POWER-SPECTRAL ANALYSIS OF HEAD MOTION SIGNAL FOR BEHAVIORAL MODELING IN HUMAN INTERACTION

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 examine whether head motion can be used for predicting human expert's judgments of behavioral characteristics relevant to the couples therapy domain. Specifically we predict "high" or "low" presence of several behavioral characteristics such as "Blame" that are discerned by human experts, through data-driven clustering of the head motion signal based on power-spectral features. We employ the distribution of motion samples in each cluster for behavior judgment prediction. We find clustering horizontal and vertical motion separately is superior to combined clustering in predicting behavior. The performance of gender-specific and gender-independent clustering of head motion is comparable in average while different for each gender. The proposed power-spectral features outperform linear prediction features in average. Using data from a clinical study of distressed couples, we empirically show that the derived clusters quantize head motion into meaningful types that relate to interpretable behavior characteristics. These findings demonstrate the feasibility of inferring behavior characteristics from head motion signals.

***Index Terms*—** Head motion; Behavioral characteristic; Power spectrum; Clustering; Human interaction

1. INTRODUCTION

Behavioral Signal Processing aims to provide a computational framework for human behavior analysis and modeling, so as to inform human assessment and facilitate decision making [1]. Exemplary applications range from user modeling in commerce to applications centered on human health and well being such as couples therapy, addiction intervention, and children's autism diagnosis/treatment. An integral aspect is to computationally model the human behavioral judgment process, *i.e.,* how a domain expert derives a behavioral characterization based on observations.

There are several possibilities for approaching this problem. Learning direct mapping from multimodal signal features (*e.g.,* pitch, energy of speech; velocity of hand and head motion) to expert judgments (*e.g., Blame*, *Approach-Avoidance*) using machine learning is straightforward [2, 3], and offers automated detection of specific well-understood behavioral constructs. However, it suffers from limitations of generalizing across domains, or learning new behavioral variants, and in interpreting the mediating processes.

An alternative approach is to first learn intermediate representations of signal cues encoding meaningful patterns of behavior, before mapping them onto domain-level descriptions. An attractive aspect is computational implementation of constructs from behavioral science and theory. For example, the notion of behavioral entrainment (or synchrony) [4] is implicated in positive/negative affective behavior in human interaction; one corresponding computational ancillary is through vocal similarity measures, that can be extracted by comparing interlocutors' speech signals in order to approximate behavioral entrainment along time. Such measures have been in turn shown to predict expert judgments of couples' affective behaviors in conversations [5] as well as therapist's empathy during addiction counseling [6].

In this work, our aim is to find patterns of head motion that relate to meaningful behavioral characteristics. Head motion has rich semantic, discourse, and communicative functions in conversation [7] as a "joint activity" of the interlocutors [8]. In terms of constituent patterns, it has been qualitatively and functionally categorized by psychologists as nodding, shaking, tilting, tossing, dipping, thrusting, dropping, *etc.*; however, it is difficult to find a complete or well-accepted inventory of head motion [9]. In addition, Hadar *et al.* has suggested five classes of head motion based on kinematics: slow, ordinary, rapid, posture shift, and tremor [10], that quantize head motion with respect to magnitude and frequency.

While head motion analysis and modeling has been researched in the engineering domain, the problem of automatically finding the classes of head motion in natural spontaneous interactions is still open. Significant effort has been made in estimating head pose; Murphy-Chutorian and Trivedi provided a survey article of those [11]. However, in addition to head pose we need to analyze the motion type. There are many methods to discern nodding *vs.* shaking, which are widely used as agreement/disagreement cues for human-computer interaction systems [12]. But in natural human interactions, many more types of motion exist.

The psychologist Birdwhistell has suggested the theory of "kinesic-phonetic" analogy, where the unit of motion is called "kineme", similar to phoneme (elementary unit of speech, such as vowels and consonants) [13]. By defining a class of kinemes, one can compose any motion with a sequence of kinemes, just like putting phonemes to words and words to language. However, the problem remains as the definition of such a class of kinemes.

