4 expressions (happiness, sadness, surprise, fear): For each $expression_i$, it returns the average number of consecutive frames in which $expression_i$ occurs

So, from the above, we get 24 means, 24 counts, 24 maxs, and 24 averages. Thus, giving us 96 individualistic features.

4.5 Mimicry based Features

Behavioral mimicry or the chameleon effect [5] occurs when an individual unconsciously mimics the actions of another individual with whom that individual is engaged in a conversation.

To extract mimicry-based features for an individual A, negotiating with another individual B. We consider individual A to be the leader during mimicry whenever individual B mimics an expression of individual A, after within X seconds of individual As enactment of the expression. This reenactment by individual B of the expression$i$ made by individual A, is counted as 1 occurrence of mimicry for expression i. Experimentally, we have chosen X to be 2 seconds, at this stage of our project. We will explore a number of options for X in future and find what may be best. We compute number of occurrences of mimicry for each of the 20 categories of AUs as well as each of the 4 categories expressions. If individual B mimics individual A at least once within X seconds, we increment the $countMimicry$ feature by 1. If individual B mimics individual A k times within X seconds, we increment the $countMultiMimicry$ feature by k.

Thus we have (20+4)*2 = 48 mimicry-based features. Hence, our feature vector consists of 144 features.

4.6 Sychrony based Features

Invented in 1970s, Dynamic Time Warping (DTW) is an algorithm which, given two time series, will stretch or compress them locally in order to make one resemble the other as much as possible. DTW outputs a score of distance, and is not bound to length of the signal nor the shape of the signal. Particularly, we use DTW here on our action units out of Facial Action Coding System. In our input generated by CLM-framework, scores of action units- either regression or classification predictions, ordered by frame are shown below: DTW gives the distance between these two time-seriess as a

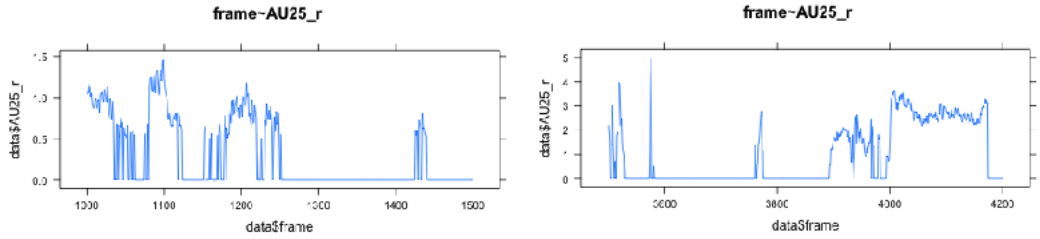

Figure 2: L: Candidates, R: Interviewer

partial presentation of difference between candidates and interviewers in experiment.

We also measure the mutuality and difference within a pair of candidate and interviewer by mutuality (Pearsons correlation across features) and difference (differences in mean and standard deviation values). Furthermore, we classify standardized scores into three categories: poor, fair, and good as

shown in the table below. After applying scoring to quantify the negotiation outcomes, we normalize the scores among the candidate group and the interviewer group by standardization.

Negotiation Score	Poor	Fair	Good
Probability	33.33%	33.33%	33.33%
Z values	< -0.431	[-0.431, -0.431]	> 0.431

Figure 3: Classification of Negotiation Outcomes

These features have not be used in our current analysis because of computational purpose. We are working on a faster implementation in order to incorporate them for the final submission.

4.7 Audio Features

We first sample the data at 16 kHz and pass it through a low pass filter to remove noise. Then, we perform a frequency analysis of the signal and find the statistics of the frequency distribution (mean, median, maximum, minimum etc.). These are indicative of emotions such as excitement, happiness, fear and so on. We used the OpenSMILE toolkit for extracting the audio features. The toolkit has been started at Technische Universitt Mnchen (TUM) and enable extraction of large audio feature spaces in realtime. SMILE is an acronym for Speech and Music Interpretation by Large-space Extraction. The main features of openSMILE are its capability of on-line incremental processing and its modularity. Using openSMILE, we extracted the following acoustic features.

