DATA DRIVEN MODELING OF HEAD MOTION TOWARDS ANALYSIS OF BEHAVIORS IN COUPLE INTERACTIONS

Bo Xiao†*, Panayiotis G. Georgiou*†*, Brian Baucom*‡*, Shrikanth S. Narayanan*†

† SAIL, Dept. Electrical Engineering, University of Southern California, Los Angeles, CA 90089
‡ Dept. Psychology, University of Utah, Salt Lake City, UT 84112, U.S.A.
boxiao@usc.edu, georgiou@sipi.usc.edu, brian.baucom@psych.utah.edu, shri@sipi.usc.edu

ABSTRACT

We propose a data driven approach for modeling head motion behavior in human dyadic interactions, by establishing a structure for unconstrained natural head movement. Using recordings of couples' conversations in real psychotherapy sessions, we first track the head of each subject, compute the head motion and detect active versus non-active intervals. For detected active intervals, we use a sliding window to collect motion sequences. Linear Prediction Coefficients are used to represent the sequence, based on which we train a Gaussian Mixture Model (GMM) such that each mixture would ideally associate with one type of prototypical movement, which we will refer to as a "kineme". For each complete interaction session, we compute the sum of posterior probabilities of all sequences over the GMM normalized by session length to predict specific "low" versus "high" expert annotated behavior code scores for *Acceptance*, *Blame*, *Positive* and *Negative* behaviors. We achieved an overall accuracy of about 70% employing these GMMs. This result shows data driven modeling of head motion provides useful information for human behavioral analysis.

Index Terms— Head motion; Kineme; Gaussian Mixture Model; Linear Prediction; Behavioral analysis

1. INTRODUCTION

Traditional methods of studying human behavior in psychology rely on manual annotation of recorded interactions, which is often costly and hard to implement on a large scale. Recently, several computational approaches have been proposed to automate the annotation process based on "Behavioral Signal Processing", *i.e.,* signal processing and machine learning techniques applied to multimodal observations, motivated by domain knowledge (*e.g.,* in psychology) and real applications [1]. For example, Rozgic *et al*. estimated the "Approach-Avoidance" behavior scores — one type of moment-by-moment behavior code manifesting the immediacy and involvement of two interlocutors — using motion-capture and acoustic cues [2]. Black *et al*. studied the problem of classifying "high" versus "low" presence of certain behaviors such as *Acceptance*, *Blame*, *etc*. as judged by human experts in couples' psychotherapy sessions, using a variety of vocal acoustic features [3]. Georgiou *et al*. addressed the same problem by utilizing lexical features obtained through an Automatic Speech Recognition system [4]. Other novel insights about behavioral dynamics can be obtained by computationally modeling and tracking subtle, often difficult to directly observe, interaction phenomena. For example, Lee *et al*. have proposed a signal-derived vocal entrainment measure and applied it to the analysis of distressed couples' interactions [5], which had only been studied qualitatively before.

This work is supported by NSF.

In addition to the verbal information conveyed through acoustic and lexical modalities, nonverbal visual behavior is an integral part of human interaction. Nonverbal communication could involve several modalities including body motion, limbic gestures, facial expression, social spacing (proxemics), and even clothing and hair that are static during the interaction [6]. Although there are many common forms of gestures and movements used by the majority of population across cultures, motion behaviors are quite variable, and not fully understood, in real life. Psychologists have been developing coding schemes to systematically model such behavior [7]. Ekman and Friesen proposed one of the most well known coding systems for this, which categorizes nonverbal behavior into five classes: emblems, illustrators, regulators, adaptors and affect displays [8]. This system focuses on the function and usage instead of a computational characterization of behaviors. Birdwhistell, one of the pioneers in establishing a structure of motion behavior, proposed the terminology of "Kinesics", as well as the theory of "kinesics-phonetics analogy" [9]. Here each elementary unit of motion is described as a "kineme", similar to a "phoneme"; and a series of "kinemes" are called "kinemorphs", analogous to morphemes. Like phonemes, kinemes are abstract and predefined (*e.g.,* right hand lift, head turn left, *etc*.), though every physical expression of a particular kineme will differ, and not be realized in the exact same way. The nonverbal language is then composed by a sequence of kinemes.

