Multimodal Prediction of Psychological Disorders: Learning Verbal and Nonverbal Commonalities in Adjacency Pairs

Zhou Yu* **Stefen Scherer†** **David Devault†** **Jonathan Gratch†**

Giota Stratou† **Louis-Philippe Morency†** **Justine Cassell***

*School of computer Science
Carnegie Mellon University
{zhouyu,justine}@cs.cmu.edu

†Institute for Creative Technology
University of Southern California
{scherer,devault,gratch,
stratou,morency}@ict.usc.edu

Abstract

Semi-structured interviews are widely used in medical settings to gather information from individuals about psychological disorders, such as depression or anxiety. These interviews typically consist of a series of question and response pairs, which we refer to as *adjacency pairs*. We propose a computational model, the Multimodal HCRF, that considers the commonalities among adjacency pairs and information from multiple modalities to infer the psychological states of the interviewees. We collect data and perform experiments on a human to virtual human interaction data set. Our multimodal approach gives a significant advantage over conventional holistic approaches which ignore the adjacency pair context in predicting depression from semi-structured interviews.

1 Introduction

Recent advances in the fields of verbal and nonverbal behavior analysis are revolutionizing our ability to analyze and understand people's behavior. One promising application is the automatic analysis of nonverbal behaviors associated with psychological disorder. Extensive research in behavioral sciences has demonstrated a link between specific psychological disorders, such as depression, and patterns of verbal and nonverbal behavior (Ellgring, 1989). Recognizing these verbal and nonverbal indicators, however, requires expert judgements from trained clinicians. The factors underlying these judgements are not easily quantifiable (Ellgring, 1989). Automatic detection of verbal and nonverbal indicators can assist clinicians by supporting their interview processes and providing more systematic, quantified measurements. Moreover, fully-automated techniques can serve as a pre-screening instrument for patients, complementing the self-reported questionnaires which are currently used for this purpose.

Psychological assessment interviews consist of a series of "question" and "response" pairs, which are consecutive utterances that we refer to as adjacency pairs (Schegloff, 2007). The potential "response" doesn't have to be a direct answer, but could be a counter-question or other form of response triggered by the "question", as long as it satisfies Grice's conversational maxim of relevance (Grice, 1975). Different adjacency pairs serve different purposes in triggering subject responses, and a model that considers context could better predict psychological disorders. We propose a computational approach that leverages the advantage of verbal and nonverbal behaviors extracted at the adjacency pair level to support a more contextualized analysis, unlike previous approaches which ignored context (Cohn et al., 2009), or only consider context in single feature analysis (DeVault et al., 2013).

Based on Hidden Conditional Random Fields (HCRFs) (Quattoni et al., 2004), we propose a new computational model, the Multimodal HCRF. HCRFs allow us to learn verbal and nonverbal commonalities among adjacency pairs automatically. For example, one specific commonality is that depressed people have a lower speech rate compared to non-depressed people in their responses to a large set of probing questions (see section 8.4 for details). In order to assess the effectiveness of incorporating adjacency pair into our analysis, we performed experiments on a corpus of 130 human to virtual human interviews, where the question was always asked by the virtual human interviewer, and the response was given by the real human. Our analysis relies on a model which brings together behaviors from multiple modalities: visual, acoustic and conversational and results showed a significant improvement for our multimodal computational model over previous models at predicting depression.

We first review previous work and our hypothe-

ses before we describe our dyadic interaction data set. After that we introduce automatically extracted multimodal features that capture verbal and nonverbal behaviors. Next, we present our computational model and experiments to validate it. Finally, we further analyze the results from our experiments.

2 Related Work

Many previous studies have examined the links between nonverbal behaviors and clinical conditions (Ellgring, 1989; Cohn et al., 2009). Little progress has been made towards identifying any clear links between patient disorders and expressed behaviors. This is due to the difficulties of manually annotating gestures and facial expressions, inconsistent measurements of nonverbal behaviors across studies and differences in social contexts of the interview processes between studies.

