Automatic Prediction of Speed-Date Outcomes via Paralinguistic Features

Blake Wulfe
blake.w.wulfe@vanderbilt.edu
April 22nd 2014

1 Abstract

In this paper we consider the automatic prediction of human interaction outcomes through the use of paralinguistic features extracted from a noisy environment. This capability carries two primary benefits. First, it furthers technology which may improve human-human and human-computer interaction. Second, predicting interaction outcomes provides insight into human relationships. This research specifically focuses on predicting whether or not participants in a speed-dating situation will mutually elect to engage in future interaction. Using audiovisual data from 283, 3-minute speed-dates from the Berlin Speed Dating Study (BSDS), we extracted a set of 12,746 paralinguistic features, which were subsequently reduced to 16 and 120 features through ad hoc and algorithmic methods respectively. Multilayer perceptron (MLP) and support vector machine (SVM) models were trained on each feature set and used to predict date outcomes. Results from classification were poor, with neither classifier performing above the baseline. The MLP suffered from issues resulting from unaccounted for class skew and learned to classify each sample negatively. The resulting misclassification error of approximately 11% matches the class imbalance closely. The SVM, using class weightings, was unable to accurately make predictions based upon extracted paralinguistic features. A Linear Kernel SVM achieved 30% misclassification error with an outlier class F1 Score of .24. These results indicate that the extracted audio features contain little useful information, likely due to the highly human-noisy data collection environment and relatively unsophisticated methods of speaker diarization preprocessing. Future research on the BSDS dataset will focus on improving audio feature extraction, gathering a wide breadth of feature types, and using more advanced methods of feature development.

2 Introduction

Social signal processing (SSP) aims to provide computers with the ability to sense and understand human social signals [27]. This capability would allow computers to automatically adapt to user behavior as well as augment human-human interaction by informing individuals with extracted information. SSP consists of four distinct **tasks**:

1. Collecting data
2. Detecting people in data
3. Extracting behavioral cues
4. Interpreting cues as social signals within specific contexts

This process focuses on four primary **domains** from which social signals arise:

1. Face and eyes
2. Vocal behavior
3. Gestures
4. Interaction geometry and synchrony

Applications of SSP are generally organized into the following four areas:

1. Analyzing interacting humans
2. Coaching
3. Social robotics
4. Interaction with virtual agents

In this paper we focus on the first application area - analyzing interacting humans - and specifically on the subproblem of automatically predicting interaction outcomes. This problem has been considered in a number of contexts, though of particular relevancy are interview outcome prediction, negotiation outcome prediction, and attraction prediction. These subject areas

are convenient to study because they inherently provided labeled data (e.g., speed-date outcomes or job interviewer impressions). In [24] the researchers used vocal tone and prosody to distinguish between highly-rated and poorly-rated candidates with 88% classification accuracy. Four primary nonverbal features were considered: (1) activity (2) engagement (3) emphasis and (4) mirroring. Each measure was derived from low-level audio descriptors and functionals. The researchers used a bayesian network classifier to conclude that activity and emphasis were good predictors of interview outcome. The research was based on the relatively small sample size (n=26), but nevertheless illustrates the efficacy of predicting outcomes based upon automatic social signal analysis.

Prediction of negotiation outcomes is considered in [18]. This research established the four signals used in the job interview analysis as potentially meaningful measures of interaction. The study, involving 38 participants, considered mock workplace negotiations dealing with workplace reassignment, salary, and health care provided to the mock employee. The developed classifier predicted negotiation outcomes with approximately the same accuracy as experts and indicated that the significance of features differed greatly depending upon which role displayed them. For example, in negotiations with positive outcomes for the lower-power negotiator, engagement and stress were critical features of the higher-power negotiator, while mirroring was a critical attribute of the lower-power negotiator.

Automatic prediction of attraction is considered within the convenient context of speed-dating, which provides each daters desired interest. Researchers in [15] performed early, automatic analysis of speed-dating. Using the same four social signal features as before, the experimenters analyzed 60, five-minute speed dates. They predicted female interest at 72% accuracy, with engagement being the strongest indicator. Use of data from individuals in both testing and training resulted in an increase of accuracy to 87.5%. The researchers also attempted to predict responses to followup questions using both linear and RBF kernel SVMs with accuracy ranging from 62% to 82%.

In a later experiment by Ranganath et al., audio from 1100 four-minute speed-dates was used in an analysis of interactional style [21]. The research focused on both linguistic information, acquired through manual transcription of audio conversations, and paralinguistic behavior, acquired automatically with the aid of manual timestamps. The researcher's primary paralinguistic focus was on prosody, but also considered total speaking time and rate of speech. Backchannel (e.g., 'yeah' or 'uh huh'), appreciations (e.g., 'nice' or 'cool'), and questions were shown to play the most vital roles. The researchers utilized a deep autoencoder in order to circumvent issues introduced by the high-dimensionality of lexical features, and in doing so reduced the 1000 most commonly used words to a set of 30 high-level features. The researchers were able to predict intended and perceived flirtation with accuracy ranging from 69% to 79.5%, which was in all cases superior to classification without the autoencoder as well as to that of human classifiers.