In our previous work [14], inspired by the kinesic analysis, we clustered head motion in a data-driven way. We segmented motion *vs.* non-motion states, extracted linear prediction features over short-time windows in motion segments, and trained Gaussian mixture models (GMMs) based on these features. We expected these features to capture the motion property, assuming head motion can be generated by an auto-regressive process in short duration where stationarity was assumed to hold. GMMs were treated as a soft way of clustering motion types, where each component may correspond to one kineme. We used the accumulated posteriors of motion events in each interaction session as cues for binary classification of "high" or "low" scores of four behavior codes (*Acceptance*, *Blame*, *Positive*, *Negative*) of interacting couples provided by expert psychologists. The automatic classification accuracies were in the range of 60% to

This work is supported by NIH, NSF and DoD.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

70% under various settings.

In this work, we expand on the data-driven analysis of head motion, addressing several problems that were not addressed in [14]. First, we propose power-spectral features besides linear prediction features used in [14] to capture motion properties. Second, we compare the performance of power-spectral *vs.* linear prediction features, gender specific *vs.* generic modeling, and separate modeling of horizontal and vertical motion directions *vs.* combined modeling. Third, we analyze the resultant clusters of motion signal in a case study, and try to find the patterns of power-spectra that indicate specific behavioral characteristics.

Additionally we also make a few methodological improvements compared to [14] that increase the robustness of the algorithm. We skip the motion/non-motion segmentation step, and expect the clustering to detect non-motion periods. We also skip the alignment of head motion direction with X-Y axes, because we observe that although the interlocutors take a steady sitting posture, the head motion direction varies continuously. Therefore assuming a global alignment direction may be disadvantageous. In addition, we avoid the GMM training and use simple K-means clustering, as we find experimentally no improvement of accuracies. The overview of the system is shown in Fig. 1.

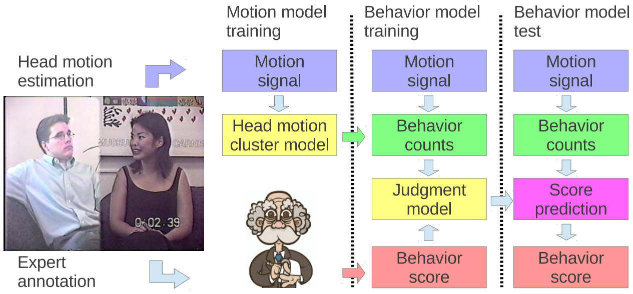


Fig. 1: Overview of the system design

2. DATASET

Psychology researchers from University of California, Los Angeles and University of Washington conducted a longitudinal psychotherapy research on couples [15]. They recorded conversations of chronically distressed couples solving problems in their marriage, where the wife and the husband chose a topic to discuss in turn, for 10 min each. The recordings were in audio-visual format, collected at three time stages during the study: before the therapy, 26 weeks into the therapy, and 2 years afterwards. The entire database contains 96 hours of recording in 574 sessions. The video format is 704 × 480 pixels, 30 frames-per-second, with a screen split and one spouse on each side taking a sitting posture (see Fig. 1).

For each session, at least three psychologists viewed the recording, and scored several aspects of each individual interlocutor's behavior on a numerical scale from 1 to 9. These are defined in two coding manuals, namely the Couples Interaction Rating System 2 (CIRS2) [16] and the Social Support Interaction Rating System (SSIRS) [17]. The CIRS2 includes 13 behavioral codes specifically designed for conversations involving a problem in relationship. The SSIRS includes 20 codes measuring the emotional component of the interaction and the topic of conversation. In this paper we select the same group of codes for experiments as in [14]: *Acceptance* and *Blame* from CIRS2, *Positive* and *Negative* from SSIRS. These codes have above 0.7 correlation among coders, indicating a high intercoder agreement. We use the average score among coders as ground truth. These codes are evaluated independently on their presence although highly correlated.

The video quality of the recording (done in various clinical locations/settings) is not ideal, and the relative positions of subjects as well as of the cameras are not available, as the database was originally intended for human analysis. We start from the same set of sessions as utilized in [3], and extract head motion signal as described in Sec. 3.1. We manually check the head tracking results on all sessions by playing the video 10 times faster. We find 221 sessions (out of 372, about 37 hours) having good tracking results for both of the spouses, which we use for experiments in this paper.