However, the kinesics perspective as a structural view of nonverbal behavior was not fully developed, partly due to the limitation of analysis capability in the time of the inventors, and partly because non-verbal language is much more unstructured and hence the kineme space can not be very well and discretely defined. In addition, it is difficult to manually categorize and label all kinds of motions in real settings, and even more difficult to reach wide agreement among practitioners in creating an acceptable inventory [7]. As Kendon commented in 1996, Birdwhistell was probably ahead of his time. Yet he suggested that with the development of signal processing and computer vision techniques, it would be much easier to examine the kinesics theory in practice [10]. Now that such computational tools are becoming increasingly robust and reliable, we would like to revisit kinesics theory in a data-driven way, by establishing the structure of motion (kinesis) classes using automatic clustering, and hence also deal with the quantization of the continuum of non-verbal language.

The most active and studied body parts are typically the hand and the head [7]. In this paper, we focus on head motion in dyadic interaction scenarios, specifically, in an audio-visual recording database of distressed couples psychotherapy sessions. Note that the popularity of head motion study is not only due to the high frequency of movement, but is also motivated by the rich linguistic and semantic meaning conveyed by head motion. McClave studied the linguistic functions of head movement in different contexts [11].

Hadar *et al*. analyzed head movement particularly in listening turns with an approach characterizing the physical movement of head [12]. Heylen addressed head motion pattern and function in social interaction contexts as a "joint activity" of the interlocutors [13].

Among various patterns of head movement, the most studied are head nods and shakes. In the engineering community there have been many studies on recognition of head motion, usually distinguishing nods versus shakes. For example, Bousmalis *et al*. did a survey on head motion and facial expression analysis towards discerning the "agreement or disagreement" attitude of users [14]. Nevertheless, there is no complete and well agreed categorization of head motion and there is limited effort from the engineering perspective. In [7] the authors have defined *six* head motion classes: Nod, Shake, Tilt, Dip, Toss and Miscellaneous. For engineering modeling, one could design classifiers for each prototypical type, yet creating training sets through labeling would be tedious and inaccurate especially due to segmentation and human labeler disagreements.

Given the need for establishing a structural model of head motion, notably in the presence of the difficulty of direct reference to motion types, we propose a data driven approach. The approach relies on using sliding windows of movement along short durations for automatic feature extraction and clustering. By clustering in a probabilistic sense, *e.g.*, using a *Gaussian Mixture Model* (GMM), we implicitly attempt to model various types of motion. Ideally one cluster could correspond to one kineme, or could be interpreted more generally as being possibly generated by different kinemes weighted by the posterior probability. We focus on modeling the head motion of an individual interlocutor. To validate this method we examine the prediction of "high" versus "low" behavior code values provided by human experts. Specifically we consider automatically categorizing *Acceptance*, *Blame*, *Positive* and *Negative* behavior codes in a binary classification task.

2. DATASET

The corpus used in this work comprises audio-visual recordings of seriously and chronically distressed couples having dyadic conversations on solving a problem in their marriage, collected by the University of California, Los Angeles and the University of Washington [15]. Each couple talked about two separate problems one chosen by the wife and one by the husband, for 10 minutes each. These discussions took place at three points in time during the therapy process: before the psycho-therapy began, 26 weeks into the therapy and 2 years after the therapy session finished. The database is 96 hours long and contains 574 sessions. The video format is 704×480 pixels, 30 fps, with a screen split and one spouse on each side.