There is a general consensus regarding the relationship between certain clinical conditions (especially depression and social anxiety) and associated verbal and nonverbal cues. Emotional expressivity, such as the frequency or duration of smiles, is diagnostic of psychological disorders. For example, depressed patients frequently display flattened or negative effects, including less emotional expressivity (Perez and Riggio, 2003; Bylsma et al., 2008), fewer mouth movements (Fairbanks et al., 1982; Schelde, 1998), more frowns (Fairbanks et al., 1982; Perez and Riggio, 2003) and fewer gestures (Hall et al., 1995; Perez and Riggio, 2003). Some findings suggest that the quantity of expressions may not be as important as their dynamics. For example, depressed patients may frequently smile, but these smiles are perceived as less genuine and often shorter in duration (Kirsch and Brunnhuber, 2007). Social anxiety and PTSD share some features with depression, such as a tendency for heightened emotional sensitivity and more energetic responses. Such responses can include startlement and a greater tendency to display anger (Kirsch and Brunnhuber, 2007) or shame (Menke, 2011). Cohn and colleagues have identified increased speaker-switch durations as indicators of depression, and have explored the use of these features for classification (Cohn et al., 2009). Our current research builds on these findings as a step to overcome the difficulty of manually annotating human behavior.

Scherer et al. (2013b) explore the correlation between automatically quantified acoustic and visual features with psychological disorders. Stratou et al. (2013) find that the subject's gender plays an important role in automatic assessment of psychological conditions when analyzing automatically extracted visual features. DeVault et al. (2013) investigate the correlation between conversation features and psychological disorders, but don't take visual features into consideration. Cohn et al. (2009) use both facial expression and vocal prosody in identifying depression, however, they do not include more features which are predictive of depression. In summary, there is a lack of models that combine comprehensive conversational, visual and acoustic features related to depression. Also, the prediction methods used in previous works do not take the contextual information of the interview into account.

We include contextual information by modeling nonverbal behavior at the adjacency pair level. We apply HCRFs for classification, as opposed to Naive Bayes used in DeVault et al. (2013) and Stratou et al. (2013) because HCRFs model time contingency. HCRFs have been successfully used to tackled problems in computational vision and speech. For instance, Quattoni et al. (2004) applied HCRFs to model spatial dependencies for object recognition in unsegmented cluttered images.

3 Research Hypotheses

Interviews typically consist of a series of question and response pairs which we refer to as *adjacency pairs*. We present the two consecutive utterances as a tuple (q_i, r_i), where q is the "question" and r is the "response".

For each adjacency pair, subjects exhibit different verbal and nonverbal behaviors, for example, a different speech rate or facial expression. We hypothesize that:

1. We can better predict depression with a computational model that takes advantage of context by considering features quantified at the **adjacency pair level** rather than models using features extracted from the whole interaction. For example, we consider the speech rate in the response of the subjects in different adjacency pairs as opposed to the speech rate over the whole interaction in our model. The change of nonverbal behaviors exhibited in human responses to different stimuli (i.e. positive questions versus negative questions) are known to be significantly different between groups with and without psychological

disorders (Bylsma et al., 2008).

2. Adjacency pairs which serve the same probing purpose **share commonalities** in human verbal and nonverbal responses. By allowing our model to learn these commonalities we can improve prediction accuracy. For example, one commonality could be that for a set of adjacency pairs which concern a client's personal experience, people with psychological disorders have a longer latency in speech onset time to respond to the questions.

3. A comprehensive set of features from **multiple modalities** improves computational performance in predicting depression compared to a single or bi-modal approach. Previous works (Cohn et al., 2009; Scherer et al., 2013b; Stratou et al., 2013) combine different multimodal features, but none of these approaches make use of all three modalities (conversational, visual and acoustic). According to our previous research, multimodal features also improve friendship prediction (Yu et al., 2013). Although the tasks are different, we believe that leveraging multiple information channels can benefit depression prediction.