The research discussed has considered audio features and to a lesser extent dyadic features (e.g., turn taking), but speed-dating research has additionally been conducted that considers network features [19] as well as interaction analysis based upon top-down video [21]. To the best of the author's knowledge, no frontal-video speed-dating automatic analysis has been performed at large scale.

Presented research analyzes speed-dating interaction using exclusively paralinguistic features. Realistically, the contributions of this research are so far minimal; however, it does considered two interesting problems, draw valuable conclusions about paralinguistic features, and establish a basis from which future related research may grow. Two notable problems this paper considers are (1) how to best perform automatic speaker diarization in audio extracted from human-noisy environments with exactly two non-collocated microphones and (2) how to best predict skewed-class interaction outcomes based upon sparse interaction patterns. The valuable conclusions this paper draws with respect to BSDS-specific research are (1) that paralinguistic features extracted through current means from the noisy environment of BSDS data are unlikely to provide sufficient information if any at all and (2) any method of interaction classification applied to BSDS data must be highly robust against noise. The remainder of this paper is organized as follows: (1) BSDS dataset (2) feature definitions (3) feature extraction (4) feature selection (5) classification models and techniques (6) results and baselines (7) discussion and (8) conclusion and future work.

3 Dataset

The Berlin Speed Dating Study (BSDS) dataset, described in detail in [2], was collected at Humboldt University, Germany during a five-month period in which 17 speed-dating sessions were held. 190 men and 192 women aged 18–54 years (mean=32.8, SD= 7.4) participated in the study, with each session involving 17-27 participants. Before each session, all participants had audio and visual samples taken in addition to filling out a pre-event survey, which collected primarily demographic information (e.g., age, level of education, income) in addition to other personal information (e.g., openness to experience, extraversion, shyness). Dates were carried out in booths containing two chairs, with a microphone positioned next to each participant and a video camera positioned across from each participant at an angle. During the event, male participants moved between booths to participate in a 3-minute speed date with each female participant. After each date, both participants recorded their interest in future dates. At the end of the event, participants were allowed to reconsider their choices. Two follow-up surveys were conducted, one 6 months after the event and one 12 months after. Both collected information about interactions resulting from the speed-dating event.

The BSDS data is unusual not only in it's breadth and depth of information collected about participants, but also in it's sample size. To the best of the author's knowledge, the BSDS is the largest speed-dating event with at least audio data, containing 2160 speed-dates. Unlike the studies addressed in the related works section, the BSDS collected frontal video of participants. A large subset of these videos were analyzed by an unrelated group of non-german-speaking participants, providing a human baseline for date-outcome prediction [20].

The BSDS data differs from the SpeedDate Corpus used in [21] in that successful speed-dates occurred with much less frequency. Approximately 11% of the BSDS dates resulted in mutual interest, which translates into significantly imbalanced outcome classes. For long-term prediction this percentage falls even further to 6%. In contrast, the SpeedDate corpus likely had a much higher positive outcome ratio. 56.3% of men and 37.4% of women responded affirmatively in the SpeedDate Corpus, while in the BSDS, 36.8% of men and 32.4% of women responded affirmatively [19]. This discrepancy could be a result of a number of factors, for example the differing average age of subjects or the environment of the study.

4 Feature Descriptions

The INTERSPEECH 2013 Computational Paralinguistic Challenge feature set was used as the full size feautre set[23]. The set includes energy, spectral, cepstral, voicing, harmonic-to-noise, spectral harmonicity, and psychoacoustic spectral sharpness related features. In total it contains 6,373 features, which are primarily functionals (e.g., mean, std dev) applied to low level descriptors (LLDs) and their derivatives. LLDs are defined as values extracted directly from "raw" data and generally reflect information computed on "chunks", or temporally-defined subsets, of data [22].

Audio LLDs are generally categorized into prosodic, voice quality, and spectral features. Prosodic features encompass fundamental frequency (F0), speech energy, duration and other qualities, and describe the manner in which a person talks. F0 is the lowest frequency present in a waveform and in speech reflects the vibration of vocal cords. F0 plays a critical role in the manner in which people perceive the pitch of speech [8]. Speech energy refers literally to the amount of energy carried by a wave and is reflected in its amplitude. Energy is used for a number of speech analysis tasks, but is particularly common within speech segmentation [13]. Voice quality measures deviations of speech from the underlying wave signal. Major features include jitter, shimmer, and harmonics to noise ratio (HNR). Jitter reflects F0 variation over wave period while shimmer reflects wave amplitude variation [10]. HNR refers to the portion of an audio signal contributing to the underlying periodic wave. Signal spectrum refers to the distribution of signal energy as a function of frequency. Common features include Mel-frequency cepstrum (MFC) and corresponding coefficients (MFCC).

