An Investigation of Vocal Arousal Dynamics in Child-Psychologist Interactions using Synchrony Measures and a Conversation-based Model

Daniel Bone[1]*, Chi-Chun Lee*[2]*, Alexandros Potamianos*[1]*, Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA, USA
[2]Electrical Engineering Department, National Tsing Hua University, Taiwan.
http://sail.usc.edu

Abstract

Researchers from various disciplines are concerned with the study of affective phenomena, especially arousal. Expressed affective modulations, which reflect both an individual's internal state and external factors, are central to the communicative process. Bone et al. (2012) developed a robust, unsupervised (rule-based) method which provides a scale-continuous, bounded arousal rating from the vocal signal. In this study, we investigate the joint-dynamics of child and psychologist vocal arousal in autism spectrum disorder (ASD) diagnostic interactions. Arousal synchrony is assessed with multiple methods. Results indicate that children with higher ASD severity tend to lead the arousal dynamics more, seemingly because the children aren't as responsive to the psychologist's affective modulations. A vocal arousal model is also proposed which incorporates social and conversational constructs. The model captures conversational signal relations, and is able to distinguish between high and low ASD severity at accuracies well-above chance.

Index Terms: vocal arousal, interaction, synchrony, autism spectrum disorders

1. Introduction

Arousal (also referred to as activation and excitation) is a primary component in dimensional theories of emotion [1, 2], and continues to be the focus of interdisciplinary work in domains such as psychology, engineering, linguistics, and biology. Arousal is an internal state that, like other affective constructs, influences our thoughts and actions. For instance, a person's arousal level can affect the decisions they make [3, 4] and their performance on certain tasks [5].

The transmission of multimodal affective cues is a core facet of human communication. Perceptual tests and engineering systems have established that expressed affective cues can be discerned from speech [6, 7, 8, 9, 10]. Although arousal is reliably decoded from vocal cues, engineering tools that are broadly applicable to unlabeled databases are lacking. State-of-the-art supervised systems usually incorporate thousands of features (e.g., openSMILE [11]); while large feature sets increase capacity for modeling behavior, they reduce interpretability and risk overfitting in cross-corpora experiments. Bone et al. (2012) developed and validated an alternative unsupervised (rule-based) framework for vocal arousal rating that utilizes three features (pitch, vocal intensity, and the ratio of high- and low-frequency energy) [12]. This framework enables general, interpretable study of vocal arousal with few constraints.

In human-human interaction, behavioral synchrony (or entrainment) between participants is central to perceptions of overall quality. Harrist & Waugh (2002) define synchrony as a "type of interaction between two people ... an observable pattern of dyadic interaction that is mutually regulated, reciprocal, and harmonious" [13]. Several studies have concentrated on infants and adolescents to understand the development and importance of entrainment processes [13, 14, 15]. Such work has reported a higher dyadic synchrony when mothers interact with their own infant, rather than an unfamiliar infant [14]; and that parent-infant synchrony is predictive of symbolic play complexity [15]. Overall, the development of behavioral synchrony is seen as key to establishing significant dyadic relationships, enabling the child to grow socially and emotionally [13].

Behavioral synchrony computation often relies on hand-coded behavioral signals like gaze, vocalizations, and affect [16]. There is an apparent need for scalable automatic techniques. Engineering methods in computing synchrony are gaining attention. Studies have used automatic measures of heart rate [16]; facial features such as smile strength, eye constriction, and mouth opening [17]; and prosodic and spectral features [18]. The present work investigates the joint-dynamics of automatically-generated vocal arousal contours.

Autism spectrum disorder (ASD) is a highly heterogeneous and highly prevalent (1 in 88 children [19]) neurodevelopmental disorder characterized by social communication deficits, social impairments, and the presence of restricted, repetitive, and/or stereotyped behaviors [20]. Computational models of social behavior have the potential to aid clinicians in diagnosis, intervention, and long-term monitoring. Further, neurobiological studies in autism need quantitative, dimensional measures of behavior for improved stratification [21]. Toward this end, Bone et al. (2012b, 2014) found through speech-prosodic analysis of ASD diagnostic sessions that the psychologist adjusted their speech properties based on the child's social-communicative impairments [22, 23]. Bone et al. (2013) additionally investigated automatically-extracted turn-taking and language cues, observing an overall degradation of conversational quality when the psychologist interacted with children with higher ASD severity [24]. The current work extends this line of research, investigating the joint-evolution of affect as captured by vocal arousal.