Both spouses in all sessions were evaluated individually following two expert designed coding systems, the Couples Interaction Rating System 2 (CIRS2) [16] and the Social Support Interaction Rating System (SSIRS) [17]. The CIRS2 contains 13 behavioral codes and was specifically designed for conversations involving a problem in relationship, while the SSIRS consists of 20 codes that measure the emotional component of the interaction and the topic of conversation. The 33 codes are each on a numerical range from 1 to 9. At least three trained human coders were assigned to the same session, where they would watch the entire session and give an overall score on each code. In this paper we select four behavioral codes for experiments (*Acceptance* and *Blame* from CIRS2, *Positive* and *Negative* from SSIRS), which have above 0.7 correlation among coders, pointing to high inter-coder agreement. We use the average score among coders as ground truth. Note that the codes only measure how much particular aspects occur, independent of how much their opposite occur. For example, both *Positive* and *Negative* codes could have high value if they are both present in the interaction.

The video quality of the recording (done in distributed clinical settings) is not ideal, and relative positions of subjects as well as of the cameras are not available as the database was intended originally for human analysis. Therefore, we apply a preprocessing step to all sessions on the left and right split screen content of the video. First, we run an OpenCV [18] face detector on one frame per second of the video. Second, the face scale is estimated by the mode of the distribution of detected size of the face block. Third, we retain sessions that have a face detected on more than 70% of the sampled frames, and the estimated face scale is between 120 pixels to 160 pixels ($\frac{1}{4}$ to $\frac{1}{3}$ of image height). As a result, 249 sessions-by-subjects were adopted (sometimes only one side of a recording). In these recordings, the upper body of a subject is present while it is uncertain if the hands are captured (sample frame in top of Fig. 1).

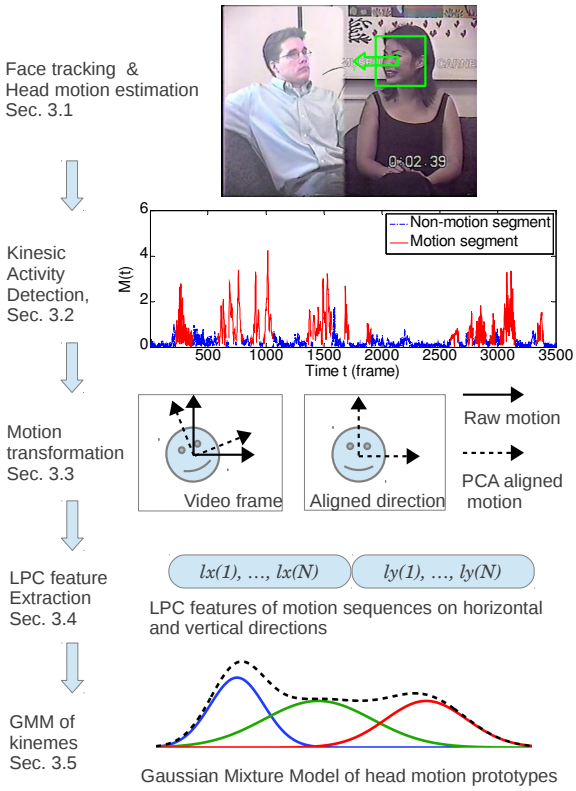

Fig. 1: Illustration of the processing steps in Sec. 3

3. MOTION MODELING AND FEATURE EXTRACTION

3.1. Motion estimation

Face tracking and head motion estimation is a necessary front-end for later steps. As this module is not the focus of our work, we utilize a simple but effective setup. We first detect the face (marked by a square) in each frame using the cascade classifiers provided in OpenCV, and approximate the face size with the side length of the detected face square. Using a 5 frame sliding window on the histogram of face size, we choose the size $\hat{S}$ that maximizes the sum of the windowed histogram. In other words, we choose the most likely face size on a smoothed histogram. We exclude outliers of face detection by rejecting faces with size $S > 1.2\hat{S}$ or $S < 0.8\hat{S}$. The central location of face is estimated by the center of accepted face squares. We again exclude faces with centers that are further than $\hat{S}$ on the horizontal axis or $0.5\hat{S}$ on the vertical axis to the estimated central location. We fill the gaps of missing-face frames by linear interpolation.

Head motion on horizontal and vertical directions are derived as the mean of horizontal and vertical components of the optical flows

over all pixels within the face square, respectively. Given that the spouses remain in a sitting position throughout the session, this simple setup satisfies our need and produces reliable results.