5 Feature Extraction

In order to derive meaningful information from the BSDS audio data, the conversational turns were extracted so that features could be computed for only one interlocutor. This task, refered to as speaker diarization, consists of two subtasks: (1) speech recognition (segmenting audio files into speech and nonspeech segments) and (2) speaker recognition (ascribing the resulting speech segments to specific speakers) [25].

Speech segmentation is traditionally performed using energy and spectrum based methods, though many approaches exist. Speaker recognition is performed on speech chunks using a variety of methods, with the most common approach being maximum likelihood classification with Gaussian Mixture Models (GMM). Speaker diarization research has traditionally revolved around audio from two domains: (1) broadcast news [5] and (2) telephone conversation [14]. More recently, speaker diarization of group meetings has become popular, specifically methods that utilize multiple audio inputs in tandem [1].

A variety of openly available software exists for the purpose of performing speaker diarization. For example, the LIUM speaker diarization toolkit includes a means of performing speaker diarization on broadcast news data [16]. Other tools include the SHoUT Toolkit, ALIZE, and DiarTK, which focus on conversational analysis [26][25].

In processing the BSDS data, however, these toolkits were either unable to accurately extract and classify speech segments, unavailable for download, or seemingly inoperable. This first issue is potentially due to four factors. First, the BSDS data contains significant background conversation that frequently becomes as loud or louder than the conversation of interest. Second, microphones capturing data during the speed-date each individually captured one interlocutor with much higher intensity than the other due to their proximity to individual participants. Third, the BSDS data may contain greater amounts of and more rapid speech overlap than anticipated by these systems. And Fourth, these systems, which were generally trained and applied to english conversation, don't work as well on conversation in German. The unequal sound intensity and presence of two microphones allowed for rough approximations of speech segments. This was done by extracting speech energy from the sound files using openSMILE [9] and then subtracting one audio file's loudness from the other. The resulting positive values above a certain threshold were then ascribed to one speaker and the negative values below a certain threshold ascribed to the other speaker. This threshold was set in a variety of ways, but ultimately a constant number of standard deviations from the mean loudness value was used. The results from this method of speaker diarization can be improved upon greatly, but, given time limitations, this approach provided a reasonable option

Once speaker diarization was complete and each date had two corresponding audio files segmented by speaker, feature extraction was performed using the openSMILE toolkit [9]. This produced 12,746 features for each speed-date, with half corresponding to each gender. These were combined with the target value - the outcome of the speed-date - to complete the sample. A similar approach was taken in [29] in order to perform the fusion task. This approach is classified under feature-level fusion, which is the aggregation of information prior to system training and classification [3]. The alternative to feature-level fusion is decision-level fusion. In this case, rather than combining features from both daters and predicting the overall outcome, decision-level fusion predicts each daters decision individually, and determines the final outcome based upon that information. Decision-level fusion is likely superior to feature-level fusion in the case of speed-dating since it mitigates imbalanced class issues, and will be used in the future.

6 Feature Selection

Feature selection is the process of selecting a subset of features for use in classification, thereby reducing the dimensionality of the feature set. Three primary approaches exist for reducing the dimensionality of the feature set or discovering low-dimensionality, high-abstraction feature sets: (1) ad hoc feature selection (2) automatic feature selection and (3) feature learning. Due to the relatively high dimensionality of the feature set, any classifier trained using the original 12746 features would likely suffer from the high dimensionality/small training set problem. Reduction of dimensionality was therefore desirable since it diminished the likelihood of overfitting as well as reduced model training time.

Ad hoc feature selection is the selection of features based upon researcher domain knowledge. This approach provides a convenient manner for dimensionality reduction though suffers from the need for experience, thereby hindering generalization of machine learning methods. A small subset of features (table 1) were selected based upon the results of other speed-dating and interaction studies [21] [29].

The second approach to dimensionality reduction is the algorithmic selection of features. Principal Component Analysis (PCA) is a commonly used method for selecting subsets of features which account for the largest amount of variability in the original data. PCA

Table 1: Ad Hoc Paralinguistic Feature Weightings

f_F0_maxPos	-0.0269
f_F0_stddev	-0.0071
f_F0_minPos	0.0455
f_RMSenergy_peakRangeAbs	0.0953
f_logHNR_amean	-0.1994
f_voicing_maxPos	-0.1182
f_mfcc_amean	-0.6102
f_mfcc_flatness	-0.8758
m_F0_stddev	-0.4524
m_F0_minPos	-0.0693
m_RMSenergy_maxPos	-0.0016
m_RMSenergy_minRangeRel	-0.0683
m_logHNR_amean	-0.0291
m_voicing_maxPos	0.1049
m_mfcc_one_amean	0.7877
m_mfcc_one_flatness	0.6176

Ad Hoc features are shown above. The Male and Female MFCC (in blue) are both weighted heavily, with female values being negative.

was used to reduce the dimension of the samples to 120 features. This is a large dimensionality reduction, reflecting two orders of magnitude fewer features, but was chosen because it empirically resulted in individual features capturing significant portions of the data variation.