4 Distress Assessment Interview Corpus (DAIC)

We use a data set that has 130 semi-structured interviews in a Wizard-of-Oz paradigm between a human and the virtual character Ellie, depicted in Figure 1. Drawing on observations of interviewer behavior in the face-to-face dialogues, Ellie was designed to serve as an interviewer who is also a good listener, providing empathetic responses, back channels and continuation prompts to elicit extended replies to specific questions. The virtual human builds rapport with the participant at the beginning of the interaction with a series of casual questions about Los Angeles. After that, the conversation transitions towards intimate questions, like, "*Do you consider yourself more shy or outgoing?*". After the intimate phase, Ellie asks questions directly related to previous experiences of psychological disorders, such as, "*Have you been diagnosed with depression before*?". A series of positive questions, for example, "*How would your best friend describe you?*" are designed to leave the participant in a positive mood. Participants for the study were recruited via Craigslist and all applicants who met the requirements (i.e. age

Figure 1: Ellie, the virtual human

greater than 18, and adequate eyesight) were accepted. The mean age of the 130 participants in our data set was 38.41 years, with 69 males and 61 females. For a measure of psychological disorders, the PHQ-9 provides guidelines on how to assess the participants' conditions based on their responses to a questionnaire. Among the 130 participants, according to the PHQ-9, 30 participants were considered to have moderate depression or above (Kroenke and Spitzer, 2002) by having a cumulative score of ten or above. We consider them depression-positive in this study.

5 Automatically Extracted Multimodal Features

In this section, we briefly describe the features used in our experiments and the literature that motivates them. We focus on three types of features: conversational (Section 5.1), visual (Section 5.2) and acoustic (Section 5.3). All the features are extracted from the "response" part of an adjacency pair, as the "question" part of an adjacency pair is spoken by Ellie and is identical for all the subjects. We include only automatically derivable features in our analysis for the purpose of reducing manual annotation. In total, we use 16 features: 5 conversational, 3 visual and 8 acoustic.

5.1 Conversational Features

The system's speech segments, including starting and ending time stamps and verbatim transcripts of system utterances, were saved from the system log files. Motivated by DeVault et al. (2013), we selected the following features:

- **Speaking Rate and Onset Time** Slowed

speech and increased onset time were observed in previous clinical interviews of depressed individuals (Hall et al., 1995). We quantify the speaking rate by counting the number of words spoken per minute, and the speech onset time as the time delay before the user responds to Ellie's question. Here we use the manual transcription of the interview. However, it is possible for the output of the automatic speech recognition (ASR) system to be used as an approximation of the transcription, thus making the speech rate and onset time automatically obtainable.

- **Number and Average Length of User Segments** The utterances are automatically segmented by identifying long pauses and the average length of the user segments is quantified in seconds.

- **Filled Pause Rate** We count the number of times any of the tokens uh, um, uhh, umm, mm, or mmm appears in each speech segment. To account for the varying length of speech segments, we define the filled pause rate as the number of those tokens divided by the duration of the corresponding segment.

5.2 Visual Features

We selected three visual features based on work in Stratou et al. (2013):

- **Expression Variability** Based on a collection of clinical observations summarized in Ellgring (1989), the homogeneity of an affective level and total facial activity are considered good indicators of psychological disorders. Specifically, reduced facial behavior, or lack of emotional variability, has been reported as an indicator of depression. Our automatic feature extraction system includes the Computer Expression Recognition Toolbox (CERT) (Littlewort et al., 2011), which measures 8 basic expressions: Anger, Disgust, Contempt, Fear, Joy, Surprise, Sadness and Neutral. We measure emotional variability by considering the variances of all these expressions.

- **Neutral Expression** The frequency of the detection by CERT of a "Neutral" expression is a good measure of emotional "flatness", which mentioned in Ellgring (1989) as well.

- **Head Rotation** Clinical observations suggest reduced motor variability or motor retardation among patients suffering from depression (Ellgring, 1989). Hence, as an aspect of motor variability we look at head rotation variability as an indicator of psychological disorders. Our system for automatic analysis provides 3D head position and orientation based on the GAVAM head tracker (Morency et al., 2008) and CLM-Z face tracker (Baltrusaitis et al., 2012). Measuring the head rotation in all three directions (yaw, tilt and roll) allows us to calculate the head rotation.

5.3 Acoustic Features

Motivated by Scherer et al. (2013a) and Cohn et al. (2009), we extracted the following acoustic features with a sample rate of 100 Hz, using the lapel microphone recordings:

- **Energy in dB** The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms (i.e. 100Hz sample rate). Each speech window is filtered with a hamming window and the energy is calculated and converted to the dB-scale.