Mel Frequency Cepstral Coefficients (MFCCs): These coefficients take into account the way the human vocal tract produces speech. The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear. We consider 13 MFCCs, 13 Delta coefficients and 13 double delta coefficients.

Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP): PLP involves warping spectra to minimize the differences between speakers while preserving the important speech information. RASTA is a separate technique that applies a band-pass filter to the energy in each frequency subband in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel. We consider 6 PLPs, 6 delta coefficient and 6 double delta coefficients.

Prosody: The prosodic features we extracted are: 1) The Fundamental Frequency 2) The voicing probability and 3) The loudness/ intensity.

The intuition behind these features is that they give information about vocal parameters like pitch , intonation and intensity which are highly indicative of human emotion. Human emotion in turn, as suggested earlier, gives us an idea of the state of the negotiation process at a given time.

5 Classifiers

5.1 Support Vector Machines

A Support Vector Machine (SVM) is a binary linear classifier. If the data is linearly separable, it can be separated using a hyperplane. For instance, if the linearly data can be represented in 2-D, it can be separated by a single line (see figure below). For data that is not inherently linear separable, SVMs can map the data into a higher dimension in which it is linearly separable. We use a kernel function for this purpose.

SVMs can be used for both classification and regression. We tackle the classification problem. There are many hyperplanes that might classify the data. Support Vector Machines pick the hyperplane which represents the largest separation, or margin, between the two classes. So, it chooses the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines

is known as a maximum margin classifier. If the training data are linearly separable, we can select two parallel hyperplanes that separate the two classes of data, so that the distance between them is as large as possible. The region bounded by these two hyperplanes is called the "margin", and the maximum-margin hyperplane is the hyperplane that lies halfway between them. These hyperplanes can be described by the equations

$$\vec{w} * \vec{x} - b = 1 \tag{1}$$

$$\vec{w} * \vec{x} - b = -1 \tag{2}$$

The points with label $y = 1$ are constrained such that $\vec{w} * \vec{x} - b > 1$, and those with label $y = -1$ are constrained such that $\vec{w} * \vec{x} - b < 1$. By a neat mathematical trick, we can write the constraint equation of the SVM as $y(\vec{w} * \vec{x} - b) > 1$. We minimize the 2-norm of $\vec{w}$ in order to maximize the separation of planes, subject to the constraint derived above. We used the Scikit-learn library to implement the SVM. [1]

5.2 Multiple Kernel Learning

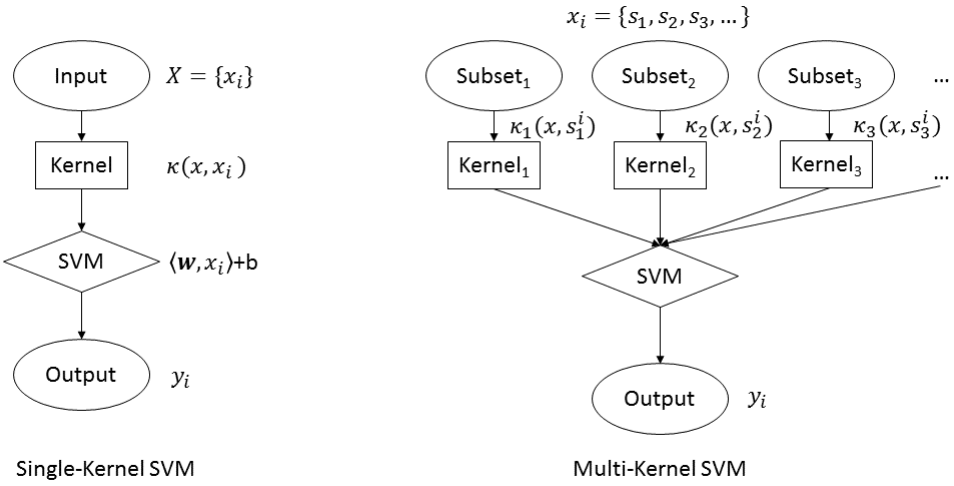

Figure 4: MKL Flow

Instead of using a single kernel function, MKL uses a combination of kernel functions which can be either a linear or non-linear combination. A formal definition (linear MKL) is as follows:

$$k_{MKL}(x_j, x_k) = \sum_i \mu_i k_i(x_j, x_k)$$

$$\mu_i > 0, \sum \mu_i^2 \leq 1$$

Hence, the new objective function is:

$$max \sum_l \alpha_l - \frac{1}{2} \sum_{j,k} \alpha_j \alpha_k y_j y_k \sum_i \mu_i k_i(x_j, x_k)$$