The third method for feature space reduction is feature learning. Feature learning, or representation learning, is a general technique referring to the act of finding data representations which exhibit certain "good" properties. These properties are covered in detail in [6], but informally, "good" properties are those that assist in the discovery and disentanglement of the causes of variation in a dataset.

7 Classification

Two models were used for classification - a multilayer perceptron (MLP) and a support vector machine (SVM). The MLP was implemented in Theano, a mathematical expression library [7] and consisted of three layers. SVMs with both linear and radial basis function (RBF) kernels were used in classification. Of the 2160 speed-dates, 283 were used in this classification task. This choice of smaller sample size was motivated by processing and quality-related factors. Each video was subdivided into 30 second segments, resulting in 1700 data points originating from 6 different speed-date sessions. Due to processing issues, there is an unknown distribution of samples containing each participant in each of the train, test, and cross validation sets. This is obviously not desirable and will be changed in future research.

Prior to classification, feature values were standardized to zero mean, unit variance due to the negative impact unscaled features have on the effectiveness of SVMs [11]. For MLP classification, a 60%-20%-20% split was used for train, cross validation, and testing data. We used stratified 5-Fold cross validation for testing the SVM models. Stratified K-Fold was used in order to prevent exasperation of issues introduced by class imbalance.

In order to assess classifier accuracy, classification error as well as precision and recall are reported. Precision refers to the ratio of items correctly labeled as a certain class to total items labeled as that class. Recall refers to the ratio of the number of samples of a class correctly classified to the total instances of that class. These measures are not susceptible to imbalanced class issue because they provide class-specific, rather than population-wide, measurements.

8 Results

Classification results are listed in table 2 along with two baseline estimates. The first baseline is that of

Table 2: Classifier Results: Paralinguistic Features

Classifier or Baseline	Overall Accuracy	Precision	Recall	F1 Score
Human Baseline	38.40%	-	-	-
Negative Baseline	88.70%	0	0	0
MLP (all paralinguistic features)	87.95%	-	-	-
MLP (pca to 120 paralinguistic features)	88.53%	-	-	-
MLP (ad hoc paralinguistic features)	87.95%	-	-	-
Linear SVM (all paralinguistic features)	88.96%	0.26	0.17	0.2
Linear SVM (pca to 120 paralinguistic features)	69.48%	0.17	0.46	0.24
Linear SVM (ad hoc paralinguistic features)	57.04%	0.1	0.49	0.17
RBF SVM (ad hoc paralinguistic features)	86.15%	0.21	0.17	0.19

Negative baseline represents predicting no for each date. Note that many overall accuracy values are similar to this baseline, indicating they may be performing the same function.

always predicting negative date outcomes. Since only 11% of dates had positive outcomes, this method results in 89% classification accuracy. The second baseline comes from human classifiers [20]. Human classifiers range from 55% to 65%accuracy depending upon which chronological section of the interaction they observe. This baseline reflects human ability to predict, based on both audio and visual information, romantic interest of individual daters. Unfortunately, the paralinguistic features collected reflect date outcomes rather than individual interest, and for this reason this baseline cannot be used in its original form. If we assume that each date is equally likely to be predicted, squaring the original value provides an approximate human baseline. Precision, recall, and F1 score for the human baseline are not available. Human raters did not perform their ranking on all the videos used in this research, so in using this baseline we make the assumption that it applies to speed-dates in general.

9 Discussion

The tendency of the MLP to approach training classification error of 11.3% indicates that it may be learning to negatively classify all examples. Since the classes are heavily imbalanced, the cost of misclassifying all of one type is overcome by the benefit of correctly classifying all of the other type. It is also possible that the MLP performs in this manner because of the feature data. In general, paralinguistic features possess predictive information [18], but poor quality speaker segmentation may render them useless.

Results from the SVM support the claim that the data does not contain sufficient information for accurate classification. Using a RBF kernel, the SVM, regardless of class weight or parameter selection, classifies all samples either negatively or positively when all 12,000 or the reduced 120 features are used. With ad hoc features the RBF kernel SVM was able to distinguish some positive and negative samples, though it still did not perform well, having an outlier class F1 score of .19. This indicates that the classifier lacks the necessary information to correctly predict date outcomes. The linear kernel SVM performed similarly to the RBF Kernel, though did achieve a higher outlier class F1 score of .24 when trained using the 120 reduced features. These results may indicate that the data does contain some useful level of information, though it may also reflect the weighting applied to the outlier class.

The linear kernel SVM ad hoc feature weightings are reported in table 1. The highest weight was applied to the first MFCC coefficient mean and flatness values (positive for male, negative for female). Flatness is generally used to measure the degree to which a signal reflects either tone or noise and in this case an inverse relationship exists between the male and female versions of these values. It's possible this relationship is a meaningless reflection of the poor data quality, a possibility made more likely by MFCC's general susceptibility to signal noise. Assuming the feature data contains useful information, however, these results may indicate that a positive relationship exists between time spent talking by males and interaction outcomes. This is infered from MFCC's role in reflecting the prescence of speech [28] (of course, only the first MFCC coefficient was used, which detracts from this claim). This conclusion is not supported by related studies, for example [12], which showed that increased time spent speaking

by males translated into a lower likelihood of romantic or friendly interest.