3.2. Kinesis activity detection

Similar to Voice Activity Detection in speech, we set up a Kinesis Activity Detection (KAD) step based on the motion estimates to remove the periods that the interlocutor does not move. Let the head motion stream be $M_x(t)$ and $M_y(t)$. We use the magnitude of motion $M(t) = \sqrt{M_x^2(t) + M_y^2(t)}$ as a 1-D feature. We use a 2-Mixture GMM to represent motion versus non-motion classes, and a 2-state Hidden Markov Model (HMM) to represent the transition between the two classes. The parameters of the GMM are initialized by selecting the top 20% high valued $M(t)$ as in the motion class while the rest are in the non-motion class, and the initial transition probability of HMM is set to 0.9 for self-transition. The Expectation-Maximization algorithm is applied for 30 iterations to obtain the Maximum Likelihood Estimation of the states. Moreover, we post-process the state sequence by smoothing over short pauses (less than 0.2 seconds) when both sides of the pause are motion sequences longer than 1 second. Then we eliminate motion sequences that are less than 1 second, which are assumed to be noise. An example of KAD result is shown in Fig. 1.

3.3. Motion transformation and windowing

Note that the spouses were sitting in arbitrary postures, so the main directions of their head movements are not necessarily 0 or 90 degrees. We apply a Principal Component Analysis (PCA) to the raw motion stream so as to align the main directions. On a 2-D plane of head motion, the two main orthogonal directions are associated with the horizontal and vertical dimensions, *e.g.,* as illustrated in Fig. 1. We also do a Z-normalization to the two dimensions of aligned motion streams (empirically found to have a distribution in bell shape with heavier tails). The motion segments may have varying durations after the KAD step, where one segment might contain a group of consecutive but different kinemes, or "kine-morphs". These heterogeneous types of motions should be analyzed separately, so further segmentation within a motion segment is needed. Since an "ideal" segmentation scheme is not directly available because of the inherent ambiguity of head motion, we apply a short time sliding window over each motion segment, with window length being 2 seconds, and window shift being 1 second. If the motion segment is less than 3 seconds then we do not further window it. Therefore, the windowed motion sequences could have length between 1 to 3 seconds.

3.4. Linear prediction features

We use Linear Prediction Coefficients (LPC) as a transformed representation of the motion sequences for several reasons. First, assuming that head motion sequences can be viewed as being generated by an auto-regressive process, then LPC would capture the dynamic properties of various motion types. Second, LPC is preferred instead of other methods such as Vector Quantization, because the motion sequences obtained through windowing are not exactly aligned with each kineme. Third, LPC provides the convenience of consistent feature dimension, while the windowed motion sequences are in varying lengths. Finally, derived forms of LPC such as Line Spectral Frequencies are often used in speech coding for better quantization property; however, in this application we found the original LPC features offer higher accuracy.

We compute the LPC for horizontal and vertical components respectively, then concatenate the two components. Therefore, let $L_j^i = [Lx_j^i \; Ly_j^i]$ be the i-th motion sequence in the j-th session, and $Lx_j^i = \{lx_j^i(n)\}_{n=1}^N$ be the horizontal component, $Ly_j^i = \{ly_j^i(n)\}_{n=1}^N$ be the vertical component, where N is the order of LPC analysis. The constants $lx_j^i(0) \equiv 1$ and $ly_j^i(0) \equiv 1$ are omitted. As a result, we have a $2N$-dim feature for each motion sequence.

3.5. Gaussian mixture of kinemes

Recall that we introduced the idea of automatically clustering motion sequences. Here we construct a GMM with LPC features. We use the posterior probability of each feature instance as a soft cluster label in associating a kineme, in order to accommodate the ambiguity of motion types that exist in practice. To train the GMM we pool all sessions together and conduct the training on all motion sequences. The K-mixture GMM is initialized by a K-means procedure, and iteratively optimized using standard EM algorithm.