- **Fundamental Frequency (f_0)** In Drugman and Abeer (2011), a method for f_0 tracking based on residual harmonics, which is especially suitable in noisy conditions, is introduced. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering. This process reduces the influence of noise and vocal tract resonances. For each $r(t)$, the amplitude spectrum is computed, showing peaks for the harmonics of f_0, the fundamental frequency. These peaks form the basis for robust f_0 estimation.

- **Spectral Stationarity (ss)** To characterize the range of the prosodic inventory used over utterances, we make use of the so-called *spectral stationarity* measure. This measurement was used in Talkin (1995) as a way of modulating the transition cost used in the dynamic programming method used for f_0 tracking. Spectral stationarity, ss, is measured using the Itakura distortion measure (Itakura, 1975) between the current current and previous frame. We use a relatively long frame length of 60 ms (with a shift of 10 ms; sampling rate 100Hz) and frames are windowed with a Hamming window function be-

fore measuring ss.

- **Normalized Amplitude Quotient (NAQ)** This feature is derived from the glottal source signal estimated by iterative adaptive inverse filtering (Alku et al., 1992). The output is the differentiated glottal flow. The NAQ is the ratio between the negative amplitude of the main excitation in the differentiated glottal flow pulse and the peak amplitude of the glottal flow pulse normalized by the length of the glottal pulse period (Alku et al., 2002).

- **Quasi-Open Quotient (QOQ) and Open-Quotient Neural Net ($\mathbf{OQ}_{NN}$):** The QOQ is also derived from amplitude measurements of the glottal flow pulse (Alku et al., 2002). The quasi-open period is measured by detecting the peak in the glottal flow and finding the time points before and after this point that descend below 50% of the peak amplitude. The duration between these two time-points is divided by the local glottal period to get the QOQ parameter. As a novel alternative of the QOQ, we extract $\mathbf{OQ}_{NN}$, a parameter estimating the open quotient using standard Mel frequency cepstral coefficients and a trained neural network for open quotient approximation (Kane et al., 2013).

- **Harmonic Amplitude Difference** The difference in amplitude levels (in dB) between the first two harmonics of the narrow band voice source spectrum, which is an alternative rough estimate of the open quotient (Henrich et al., 2001).

- **Peak Slope** This voice quality parameter is based on features derived following a wavelet-based decomposition of the speech signal (Kane and Gobl, 2011). The parameter, named *peak*, is designed to identify glottal closure instances from glottal pulses with different closure characteristics.

6 The Multimodal HCRF Modal

A semi-structured interview changes according to the behaviors of the participants and is composed of a series of adjacency pairs. From a modeling perspective, semi-structured interviews have three main components: (1) an overall goal, which is specific to each interview (e.g., assessing depression or PTSD), (2) a conversational structure where some adjacency pairs share a common purpose and (3) a variation in human behavior during different adjacency pairs or sets of adjacency pairs. We propose a computational approach which explicitly models these three main components and addresses all the research hypotheses discussed in Section 3. Our approach is based on a Hidden Conditional Random Field (HCRF) (Quattoni et al., 2007) which is a probabilistic energy model that learns hidden commonalities automatically from a series of observations from adjacency pairs and their corresponding mappings to depression assessments. Each hidden state groups together adjacency pairs with similar function for the purpose of differentiating depressed people from non-depressed. We propose to adapt HCRF to automatically predict depression over the semi-structured interviews between humans and virtual humans.

Figure 2 depicts a graphical representation of our model. We wish to learn a mapping between