10 Conclusion

The presented research illustrates that paralinguistic features extracted using rudimentary speaker diarization methods from high-human-noise environments do not provide a basis for consistent interaction classification. More accurate methods of speech segmentation and paralinguistic information extraction may improve classification, but even with these refinements, multimodal analysis (i.e., incorporation of visual, network, demographic and other features) will provide otherwise unattainable improvements in classification accuracy.

Even such a system would, however, still be reliant upon essentially ad hoc features, whether researcher or algorithmically-defined. Feature selection in this manner is undesirable not only because it requires domain-specific expertise, but also because it limits the classification system to human-defined abstractions. This limitation is particularly costly in recognizing patterns in human-interaction because of the low likelihood that any combination of human-defined abstractions would be capable of accurately accounting for the wide variety of rarely occurring, yet important information present in an interaction. For this reason, a method of applying representation learning to raw audiovisual data in order to define high-level features of human-human interaction is highly desirable. Whether such a method is feasible is a difficult assessment to make, but its development would almost certainly revolutionize interaction classification.

11 Appendix A: Outcome and Network Features

11.1 Outcome Features

Outcome features are derived from information provided after the speed-date by the participants. In the BSDS response sheet, participants provided three pieces of information: (1) their initial choice concerning the date (2) their final choice (participant's could change their decision from their initial choice) and (3) the choice they predicted the other participant, their date, would make. Based upon this information, three features were developed: (1) awareness (2) confidence and (3) decisiveness. Awareness is the fraction of correct predictions by the participant. Confidence is the fraction of affirmative predictions (i.e., predictions that the other person would request to meet again). Decisiveness is the fraction of decisions in which the participant changed his/her mind between initial choice and final choice. The assumption is that these features could also be derived from other sources (e.g., self-reporting) and could therefore be useful in practice.

A fourth feature called selectivity was initially defined which reflected the fraction of dates to which a person chose negatively. This factor was removed since it seemed to "unfairly" reflect the target value. Removing this feature ultimately made little difference in accuracy predictions (1.5% increased error without it).

11.2 Network Features

Network features are derived from information contained within the network of relationships existing in the speed dating event. While these features do not necessarily exclude information about the specific individual being considered, the features considered here do. As a result, these features make inferences about a specifc date without using any information from that date. For these features, I simply replicated the approach taken by [19], though left out components they stated had minimal impact. Three components make up each feature: (1) a weight representing the similarity between an individual and a member of the same gender (2) a value derived from the relationship between that member of the same gender and the member of the opposite gender involved in the current date and (3) a method of aggregating all combinations of (1) & (2).

1. Weights were defined in two ways. The first was a weight dependent on the number of people liked in common by you and another member of your gender. The second was a weight dependent on the number of people that like both you and the other member of your gender. These weights were developed by taking the projection of the bipartite graph on either the male or female node sets and representing the weights on the new edges with either of these two types of weights (see figure 1 and 2).

2. The value that was multiplied by the weight was defined as +1 if an edge existed between the the

date and the other person of the same gender and -1 if no edge existed.

3. Two methods of aggregation were used by [19]: (1) the max of all weight*value combinations and (2) the sum of these values. They found sum to be more predictive and so I elected to only use that method.

They also defined two other features (1) 'num that like you' and (2) 'she likes you', which I did not include.

11.3 Results

Note that for these features, I attempt to predict the response of only one individual rather than both (before I was trying to predict date outcomes. Here I only predict the choices of individuals). This approach results in more balanced classes and arguably allows for more easily interpreted results. It also results in different feature weightings for each gender. Each sample had a total of 10 features. There were 2160 samples for each gender. More samples were used in this portion than for the paralinguistic features because I have available all the information required for outcome and network samples.

The classification was performed using stratified 10-fold cross validation and since there is much less class imbalance in this case, the precision, recall, and F1 score are reported as averages over both positive and negative target values (classes). Unless otherwise noted, the best SVM classification uses an RBF kernel.

For predicting male choices, the best classifier used both outcome and network features, while for predicting females, the best predictor used exclusively network features. Male choices (72.4%) were significantly harder than female choices (83.5%) to predict. This trend aligns with the relative difficulty in predicting male and female choices based on conversational features [21]. Interestingly, predicting a males choice is more difficult than predicting a males prediction.

These classification results indicate that a male's similarity with his peers is less informative than a female's similarity with her peers (by either measure of liking the same people or being liked by them). The ability to predict female choices with 83.5% accuracy using only four network features is somewhat surprising and confirms that network features are worth pursuing.