However, since GMM approach is unsupervised with the initialization being random, a single GMM does not guarantee good performance for a discriminative problem (the behavior code classification). This issue will be discussed further in the experiments section and discussion.

Let π_k be the prior probability, $\mu_k = \{\mu_k(n)\}_{n=1}^{2N}$ be the mean vectors, and $\sigma_k = \{\sigma_k(n)\}_{n=1}^{2N}$ be the variance vector, *i.e.,* the diagonal of the assumed diagonal covariance matrix corresponding to the feature vector of dimension $2N$. The likelihood probability is given by $P(L_j^i|k) = \mathcal{N}(L_j^i; \mu_k, \sigma_k)$, and the posterior probability is:

$$P(k|L_j^i) = \frac{\pi_k P(L_j^i|k)}{\sum_{k'=1}^K \pi_{k'} P(L_j^i|k')}, k = 1 \cdots K. \quad (1)$$

We use the sum of posterior probability vectors from all sequences in each session, normalized by session duration, as the final feature for the binary behavior code classification experiments. Let the final feature be $F_j = \{F_j(k)\}_{k=1}^K$, the posterior of mixture k be $P(k|L_j^i)$, the session duration be T_j. Then

$$F_j(k) = \frac{1}{T_j} \sum_i P(k|L_j^i), k = 1 \cdots K \quad (2)$$

For unseen sessions, we extract the LPC features in the same way, and compute the posterior probabilities based on the trained GMM.

4. EXPERIMENTS

We check the effect of linear predictive analysis by comparing the original signal variance $\mathtt{V}_S$ to the residual variance $\mathtt{V}_E$, which approximates the signal-to-noise ratio. Let $\mathtt{VR} = 10\mathtt{log}_{10}\frac{\mathtt{V}_S}{\mathtt{V}_E}$. We found that VR increases with LPC order. When $N = 10$, VR for both horizontal and vertical motion directions are around 10dB. In experiments we found that in general any higher N does not improve the classification accuracy, yet consumes more computation time. In the following, we select $N = 10$.

To test the kinesis model as GMMs, we setup a binary classification problem, where we select the top 25% and bottom 25% scored sessions for each of the four behavioral codes, respectively. We test the number of mixtures K from 3 to 25, where 3 is chosen as a very small number of mixtures, and 25 is much larger than the number of head motion types in most existing coding systems.

Our experiments are with a leave-one-subject-out cross validation. Ideally one would optimize parameters on a development set, but as the first experiment we decided to avoid that due to the small amount of data (in terms of session level scores) and the computational cost. Given an unsupervised GMM training it is likely that we will end up with a partition which is uninformative for the task at hand. Therefore we followed two different approaches:

First, we repeatedly train 50 different GMMs on the training data and choose our best classifier as the one that performs best on the training data itself (Acc_1). This in a sense is using the training data as the development data, but instead of full-blown parameter optimization the "best of 50" approach is taken for computational cost