3. HEAD MOTION SIGNAL ANALYSIS

3.1. Head tracking and motion estimation

The front-end module tracks "corner-like" points in the video, then prunes it according to detected face regions, described as follows. We use OpenCV [18] implementations of the supporting algorithms.

```
Input: Video recording of one interlocutor, empty set of tracks $\mathcal{T}$
       and empty sequence of faces $\mathcal{F}$.
  1. Find pixel set $\mathbf{P}$, where $p \in \mathbf{P}$ is selected by the "good-
     features-to-track" technique [19]. Create a track $t \in \mathcal{T}$ of
     length 1 for each $p$.
  2. Estimate the corresponding pixel $p'$ for each $p$ in the next
     frame, i.e., where $p$ moves to, using optical flow estimation
     [20]. Run the algorithm in reversed time order to estimate
     the corresponding pixel $p''$ for $p'$ in the current frame.
  3. If the distance between $p$ and $p''$ is less than 1 pixel, link $p'$
     to the end of $t$; otherwise close $t$.
  4. Detect face using Haar cascade classifiers; if a face is de-
     tected, add it to $\mathcal{F}$.
  5. Repeat steps 2 to 4 for all frames, update $\mathbf{P}$ and open new
     tracks $t$ by searching feature points again at a fixed interval
     of 5 frames.
  6. Keep $t$ if $t$ is in the face region for more than $0.9\tau$ frames,
     where $\tau$ is the count of detected faces in $\mathcal{F}$ within the life
     span of $t$; otherwise remove $t$.
  7. Remove $t$ if $t$ is longer than 3 seconds and not moving, i.e.,
     range and standard deviation of its coordinates are below
     certain thresholds.
Output: Set of head motion tracks $\mathcal{T}$
```

In step 6 we set the threshold to 0.9 in order to tolerate errors in face detection. In step 7 we remove stable points on the background that are in the face region. This approach is more robust than directly detecting face in low quality video as in [14]. On average, faces are detected in 52% of frames in a session. In the worst sessions, the above rate is less than 1%. The achieved average time coverage by tracks is about 0.97. Although more advanced tracking methods can be applied, we would like to leave it for future development, and focus more on behavior analysis using this simple and satisfactory implementation.

We compute the head motion by averaging the motion of all tracks in each frame, separately on horizontal and vertical directions. Due to the variability of the subject-to-camera distance, we scale the motion signal through dividing by the average side length of the detected square face region. This results in the head motion signal on horizontal direction (X-dir) and vertical direction (Y-dir) as $M_x(n)$ and $M_y(n)$, respectively.

3.2. Power spectral analysis

Recall Hadar *et al.* [10] suggested that magnitude and frequency are plausible dimensions to quantize head motion. We compute the power spectrum of head motion signal, as it reflects the energy distribution of the signal with respect to frequency. The procedure of computing power spectrum in analysis windows is shown as follows.

Input: Motion signal $M_x(n)$, $n \in \{0, 1, \cdots, N-1\}$, where N is the session length.

1. Take an analysis window of length $W = 60$ samples (2 seconds). Compute the autocorrelation function $R(M_x, l)$ for $l \in \{-W+1, -W+2, \cdots, W-1\}$

$$R(M_x, l) = \sum_{n=0}^{W-1} M_x(n) M_x(n+l)$$

2. Compute power spectrum $S(M_x, f)$ as absolute value of 128 points Discrete Fourier Transform of $R(M_x, l)$ (length 119), where $f \in \{0, 1, \cdots, 127\}$.
3. Shift the analysis window by 30 samples (1 second), and repeat step 2 until finishing the entire session.

Output: Power spectrum $S(M_x, f)$ for every second in the session.

Note that the first 64 points of $S(M_x, f)$ samples the frequency band of 0 to 15 Hz (30 fps, 128 point DFT). We segment these 64 points into three parts: $\{f = 0\}$, $\{f = 1, 2, \cdots, 15\}$, and $\{f = 16, 17, \cdots, 63\}$, each representing DC component, the frequency band below 3.5 Hz, and the frequency band from 3.5 Hz to 15 Hz, respectively. Considering three rounds of moving in a second (3 Hz) as a reasonable upper bound of the head movement speed, we keep the second part of $S(M_x, f)$ listed above, for $f \in \{1, 2, \cdots, 15\}$. Similarly, we obtain the power spectrum for Y-dir $S(M_y, f)$, $f \in \{1, 2, \cdots, 15\}$ in each shifting window. In the end, we convert the spectrum to log scale.