The custom outcome features proved not to be very effective. One notable outcome feature result is that the most heavily weighted outcome feature for predicting female outcomes was female awareness (fraction of correct predictions), with a large negative weighting. This indicates that the worse a female individual is at predicting the male's response, the more likely she is to choose to go on another date. The highest weighted feature in predicting male choice was male confidence (fraction of times he predicted the female selected to go on another date). This indicates a male's decision is more heavily impacted by what he believes the female's decision will be than is the female's decision is by the male's - no surprise there.

Network features are a promising avenue for predicting outcomes. One highly desirable statistic given a speed-dating network would be the two individuals best suited for each other by certain metrics. It seems plausible this information can be predicted based upon network features, and I intend to investigate it next.

12 Appendix B: Data Pipeline

Data was transferred from storage in Dropbox to AWS EC2 instances. Movie files were converted to audio using FFmpeg and this audio was converted to a format usable by openSMILE by SoX (Sound eXchange).

Table 3: Classifier Results: Network & Outcome Features

Classifier	Overall Accuracy	Precision	Recall	F1 Score
SVM (female, network)	83.50%	0.82	0.77	0.79
SVM (female, outcome)	68.60%	0.51	0.41	0.45
SVM (female, network + outcome)	83.01%	0.81	0.80	0.80
SVM (male, network)	61.85%	0.44	0.29	0.35
SVM (male, outcome)	56.90%	0.43	0.47	0.45
SVM (male, network + outcome)	72.40%	0.70	0.76	0.72

Female choice prediction is performed more accurately with network features than male choice prediction. SVM (male, network) has an unusually low yet correct recall value.

Figure 1: Directed, Bipartite Graph

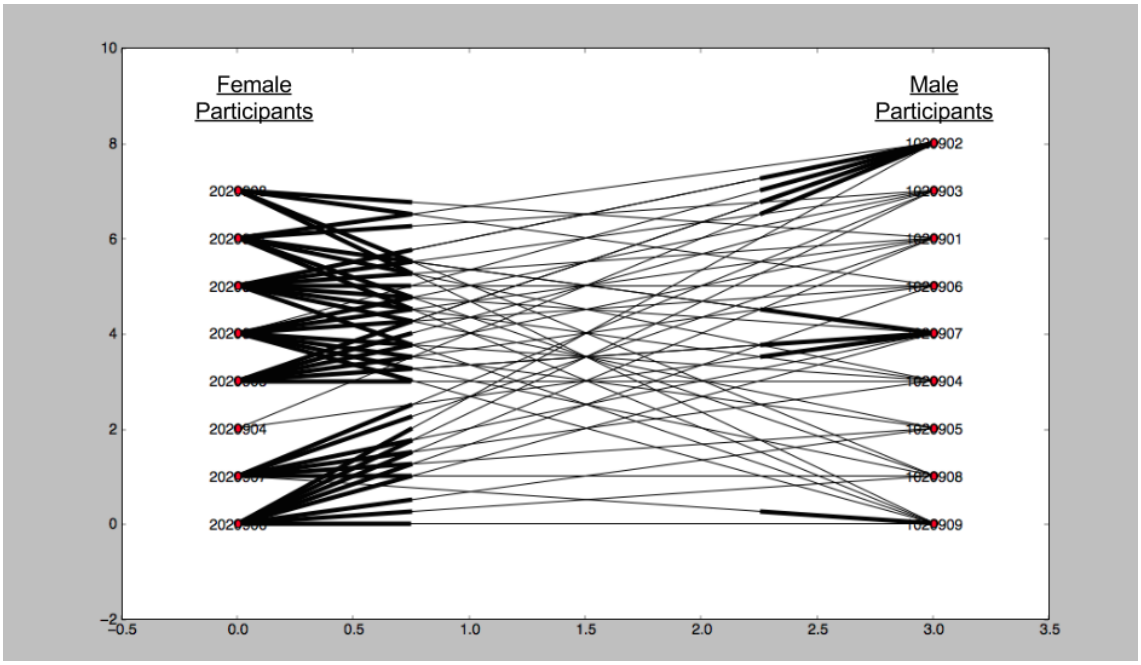
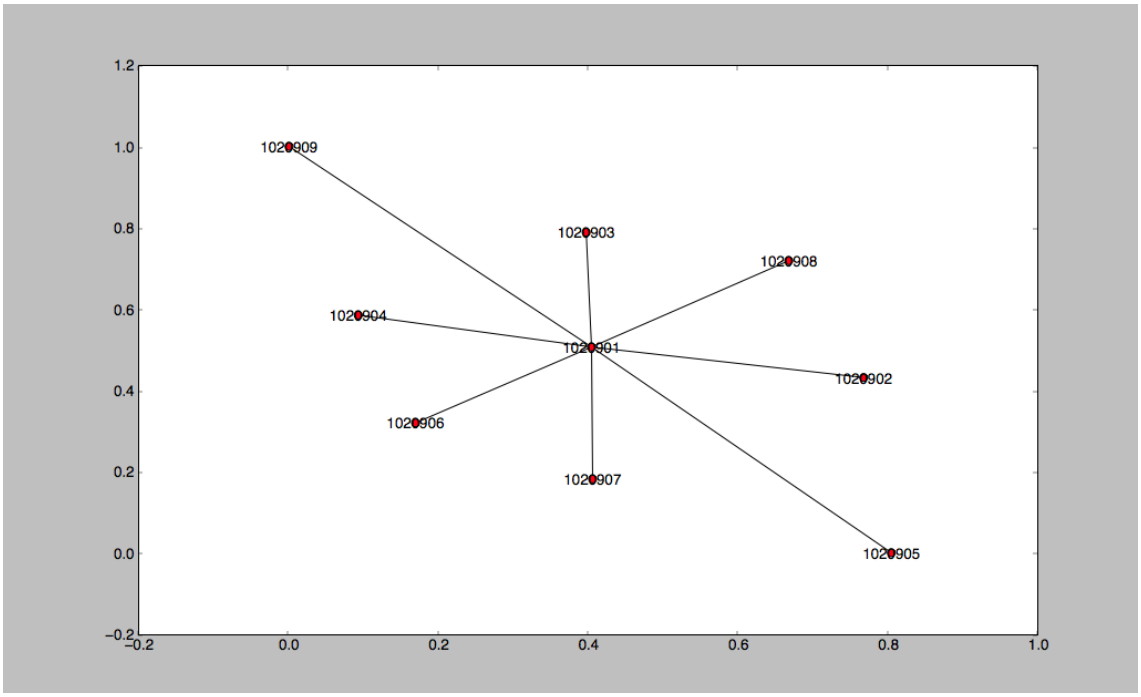