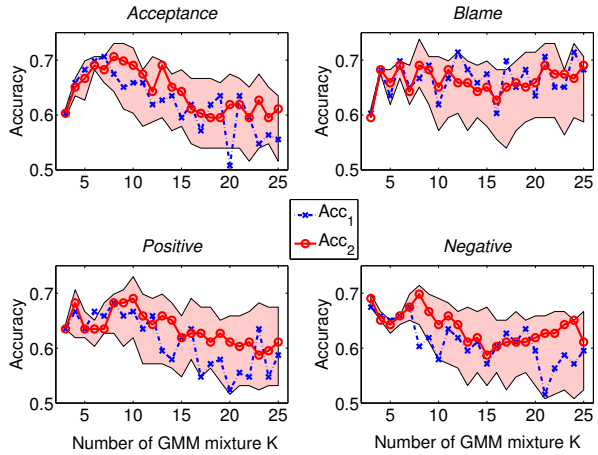

Fig. 2: Classification results on behavior codes

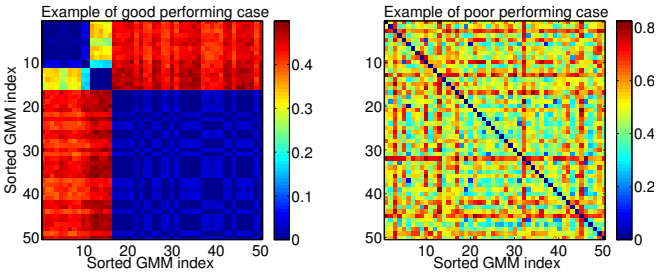

Fig. 3: Divergence of GMMs in the ensemble

reduction. Note that optimization would take place on the clustering initialization parameters and as we will show later this has little impact in some cases, as it should when clustering cost is smooth with single global minimum.

The second approach is to predict using all the 50 GMMs and employ majority-voting as the decision (Acc_2). The assumption here is that the clustering that converges at other local minima is detecting events independent to the behavioral codes of interest, hence averaging would cancel those out. The number 50 is chosen empirically as a trade-off between having an adequate number to increase robustness and assuring affordability in terms of computation complexity.

In both of the above cases for the training stage and for every cross validation round, behavior code and K value, we have an ensemble of 50 GMMs. For each behavior code there are examples from about 80 distinct subjects in our corpus, yielding about 1.2×10^4 instances of motion sequence in total. We use F_j as feature for classification, and linear SVM [19] as the binary classifier when a particular GMM is considered.

In Fig. 2 we show the results of behavior codes classification. The pink areas in the background illustrate the range of test accuracy using the m-th GMM in each cross validation round, with m ranging from 1 to 50. The range tends to increase when K is large. We see that for all codes the highest Acc_2 is about 0.7, while for most cases Acc_2 is above 0.6 which is significantly higher than chance level ($p < 0.01$ in binomial test). The first approach is in general not as robust as the second one, likely due to the lack of a distinct development set. We leave the problem of selecting an optimal K for a particular code for future work. Alternatively, a non-parametric Bayesian approach may be considered where the prior on K is obtained from domain knowledge or previous experimental result.

5. DISCUSSION

5.1. Comparing the GMMs of two representative cases

In Fig. 3 we compare two representative cases, $K = 7$ and $K = 20$ for code *Acceptance* with the first subject left-out. Note that $\text{Acc}_1(K = 7) = 0.71$, and $\text{Acc}_1(K = 20) = 0.51$ (see the upper-left plot of Fig. 2). We compute pair-wise approximated KL-divergence between GMMs using the Monte Carlo method introduced in [20]. In both plots the GMMs are sorted by their accuracies on the training set. So the upper-left corner corresponds to the divergences among higher accuracy GMMs, and the lower-right corner corresponds to that of the lower accuracy ones.

The figures suggest that the data can be generalized in many different ways, *i.e.,* there are many local minima in the EM procedure. The left figure suggests that the best 9 classifiers model data in a way that is most meaningful for the behaviors of interest (for instance the mixtures posterior are correlated with the behavior occurrences). It also shows that all GMMs are well defined. This might imply a relatively smooth likelihood function in EM with local minima close to the global minimum so that even without optimization the three resulting partition types (each chosen 9, 5, 36 times respectively) are reasonable and close together (greatest KL div.$<$0.5).

On the other hand, the right figure suggests that the likelihood function is very noisy with a large number of local minima and that the random initialization and lack of optimization lead to a lack of convergence towards an appropriate clustering. The KL divergence is much higher which denotes far less convergence to a global optimum. This may be due to a wrong choice of the number of clusters (although in reality human movement is unstructured and continuous, and definition of the correct number of clusters is ill conditioned), or due to the larger search space in this higher dimensional space. Therefore, in this case the selection of best performing GMM is not reliable, and it might partly explain why the majority voted Acc_2 outperforms Acc_1.

Addressing the above issue through detailed error analysis and optimization will be a thrust of our future work.