For comparison, we also compute the linear prediction features as suggested in [14], specifically Linear Spectral Frequencies (LSF) which have better quantization properties [21]. We do the shifting window for LSF in the same way, and compute 10-order LSF features $L_x(j)$ and $L_y(j)$, $j \in \{1, 2, \cdots, 10\}$, for X-dir and Y-dir in each window respectively.

3.3. Head motion clustering

We apply data-driven clustering to the motion representations in order to quantify motion into certain types. Since the best number of clusters is unknown, and may be subject to target behavior, we try different number of clusters as follows.

Input: Set of motion representations $\mathcal{M}$ by aggregating samples of $S(M_d, f)$ or $L_d(j)$, $d \in \{x, y\}$ in all training sessions.

1. Find mean μ, variance σ^2 of $\mathcal{M}$. Apply zero-mean, unit-variance normalization to obtain $\overline{\mathcal{M}}$.
2. Randomly select 10% of data in $\overline{\mathcal{M}}$ and run a smaller scale K-means clustering to initialize the cluster centroids C^i for $i \in \{1, 2, \cdots, K\}$, and then optimize on $\overline{\mathcal{M}}$ till the K-means algorithm converges (using Euclidean distance).
3. Repeat step 2 for 5 times, and keep the resulting $\{C^i\}_{i=1}^K$ with minimum total distance to associated samples in $\overline{\mathcal{M}}$.
4. Repeat step 2 and 3 for $K \in \{4, 5, \cdots, 25\}$.

Output: Feature mean μ and variance σ^2. Groups of cluster centroids $\{C_K^i\}_{i=1}^K$, for $K \in \{4, 5, \cdots, 25\}$.

In step 4, K is examined from a very small number to a number that is much larger than usually defined by psychologists in coding head motion. By automatic clustering, we hope that finer structure can be identified. For unseen sessions, we assign any normalized motion sample to the cluster with the nearest centroid.

In the context of behavior modeling of couples' conversations, we design three methods of building motion models, from more generic to more specific.

M1 On all spouses using both X-dir and Y-dir combined (feature vectors of X-dir and Y-dir concatenated).
M2 On all spouses for X-dir and Y-dir separately.
M3 For wife and husband separately, and for X-dir and Y-dir separately.

3.4. Behavior modeling

We use the histogram of cluster IDs (*i.e.,* the counts of motion samples in each cluster) for each spouse in a session as the final feature vector for classification, akin to a "bag-of-words" model. For M2 and M3, we concatenate the histograms derived from the models of X-dir and Y-dir to form a single feature vector. For the classification experiments, we consider a gender specific setting. We divide our data into three parts given each gender and behavior code, *i.e.,* $\mathcal{D}_w^1$ and $\mathcal{D}_w^{-1}$ being the top and bottom 25% sessions of the behavior code for the wife, while $\mathcal{D}_w^0$ being the middle 50% (the split may change for different codes). In the same manner denote $\mathcal{D}_h^1$, $\mathcal{D}_h^{-1}$, and $\mathcal{D}_h^0$ for the husband. We train the cluster models on $\mathcal{D}_w^0 \cup \mathcal{D}_h^0$ for M1 and M2, and on $\mathcal{D}_w^0$, $\mathcal{D}_h^0$ separately for M3.

We examine behavior code prediction on $\mathcal{D}_{w,h}^1$ and $\mathcal{D}_{w,h}^{-1}$ as a binary classification problem, where the behavior characteristics are more prominent. Due to data sparsity, we only have 112 sessions in $\mathcal{D}_w^1 \cup \mathcal{D}_w^{-1}$, and similarly for $\mathcal{D}_h^1 \cup \mathcal{D}_h^{-1}$. In order to find the best performing clustering model with different K, we conduct a two-level leave-one-session-out cross-validation. First, the outer level leaves one session out. Second, the inner level conducts leave-one-session-out over the rest 111 sessions in order to find the best performing K. Third, the selected cluster model is used to classify the left out session in the outer level. Finally, repeat for all sessions in the outer level to find the averaged prediction accuracy.