As observed, the function is identical to that of an SVM expcept for that fact that the kernel function is now replaced by a sum over kernels. Training SVM is identical to traditional SVM solvers, only with the replaced kernel. However, it brings up an additional training task that we need to learn- to find the values of the $\mu_i' s$. To solve, we can easily implement a gradient descent approach.

$$J(\mu) = \sum_l \alpha_l - \frac{1}{2} \sum_{j,k} \alpha_j \alpha_k y_j y_k \sum_i \mu_i k_i(x_j, x_k)$$

By the theorem of Bonnans and Shapiro:

$$\frac{\partial J(\mu)}{\partial \mu_i} = -\frac{1}{2}\sum_{j,k} \alpha_j \alpha_k y_j y_k k_i(x_j, x_k)$$

$\mu_k^1 = \frac{1}{M}$ for $k = 1, ..., M$
for $t = 1, 2, ...$ **do**
solve the classical SVM problem with $K = \sum_k^M \mu_k^t K_k$
compute $\frac{\partial J}{\partial \mu_k}$ for $k = 1, ..., M$
compute descent direction D_t and optimal step γ_t
$\mu_k^{t+1} \leftarrow \mu_k^t + \gamma_t D_{t,k}$
if stopping criterion **then**
break
end if
end for

Figure 5: Pseudo code for MKL Learning

The data is tested as follows:

$$f(x_{new}) = -\sum_l \alpha_l y_l k_{MKL}(x_l, x_{new}) + b$$

6 Experiments and Results

As mentioned earlier, our dataset contains 42 dyads (84 persons) of 19 low performers, 50 moderate performers and 15 high performers. 12 dyads each have either both members as males or both as females and 18 have a mixed gender interaction. We created a training data set such that the number of male and female participants was equal, so as to eliminate any possible bias. This is especially important with audio features which are significantly different according to gender. Since we have limited data samples, we use Leave one out Cross-Validation in order to train the SVM parameters both for the linear kernel as well as Multiple Kernel approach. Effectively, we are using (35 traning samples + 1 validation sample) for training and 6 samples for testing. The training and test sets are also iterated five times to yield five different accuracy measure. An average is taken as the final classification accuracy.

6.1 Baseline Model using Linear SVMs

Video Only	Best C	Audio only	Best C	Audio + Video	Best C
50	1.5	41	0.1	41.67	0.1
41.67	0.1	58.33	0.8	58.33	20
50	0.2	66.67	20	50	0.5
50	0.1	33.33	1.5	41.67	2
50	15	41.67	0.1	75	0.1
48.334		48.2		53.334	

Figure 6: Baseline Model Using Linear SVMs

For the baseline method, we implemented Linear SVM and trained 3 classifiers on (i) video features only, (ii) audio features only, (iii) both audio and video features. Since, we only have 6 test samples,

to ensure that our results our statistically conclusive, we trained and ran the SVM on 5 test sets, and reported the average accuracies. For each of these test sets, we carried out leave-one-out-cross-validation to get the C parameter. During cross-validation, we also found that L1 regularization typically performs better than L2 regularization in our case.

6.2 Final Model using Multiple Kernels and incorporating Audio Features

Weights	Accuracy on Audio + Video
[0.33720114879927737, 0.49837201148799276]	0.6666666667
[0.34381024343218369, 0.49843810243432191]	0.5833333333
[0.34348235887141965, 0.49843482358871422]	0.5
[0.34348235887141965, 0.49843482358871422]	0.5833333333
[0.32649735157859305, 0.49826497351578591]	1
	0.630952381

Figure 7: Final Model Using Multiple Kernel Learning

Here the two values for the weights μ indicate the weight the audio and video models. Clearly MKL performs better than a linear SVM by 10 percent. Higher weights are usually assigned to the audio model. This indicates that the audio features are more discriminating but we will need to investigate a little further before drawing any certain conclusions.