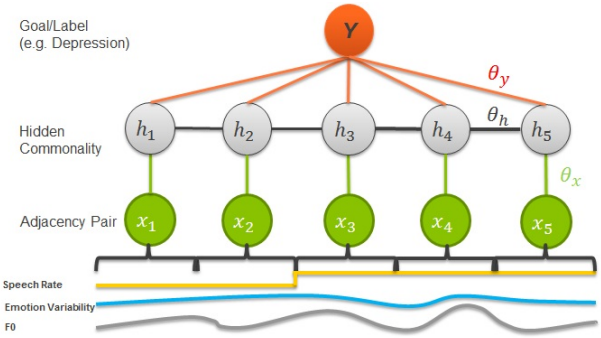

Figure 2: Multimodal HCRF

multimodal features $\boldsymbol{x}$=$\{x_1, x_2, ..., x_n\}$, defined in Section 4 and extracted at the adjacency pair level, and the class label y ∈ Y, which is either depressed or not. Our model is defined as

$$P(y|\boldsymbol{x}, \theta) = \frac{\sum_{\mathbf{h}} e^{\psi(y,\mathbf{h},\mathbf{x};\theta)}}{Z(\boldsymbol{x}, y)}$$

where $\mathbf{h} = \{h_1, h_2, ..., h_m\}$ are hidden states representing the commonalities between adjacency pairs. H is the set of hidden commonalities. The constant Z($\boldsymbol{x}$,y) is a partition function that serves as a normalization factor. The most important parts of the model are the potential functions, $\psi(y, \mathbf{h}, \mathbf{x}; \theta)$, parameterized by [θ_x θ_y θ_h]. We visualize these parameters in Figure 2 and describe them below:

1. The parameter θ_x models the relationship between multimodal features x_j and hidden states (commonalities) h_j. By analyzing the

amplitude of each of the weights in θ_x, it is possible to learn the relative importance of each feature for each hidden state. Adjacency pairs that map to the same hidden state form a group which share commonalities.

2. The parameter θ_y models the relationship between the hidden states h_j and the label y. By analyzing the weights of θ_y, it is possible to see which groups of adjacency pairs are important to predict depression.

3. The parameter θ_h represents the links between hidden states. It models the temporal dynamics in the hidden states (commonalities) of adjacency pairs.

In our experiments we used a Quasi-Newton optimization technique implemented in HCRF toolbox [1].

7 Experiments

We designed our experiments to evaluate our three hypotheses: (1) the effect of modeling semi-structured interviews at the adjacency pair level, (2) the importance of explicitly learning the commonalities between adjacency pairs, and (3) the importance of multimodal features. In this section, we introduce our baseline models and the methodology of our experiments. Furthermore, we compare our model against various baseline approaches.

7.1 Baseline Models

We select two baseline models: (i) a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with a linear kernel, which is widely used as a discriminative model, (ii) a Maximum Entropy Model, which is an energy model similar to the HCRF but without the hidden states assumption. We used MaxEnt models instead of CRF models (Lafferty et al., 2001), as CRFs are designed to predict a sequence of labels while our task contains only one label for the entire interaction.

Support Vector Machine (SVM)

We use the implementation of SVM from the libsvm package (Fan et al., 2008). The parameter that controls the scale of the soft margin was obtained automatically using cross validation. We train two SVM models: one using the averaged features extracted over the entire interview(SVM Holistic), and the second using features from each adjacency pair stacked into a large feature vector (SVM AP).

[1] http://sourceforge.net/projects/hcrf/

Maximum Entropy Model (MaxEnt)

MaxEnt is implemented based on Ratnaparkhi (1996). We trained two models: MaxEnt Holistic, MaxEnt AP, following the same technique described for SVM.

7.2 Experiment Settings

All models in this paper were evaluated with the same cross validation and training-testing splits. We use a 4-fold testing and 3-fold validation with retraining. Validation of all model hyperparameters (regularization terms and number of hidden states) was performed automatically. For HCRF, we perform grid search over the regularization constant, 0, 1, 10, 100, 1000, and the number of hidden states, 2, 3, 4, 5. We found the best hyperparameter setting to be 1 for the regularizer and 4 for the number of hidden states. The reported model parameters are calculated using all available data, with 5-fold cross validation.

We compute precision by taking the number of correctly predicted depressed subjects divided by the total number of subjects that are predicted as depressed. Likewise, recall is computed as the number of correctly predicted depressed subjects divided by the actual number of depressed subjects. The F1 measure is the harmonic mean of the precision and recall in multimodal analysis (Stratou et al., 2013), which is a standard measure to capture the joint performance of precision and recall.