5.2. Comparison to previous work

A summary of related work was provided in Sec. 1. Compared to earlier results on the same line of behavior code classification problems, the use of this specific visual modality (head motion) achieves an accuracy close to that with the acoustic modality [3], and the lexical modality with automatically extracted transcripts, but less than the lexical modality with accurate transcriptions [4]. This is intuitive since lexical information is more directly in relation to the human judgment of behaviors under question which serves as the baseline, while expressions in audio-visual channels are more implicit and predictions are directly signal driven.

6. CONCLUSION

In this paper we proposed a data driven approach to construct a GMM for head motion, so as to establish a structure of head movement behavior in dyadic human interactions. LPC features were adopted to represent the motion sequence, and the sum of posterior probabilities over the GMM normalized by session length was used as the session level feature in a binary classification task aimed at predicting expert specified behavior codes. Experiment results show that the proposed model is able to predict extremes of behavior codes with an overall accuracy around 70%. The results also appear consistent with previous (theoretical) proposals of *six* classes of head motion [7], since K being 5$\sim$10 yields better performance. These suggest that data driven modeling of head motion can provide useful information for human behavioral analysis.

In the future we plan to investigate error analysis, feature design for better interpretability, optimization based on development set, multimodal fusion using salient modality selection, and dynamic modeling of the interaction.

7. REFERENCES

[1] S. Narayanan and P. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceeding of IEEE*, 2012.

[2] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. Georgiou, and S. Narayanan, "Estimation of ordinal approach-avoidance labels in dyadic interactions: ordinal logistic regression approach," in *Proc. ICASSP*, 2011, pp. 2368–2371.

[3] M.P. Black, A. Katsamanis, Baucom B.R., C.C Lee, A.C. Lammert, A. Christensen, P.G. Georgiou, and S.S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, 2011.

[4] P. Georgiou, M. Black, A. Lammert, B. Baucom, and S. Narayanan, ""that's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?," in *Proc. ACII*, 2011, pp. 87–96.

[5] C.C. Lee, A. Katsamanis, M.P. Black, B.R. Baucom, A. Christensen, P.G. Georgiou, and S.S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, 2012.

[6] M.L. Knapp and J.A. Hall, *Nonverbal Communication in Human Interaction*, Wadsworth, Cengage Learning, Boston, 7 edition, 2007.

[7] J.A. Harrigan, R. Rosenthal, and K.R. Scherer, *The new handbook of Methods in Nonverbal Behavior Research*, pp. 137–198, Oxford University Press, New York, 2005.

[8] P. Ekman and W.V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Nonverbal communication, interaction, and gesture*, pp. 57–106, 1981.

[9] R.L. Birdwhistell, *Kinesics and context: essays on body motion communication*, vol. 2, University of Pennsylvania Press, 1970.

[10] A. Kendon and S.J. Sigman, "Commemorative essay. Ray L. Birdwhistell (1918-1994)," *Semiotica*, vol. 112, no. 3-4, pp. 231–262, 1996.

[11] E.Z. McClave, "Linguistic functions of head movements in the context of speech," *Journal of Pragmatics*, vol. 32, no. 7, pp. 855–878, 2000.

[12] U. Hadar, TJ Steiner, and F. Clifford Rose, "Head movement during listening turns in conversation," *Journal of Nonverbal Behavior*, vol. 9, no. 4, pp. 214–228, 1985.

[13] D. Heylen, "Challenges Ahead: Head movements and other social acts in conversations," *Virtual Social Agents*, pp. 45–52, 2005.

[14] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *Proc. ACII*. IEEE, 2009, pp. 1–9.

[15] A. Christensen, D.C. Atkins, S. Berns, J. Wheeler, D.H. Baucom, and L.E. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *Journal of consulting and clinical psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[16] C. Heavey, D. Gill, and A. Christensen, "Couples interaction rating system 2 (CIRS2)," *University of California, Los Angeles*, 2002.

[17] J. Jones and A. Christensen, "Couples interaction study: Social support interaction rating system," *University of California, Los Angeles*, 1998.

[18] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[19] C.C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[20] J.R. Hershey and P.A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. ICASSP*. IEEE, 2007, vol. 4, pp. 317–320.