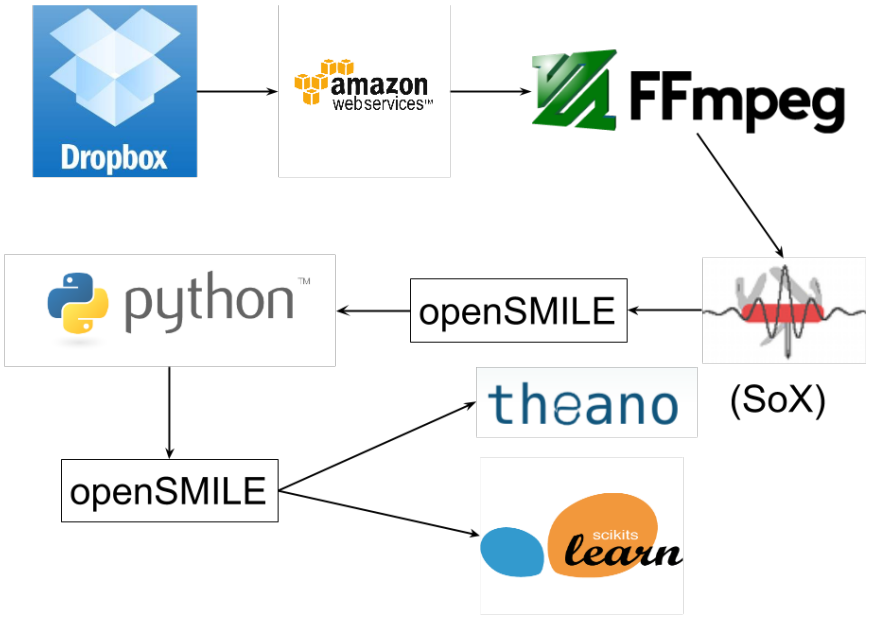
Figure 2: Corresponding Undirected, Weighted Projection



OpenSMILE was used initially to extract speech energy from each pair of audio files from each date. Speaker diarization was performed using speech energy values in python. Segmented files were then passed back to openSMILE in order to extract audio features. Extracted features were then passed to theano for MLP classification and scikit-learn for SVM classification.

Converting the 40GB of video data into a single file containing data samples (features and corresponding target value) required about 60 hours of processing time. Profiling the code showed that most of that time was spent running openSMILE and FFmpeg or in file input/output. The actual data samples file ultimately reflected only a fraction of the dates due to a variety

Figure 3: Data Process Flow



of issues and mistakes on my part.

13 Appendix C: Future Work

Areas of future work include improving audio feature extraction, extraction of visual and dyadic features, proportional weighting of data samples, and representation learning.

13.1 Improving Audio Extraction

The speaker diarization systems mentioned in the paper are the best option for improving sound extraction. I plan to take another look at DiarTk. It will require training a speaker recognition model. I thought this would require too much time during the semester, though in hindsight not trying it may have been more costly. Either way I'm looking forward to trying it out.

13.2 Visual Features

Visual feature analysis would entail analyzing the date videos in to order to extract visual features. This problem consists of first detecting people in images and then extracting features. Visual analysis is generally grouped into face & eyes and positioning & gestures. I intend to focus initially on face & eyes. Facial features include (1) facial action units (2) positioning (3) direction and (4) mouth openness. Features related to eyes include (1) gaze direction (2) motion and (3) openness. There are a number of approaches to extracting these features as well as many tools. I intend to use OpenCV.