7 Analysis and Future Work

Athough we get acceptable results, our training accuracy is at times as high as 100 percent on certain runs. Hence, its likely that our model overfits the training data. This could be because our number of features are far more than the number of training samples. In future, it would be interesting to see if our results improve when we use feature selection methods like Kernel PCA. We would also like to explore better behavioral representations for synchrony-based features such as Cross-recurrence quantification analysis []. This representation has been found to be particularly useful for continuous signals like electrodermal activity, heart rate, and EMG [cite2]. We also intend to explore different classification approaches based on Kernel Canonical Correlation Analysis (KCCA) and Multi-view Hidden Conditional Random Fields (MV-HCRF) proposed by Song et al [6]. We can also train the model using Deep Neural Networks given the richness and diversity of our features.

Acknowledgments

We would like to thank Prof. Laura Dabbish and Dr. Maria Tomprou for providing us access to this dataset, and giving us the opportunity to work on this project. We would like to thank Professor Tom Mitchell for his instruction and support, and for providing us with the flexibility to choose this topic, and Prof. LP Morency for suggesting the use of Multiple Kernel Learning for our project. We are also grateful to Brynn Edmunds, Abhinav Arora and the team of TAs for helping us out and for facilitating the project.

Division of Work

Work other than that specifically listed below has been completed together.
Prerna: Video Pre-processing using CLM-tracker, Individualistic features and mimicry-based features from video, Linear SVM, MKL idea and modeling
Jianlin: DTW for synchrony-based features from video, leave one out cross validation, MKL implementation
Chaitanya: Audio Pre-processing, and Audio feature extraction, Linear SVM and MKLimplementation

References

1. Fabian Pedregosa, Gal Varoquaux et al. (2011) Scikit-learn: Machine Learning in Python.
2. Anita Williams Woolley, Christopher F. Chabris et al. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. VOL 330 SCIENCE
3. Tadas Baltrusaitis, Peter Robinsonand Louis Phillipe Morency.(CVPR 2012). 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking
4. Jeffrey F. Cohn, Zara Ambadar and Paul Ekman. Observer-Based Measurement of Facial Expression With the Facial Action Coding System
5. Tanya Chartrand and John Bargh. The Chameleon Effect: The Perception-Behaviour Link and Social Interaction
6. Song, Yale, Louis-Philippe Morency, and Randall Davis. (ACM, 2012.) Multimodal human behavior analysis: learning correlation and interaction across modalities.
7. Sunghyun Park, Stefan Scherer, Jonathan Gratch, Peter Carnevale, and Louis-Philippe Morency. (2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.) Mutual Behaviors during Dyadic Negotiation: Automatic Prediction of Respondent Reactions
8. Thomas Polzin and Alex Waibel. Detecting Emotions from Speech.
9. Nouri, Park et al. Prediction of Strategy and Outcome as Negotiation Unfolds by Using Basic Verbal and Behavioural Features
10. Nouri, Traum. Initiative Taking in Negotiations (SIGDIAL 2014 Conference)
11. Hernandez, J., Riobo, I., Rozga, A., Abowd, G. D., and Picard, R. W. (2014, September). Using electrodermal activity to recognize ease of engagement in children during social interactions. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 307-317). ACM.
12. Dyrba, M., Grothe, M., Kirste, T., and Teipel, S. J. (2015). Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. Human brain mapping, 36(6), 2118-2131.
13. Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., ... and Ferrari, R. C. (2015). Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces, 1-13.
14. Fusaroli, R., Konvalinka, I., and Wallot, S. (2014). Analyzing social interactions: the promises and challenges of using cross recurrence quantification analysis. In Translational Recurrences (pp. 137-155). Springer International Publishing.
15. Mnster, D., Hkonsson, D. D., Eskildsen, J. K., and Wallot, S. (2016). Physiological evidence of interpersonal dynamics in a cooperative production task. Physiology and Behavior.