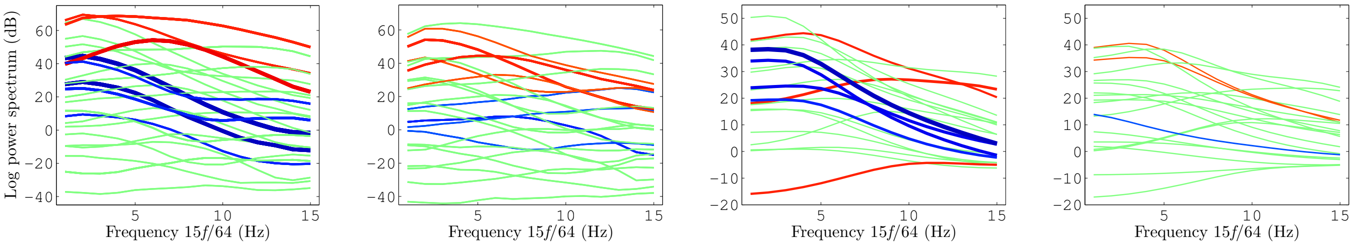
4. EXPERIMENT RESULTS

We show the overall behavior code classification accuracies obtained as described in Sec. 3.4 in Table. 1. These are shown for all methods (M1, M2, M3 as in Sec. 3.3), features (PS for power spectrum and LSF for line spectral frequencies), and interlocutors (W for Wife, H for Husband). We can see that the accuracies are mostly above 0.62 ($p < 0.01$ in binomial test), but that variation among codes and settings is also large.

Table 1: Behavior codes binary classification accuracies. The best performing cases are highlighted with bold. The case of M2/Wife/*Blame* examined in Sec. 5 is highlighted with dark red.

Mod.	Fea.	Gen.	*Accep.*	*Blame*	*Posit.*	*Negat.*
M1	PS	W	0.71	0.63	0.72	0.70
M1	LSF	W	0.74	0.66	0.76	0.66
M1	PS	H	0.64	0.71	0.53	0.57
M1	LSF	H	0.68	0.67	0.56	0.63
M2	PS	W	0.76	0.75	**0.85**	0.72
M2	LSF	W	0.75	**0.81**	0.83	**0.79**
M2	PS	H	0.73	**0.84**	0.57	0.76
M2	LSF	H	0.58	0.62	0.66	0.68
M3	PS	W	**0.77**	0.67	0.81	0.66
M3	LSF	W	0.76	0.79	0.76	0.67
M3	PS	H	**0.79**	0.80	0.65	**0.85**
M3	LSF	H	0.72	0.64	**0.72**	0.71

We compare the performance of different modeling methods, averaged across all features, codes and interlocutors. The averaged



(a) PS, X direction

(b) PS, Y direction

(c) LSF, X direction

(d) LSF, Y direction

Fig. 2: Power spectrum of cluster centroids, red – larger counts in high *Blame* sessions, blue – larger counts in low *Blame* sessions, green – no significant difference (for X: $p > 0.01$, for Y: $p > 0.1$), width – wider for higher significance in ANOVA test

Table 3: Summary of the properties of *Blame*/Non-*Blame* indicating clusters

Direction	*Blame*	Power-spectrum	Head motion types
horizontal	high	large value, rising or concave	head swing, shaking widely, moving actively with speech
horizontal	low	small value, falling	head tilting, turning a small angle, small quick shaking
vertical	high	large value, falling	leaning forward or backward, moving actively with speech
vertical	low	small value, rising or concave	head dipping, tilting downward, small nodding

Table 2: Comparison of average accuracies by PS and LSF

Feature	M2-W	M2-H	M3-W	M3-H	Average
PS	0.77	0.73	0.73	0.77	0.75
LSF	0.80	0.64	0.75	0.70	0.72

accuracies for M1, M2 and M3 are 0.66, 0.73, and 0.74, respectively. The performances of both M2 and M3 significantly exceed M1. This suggests that clustering head motion along X-dir and Y-dir separately is better than clustering X-dir and Y-dir combined. Though X-dir and Y-dir combined clustering is more comprehensive and should describe the motion types more accurately, it may suffer from data sparsity as the possible set of clusters may be a product of those separately for X-dir and Y-dir. It also makes the assumption implicitly that X-dir and Y-dir are perceptually equal if no weighting is added for either direction.

In Table. 2 we compare the performance of PS and LSF with respect to gender, averaged across all codes. We can see that they have comparable performances, while PS is slightly better than LSF ($p < 0.01$) by the overall average. Moreover, M3 has no significant improvement over M2 in average (0.74 *vs.* 0.73, $p > 0.1$); but for the wife M2 is slightly better ($p < 0.05$) than M3, while for the husband the preference is reversed. Such trend is exhibited with both PS and LSF. This may suggest that proper model selection is required for different problems.