13.3 Dyadic Features

Dyadic features describe the interaction of two individuals. For example, [29] utilized dyadic features such as mutual/nonmutual lean forward or smiling in an attempt to predict friendship. These features can be automatically extracted in certain cases provided necessary low level features are available. I'll extract dyadic features once I've completed visual features.

13.4 Proportional Weighting of Data Samples

Machine learning systems are often evaluated on their ability to generalize to data they have not encountered. Emphasizing this metric generally translates into training classifiers on data from individuals who do not also appear in testing data. It may be desirable, however, to develop a system capable of adaptation to patterns exhibited by a single individual in situations where that individual may be frequently reusing a system, for example SSP. One way to adapt a system in this way is

to use that person's data in training and to weight that data so as to emphasize it's significance to the model. This approach can also be used on data that is, by some chosen metric, similar to the user data (as seen in the network section).

13.5 Representation Learning

Representation learning consists of a set of method for creating low-dimensional, high-level features from low-level data that accurately reflect variation. This includes, but is not limited to, deep learning methods. A couple of the papers referenced in this report used representation learning and saw significant classification improvement. Making interaction outcome prediction amenable to representation learning may not be possible given current capabilities (it's almost certainly not in the broad sense), but it is nevertheless an interesting prospect and something I intend to pursue.

14 Acknowledgements

The Cardiff Conversational Database (CCDb) [4] was helpful in developing the system.

References

[1] Xavier Anguera, Chuck Wooters, and Jose M Pardo. Robust speaker diarization for meetings: Icsi rt06s meetings evaluation system. In *Machine Learning for Multimodal Interaction*, pages 346–358. Springer, 2006.

[2] Jens B Asendorpf, Lars Penke, and Mitja D Back. From dating to mating and relating: Predictors of initial and long-term outcomes of speed-dating in a community sample. *European Journal of Personality*, 25(1):16–30, 2011.

[3] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[4] Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 277–282. IEEE, 2013.

[5] Claude Barras, Xuan Zhu, Sylvain Meignier, and J Gauvain. Multistage speaker diarization of broadcast news. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1505–1512, 2006.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. 2013.

[7] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for scientific computing conference (SciPy)*, volume 4, 2010.

[8] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[10] Mireia Farrús, Javier Hernando, and Pascual Ejarque. Jitter and shimmer measurements for speaker recognition. In *INTERSPEECH*, pages 778–781, 2007.

[11] George Forman, Martin Scholz, and Shyamsundar Rajaram. Feature shaping for linear svm classifiers. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 299–308. ACM, 2009.

[12] Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646. Association for Computational Linguistics, 2009.

[13] Thomas Kemp, Michael Schmidt, Martin Westphal, and Alex Waibel. Strategies for automatic segmentation of audio data. In *Acoustics, Speech,*

and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, volume 3, pages 1423–1426. IEEE, 2000.

[14] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo. Diarization of telephone conversations using factor analysis. *Selected Topics in Signal Processing, IEEE Journal of*, 4(6):1059–1070, 2010.

[15] Anmol Madan, Ron Caneel, and Alex Pentland. Voices of attraction. 2004.

[16] Teva Merlin, Elie Khoury, Mickael Rouvier, Gregor Dupuy, Sylvain Meignier, and Paul Gay. An open-source state-of-the-art toolbox for broadcast news diarization. In *INTERSPEECH*, number EPFL-CONF-192762, 2013.

[17] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S Huanag. Human-centred intelligent human? computer interaction (hci^2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

[18] Alex Pentland. Social dynamics: Signals and behavior. In *International Conference on Developmental Learning*, volume 5, 2004.

[19] Emma Pierson and Andrew Suciu. Predicting speed-date outcomes with conversational and network features.

[20] Skyler S Place, Peter M Todd, Lars Penke, and Jens B Asendorpf. The ability to judge the romantic interest of others. *Psychological Science*, 20(1):22–26, 2009.

[21] Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. It's not you, it's me: detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 334–342. Association for Computational Linguistics, 2009.

[22] Björn Schuller, Florian Eyben, and Gerhard Rigoll. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In *Perception in multimodal dialogue systems*, pages 99–110. Springer, 2008.

[23] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings of Interspeech*, 2013.

[24] Vikrant Soman and Anmol Madan. Social signaling: Predicting the outcome of job interviews from vocal tone and prosody.

[25] Sue E Tranter and Douglas A Reynolds. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557–1565, 2006.

[26] Deepu Vijayasenan and Fabio Valente. Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings. In *INTERSPEECH*, 2012.

[27] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[28] Thurid Vogt and Elisabeth André. Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*. Citeseer, 2006.

[29] Zhou Yu, David Gerritsen, Amy Ogan, Alan W Black, and Justine Cassell. Automatic prediction of friendship via multi-model dyadic features.