Z-score normalization is performed for each conversation to scale all the features into the same range, making the learned weights comparable. All multimodal features defined in Section 4 are concatenated into one feature vector per observation, in an early fusion fashion. The distribution of depressed and non-depressed subjects is skewed (30 depressed versus 100 non-depressed).

8 Results and Discussion

In this section, we present the results of our three experiments, looking at the effects of adjacency pairs, hidden commonalties and multiple osmolalities of the features. We further analyze the weights learned from our multimodal HCRF model to draw knowledge and implications from our interview corpus.

8.1 Effect of Using Adjacency Pairs

In order to show the benefits of modeling features at adjacency pair level, we compared the holistic approaches (SVM Holistic and MaxEnt Holistic)

Model	F1	Precision	Recall
HCRF	**0.664**	**0.767**	**0.585**
SVM Holistic	0.417	0.500	0.357
SVM AP	0.449	0.533	0.381
MaxEnt Holistic	0.523	0.567	0.486
MaxEnt AP	0.603	0.733	0.512

Table 1: Comparison of our approach with baseline models. 'Holistic' stands for models with features extracted over the whole interaction, 'AP' stands for models with features extracted at adjacency pair level.

with the adjacency pair approaches (SVM AP and MaxEnt AP) by performing pairwise T-tests on a 4 fold testing set. By F1 measure, the adjacency pair approaches are significantly better than holistic approach for both SVM and MaxEnt ($p < .05$ respectively). Detailed numbers are shown in Table 1. This shows that using features extracted at each adjacency pair level is better than extracting features over the whole interaction in the task of depression prediction as we have hypothesized in H1 of Section 3. Extracting features at the entire interview level ignores discriminative information within each adjacency pair as well as the dependence between consecutive pairs.

8.2 Effect of Learning Commonalities among Adjacency Pairs

Multimodal HCRF automatically learns the commonalities among different adjacency pairs by assigning them to the same hidden state. Each hidden states is a similar set of questions designed to serve similar purpose. We see from Table 1 that our approach outperforms all the baselines. Four paired T-tests are performed on the F1 measures, between the HCRF and each baseline model (SVM Holistic, SVM AP, MaxEnt Holistic and MaxEnt AP) on a 4-fold testing set and found statistical significance in all the four pairs with $p < .05$. These results suggest the advantage of learning commonalities among adjacency pairs, as we have hypothesized in H2 of Section 3.

8.3 Effect of Using Features Extracted from Three modalities

Figure 3 shows that the use of features from three modalities statistically outperforms (paired T-test with $p < .05$) all other possible combination of modalities using HCRFs in terms of the F1 measure, as we have hypothesized in H3 of Section 3. These results confirm the advantage of combining features from three modalities in the depression prediction task suggested in our third hypothesis. Yu et al. (2013) reported similar trends in friendship prediction.

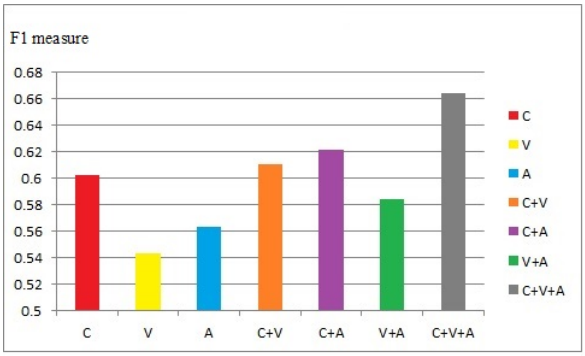

Figure 3: Comparison of our comprehensive multimodal approach against other set of features using HCRF, 'C' stands for conversational features, 'V' stands for visual features and 'A' stands for acoustic features, '+' stands for combination

8.4 Analysis of the Learned Multimodal HCRF

Figure 4 illustrates the learned Multimodal HCRF model with its optimized parameters. The learned model has four hidden states, which means that the adjacency pairs are clustered into four groups. By analyzing θ_y, we observe that depressed individuals are more tightly associated with the verbal and nonverbal behaviors manifested in the first and the last hidden states, while non-depressed individuals are more tightly associated with the second and third hidden states. We obtain the set of the most predictive features for each hidden state by selecting features with associated weights higher than 0.15. For example, in hidden state 1, "speech onset time", "neutral expression", "energy in dB" and "peak slope" stand out as the top ranked features. We show the top ranked features of each hidden state in Figure 4.