5. CASE STUDY OF HEAD MOTION CLUSTERS

We aim to investigate what clusters have been found in the data-driven process, and how they link with behavioral characteristics. In our discussion below we employ M2 and analyze the *Blame* behavior of the wife. We choose this because of its overall performance and especially high accuracy (0.75 for PS, 0.81 for LSF) for the specific conditions of interest, marked in dark red in Table 1.

Recall we conducted two-level cross-validation (Sec. 3.4). For M2/Wife/*Blame* we found that the most frequently selected clustering models are $\{C_{25}^i\}_{i=1}^{25}$ for PS ($K = 25$) and $\{C_{20}^i\}_{i=1}^{20}$ for LSF ($K = 20$). In Fig. 2a and Fig. 2b we plot the cluster centroids in PS-$\{C_{25}\}$. We compute the power spectra of the impulse responses generated by the all-pole filters that LSF-$\{C_{20}\}$ centroids describe, and plot the same frequency band as PS in Fig. 2c and Fig. 2d.

For each cluster, we check if the mean count of motion samples in high score sessions D_w^1 exceeds that in low score sessions D_w^{-1} (and vice versa), using ANOVA test of mean value. The test results are illustrated in Fig. 2. Clusters that show up more in D_w^1 or D_w^{-1} are colored in red or blue, respectively. Such *Blame*-indicating properties lend support to the prediction power of the cluster based cues. Interestingly, we see that the clusters derived from PS and LSF features share similar behavior indications.

In Table. 3 we summarize the properties of the clusters with respect to the association with *Blame*, power-spectral shape, and the associated head motion types (partial empirical list) observed by visually checking typical motion samples of the cluster.

6. CONCLUSION

In this work we clustered head motion and used the counts of motion samples in each cluster to build predictive models of expert judgments on couples' interactive behavior characteristics. We found that motion clustering along horizontal and vertical directions separately was better than when combined. Gender specific motion models had similar performance as gender-independent models in average, while gender-wise differences of model performances existed. Power-spectral features had slightly better performance than line spectral frequency features in terms of averaged accuracy. We interpreted the properties of the derived clusters in a case study, and tested their relations to behavioral characteristics.

Future work will focus on ways to improve the motion model, *e.g.,* tracking the head angle to achieve less interference between horizontal and vertical motion; (semi-)supervised training with manually labeled typical motion samples; using up-to-date devices to record high quality interaction data for faithful motion signal extraction. For improving behavior judgment prediction, several steps can be taken including capturing dynamic relation of the couples' behaviors, fusion of multimodal observation cues, and regression of behavior score instead of binary classification.

7. REFERENCES

[1] S. Narayanan and P. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceeding of IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[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*, vol. 55, no. 1, pp. 1–21, 2013.

[4] T. Wheatley, O. Kang, C. Parkinson, and C.E. Looser, "From mind perception to mental connection: Synchrony as a mechanism for social understanding," *Social and Personality Psychology Compass*, vol. 6, no. 8, pp. 589–606, 2012.

[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] B. Xiao, P.G. Georgiou, Z.E. Imel, D.C. Atkins, and S.S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling," in *Proc. Interspeech*, 2013.

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

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

[9] 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.

[10] U. Hadar, T.J. Steiner, E.C. Grant, and F.C. Rose, "Kinematics of head movements accompanying speech during conversation," *Human Movement Science*, vol. 2, no. 1, pp. 35–46, 1983.

[11] E. Murphy-Chutorian and M.M. Trivedi, "Head pose estimation in computer vision: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 607–626, 2009.

[12] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behavior: A survey of related cues, databases, and tools," *Image and Vision Computing*, 2012.

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

[14] B. Xiao, P.G. Georgiou, B. Baucom, and S.S. Narayanan, "Data driven modeling of head motion towards analysis of behaviors in couple interactions," in *Proc. ICASSP*, 2013.

[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] J. Shi and C. Tomasi, "Good features to track," in *Proc. CVPR*. IEEE, 1994, pp. 593–600.

[20] J.-Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 5, 2001.

[21] P. Kabal and R.P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 6, pp. 1419–1426, 1986.