By performing inference on the learned model parameters, we can recover a list of the adjacency pairs most strongly associated with each hidden state for each participant. Then we hold a majority vote for each adjacency pair with all 130 participants to determine its most strongly associated hidden state overall. The first hidden state was most strongly associated with the responses to the questions *"How would your best friend describe you?"*, *"Tell me about the last time you felt really happy?"*, and *"I'm sure you can tell by my shoes. I'm not much of a world explorer. Do*

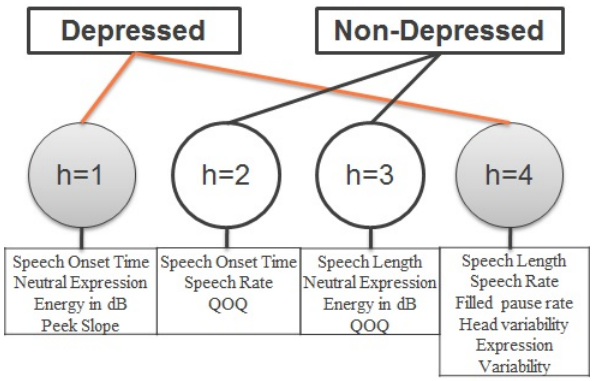

Figure 4: the Multimodal HCRF model for depression prediction. Hidden state 1 and 4 are more correlated with depressed people, while hidden state 2 and 3 have relatively larger influence on non-depressed people. We also listed features with weights higher than 0.15.

you travel a lot?". It is interesting to see that all of these questions are designed to build up intimacy between clinicians and patients. We found that "speech onset time" is negatively correlated with depression for all three adjacency pairs mentioned above. This is consistent with the findings in Cohn et al. (2009), where increased speaker-switch duration in conversation is found in the depressed group. However, there are other features that are only salient for one adjacency pair but not for the others. For instance, "peak slope" and "energy in dB" are only salient for the first question's response, but not for the others. The "peak slope" feature has been identified as a good indicator of depression, and as Scherer et al. (2013b) suggests, depressed patients tend to have tighter glottal flow than healthy individuals. Lower "energy in dB", meaning quieter speech, is correlated with depression. In addition to the above observations, we find that the "neutral expression" feature is not salient. This is despite the feature being the second most heavily weighted feature associated with the first hidden state. We believe that clustering adjacency pairs together through the hidden states provides more predictive power than using the features themselves. A previous study also found that "neutral expression" is a good indicator of depression through a holistic analysis (Stratou et al., 2013).

For the fourth hidden state of our model, the adjacency pairs with questions *"What are things you really like about LA?"*, *"How are you doing?"* , *"Where are you originally from?"*, and *"Sometimes when I'm feeling tense, I turn on the fish tank screen saver. Hey I know it's not Hawaii but it's the best I've got. What do you do to relax?"* appear to be the most relevant according to majority vote. All of these questions are from the rapport building phase of the interview. We found that for all four questions, depressed participants respond with shorter speech length. This finding is correlated with a previous report that depressed people are less expressive in the rapport-building phase of the conversation (Bylsma et al., 2008). In addition to shorter "speech length", lower "speech rate" is also a salient indicator of depression in response to the first three adjacency pairs we mentioned above, which correlates with findings of a previous study (Teasdale et al., 1980).

To sum up, our analysis suggests that clinicians should focus on different verbal and nonverbal behaviors in response to different questions. For example, "speech onset time" is very crucial for evaluating responses triggered by intimate questions, while "speech length" is very important for rapport building questions.

9 Conclusion

We introduced the Multimodal HCRF, a computational model which explicitly considers the context and the commonalities among the adjacency pairs in an interview. By combining conversational, visual and acoustic features, our model outperforms the use of any other combination of the modalities. The saliency of the verbal and nonverbal features extracted from the adjacency pairs is related to the content and purpose of the probing questions. For future work, we plan to incorporate linguistic cues, such as sentiment analysis, syntactic structure and lexical features into our computational model.

10 Acknowledgement

We thank the Richard King Mellon Foundation for their generous funding. This work is also partially supported by DARPA under contract (W911NF-04-D-0005) and U.S. Army Research, Development, and Engineering Command. The content of the paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We also thank Alan W Black, Carolyn Rose, Yanchuan Sim, Shoou-I Yu, Elijah Mayfield and Brian Coltin for their insightful suggestions and Ramon Paz, Rob Groome and Ben Farris for their technical support.

References

P. Alku, T. Bäckström, and E. Vilkman. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

T. Baltrusaitis, P. Robinson, and L. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *CVPR, 2012 IEEE Conference*, 2012. doi: 10.1109/CVPR.2012.6247980.

L. Bylsma, B. Morris, and J. Rottenberg. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical psychology review*, 28 (4):676–691, 2008.

J. F. Cohn, T. S. Kruez, I. Matthews, Y. Ying, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, 2009.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

D. DeVault, K. Georgilia, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *to appear in Proceedings of SigDial 2013*, 2013.

T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech 2011*, pages 1973–1976. ISCA, 2011.

H. Ellgring. *Nonverbal communication in depression*. Cambridge University Press, Cambridge, 1989.

L. A. Fairbanks, M. T. McGuire, and C. J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology*, 91(2):109–119, 1982.

R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

H. P. Grice. Logic and conversation. *1975*, pages 41–58, 1975.

J. A. Hall, J. A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician-patient interaction. *Applied and Preventive Psychology*, 4(1):21–37, 1995.

N. Henrich, C. d'Alessandro, and B. Doval. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. *Proceedings of EUROSPEECH, Scandanavia*, pages 47–50, 2001.

F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23:67–72, 1975.

J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech 2011*, pages 177–180. ISCA, 2011.

J. Kane, S. Scherer, L.-P. Morency, and C. Gobl. A comparative study of glottal open quotient estimation techniques. In *to appear in Proceedings of Interspeech 2013*. ISCA, 2013.

A. Kirsch and S. Brunnhuber. Facial expression and experience of emotions in psychodynamic interviews with patients with ptsd in comparison to healthy subjects. *Psychopathology*, 40 (5):296–302, 2007.

K. Kroenke and R. L. Spitzer. The phq-9: A new depression and diagnostic severity measure. *Psychiatric Annals*, 32:509–521, 2002.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL `http://dl.acm.org/citation.cfm?id=645530.655813`.

G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *FG 2011, 2011 IEEE International Conference on*, 2011. doi: 10.1109/FG.2011.5771414.

R. Menke. *Examining nonverbal shame markers among post-pregnancy women with maltreatment histories*. PhD thesis, Wayne State University, 2011.

L. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *FG '08. 8th IEEE International*

Conference on, 2008. doi: 10.1109/AFGR.2008.4813429.

J. E. Perez and R. E. Riggio. *Nonverbal social skills and psychopathology*, pages 17–44. Nonverbal behavior in clinical settings. Oxford University Press, 2003.

A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Advances in neural information processing systems*, pages 1097–1104, 2004.

A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(10):1848–1852, 2007.

A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142, 1996.

E. A. Schegloff. *Sequence organization in interaction: Volume 1: A primer in conversation analysis*, volume 1. Cambridge University Press, 2007.

J. T. M. Schelde. Major depression: Behavioral markers of depression and recovery. *The Journal of Nervous and Mental Disease*, 186(3): 133–140, 1998.

S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Proceedings of Interspeech 2013*. ISCA, 2013a.

S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*. IEEE, 2013b.

G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *to appear in Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013.

D. Talkin. A Robust Algorithm for Pitch Tracking. In W. B. Kleijn and K. K. Paliwal, editors, *Speech coding and synthesis*, pages 495–517. Elsevier, 1995.

J. D. Teasdale, S. J. Fogarty, and J. M. G. Williams. Speech rate as a measure of short-term variation in depression. *British Journal of Social and Clinical Psychology*, 19(3):271–278, 1980.

Z. Yu, D. Gerritsen, A. Ogan, A. W. Black, and J. Cassell. Automatic prediction of friendship via multi-model dyadic features. 2013.