In this study, we examine child-psychologist affective synchrony in relation to ASD severity. We note that vocal arousal is a function of both internal state and external social and conversational factors. Thus, we additionally propose a model of vocal arousal dynamics which incorporates both internal (self-evolution) and external social and contextual factors. This model is then utilized to discriminate between groups of interaction sessions, divided according to the child's ASD severity.

2. Experimental Design

2.1. The USC CARE Corpus

Child-psychologist interactions are studied in the context of the Autism Diagnostic Observation Schedule (ADOS, [25]). The present work focuses on the ADOS Module 3, designed for subjects who are verbally fluent as judged by the psychologist. During administration, the psychologist leads the child through a variety of activities, or subtasks, designed to elicit social responses. The psychologist then scores 28 codes representing the child's behaviors in the domains of Social Interaction, Communication, and Restricted, Repetitive Behaviors. This analysis uses the revised ADOS Module 3 algorithms [26], and the transformation of the ADOS total to an ASD severity score [27]. The ASD severity score is in the range 1 to 10, with higher scores indicating higher severity of ASD symptoms.

The USC CARE Corpus [28] is comprised of audio-video recordings (2 HD cameras and 2 high-quality far-field microphones) of ADOS administrations. Sessions are lexically transcribed based on the SALT transcription manual [29], and temporally marked for utterance boundaries. Demographics of the 29 participants for this study are presented in Table 1.

As with previous studies conducted with this corpus [22, 23, 24], we examine both child and psychologist behavior. Three licensed, research-certified psychologists with extensive clinical experience with ASD children administered the ADOS. Two of the psychologists were bilingual in English and Spanish; bilingual participants were evaluated by bilingual psychologists. Administrations were conducted in English, so small portions of Spanish conversation are disregarded; one subject (of 30) was excluded due to a primarily Spanish discourse.

Table 1: *Demographic statistics of the 29 recorded children in this study that were administered Module 3 of the ADOS.*

Category	Count/Statistic
Age (years)	mean: 10.0, std. dev.: 2.6, range: 5.8-15.0
Gender	male: 23, female: 6
Native language	Spanish: 9, English: 10, Sp&Eng: 4, unk: 6
Ethnicity	Hispanic: 20, White/+Other: 8, AF-AM: 1
ADOS module	#3: 29
ADOS diagnosis	autism: 18, ASD: 5, below ASD cutoffs: 6

2.2. Vocal Arousal

Expressed vocal arousal is quantified using a method proposed in Bone et al. (2012a) [12]. This rule-based method can provide a scale-continuous arousal rating, bounded in the range [-1,1], from the vocal signal without the need for any manual labeling. The arousal rating is built upon three knowledge-inspired features, whose individual scores are fused to improve reliability. The method performs consistently across multiple corpora, achieving both high correlations with labels and impressive binary classification performance.

The framework tracks a speaker's variation from a baseline (e.g., higher pitch indicates higher arousal). This baseline data can be neutral-labeled or unlabeled (global normalization). In the case of neutral-data normalization, positive (negative) ratings can be interpreted as higher-than-neutral (lower-than-neutral) arousal. If no labeled data is available, the method is still able to rank instances according to perceived arousal via global normalization; here, the relative value of instances still has meaning, but the absolute value is less interpretable.

In this corpus, vocal arousal is extracted for both speakers. The data is organized into turns. A turn consists of consecutive, uninterrupted utterances from a single speaker. Utterances that are entirely overlapped by another speaker's turn are excluded (rather than performing source separation), while overlaps that represent speaker changes are maintained. Sessions vary from 76-326 turns for each participant. Features are extracted within the voiced frames of a turn. Vocal arousal is computed for each turn with global speaker-normalization.

Sample vocal arousal streams for child and psychologist are shown in Figure 1. Coupling is apparent in this 50 turn sample; the Spearman's rank-correlation coefficient between the two arousal contours is ρ_S=0.66. A varying lead-lag relationship is also evident: In the first segment ("Psych leads"), the psychologist's arousal precedes the child's arousal by one turn. (*By convention, time slot t contains the arousal from the psychologist at turn k and the following turn from the child,* k+1.) In the second segment ("Child leads"), the child's arousal precedes the psychologist's arousal by approximately two turns.

2.3. Synchrony Measures

We consider two measures of synchrony: cross-correlation with peak-picking and Granger causality. Windowed cross-correlation with peak-picking [30] is used to evaluate both the magnitude of interaction between vocal arousal streams (peak absolute value of cross-correlation), and the lead/lag relationship (corresponding peak index). Correlation can be either positive or negative, reflecting synchronous or asynchronous interaction; we consider the absolute magnitude of the correlation.

Granger causality [31] is a statistical method for determining if one signal is useful in predicting another; in this case, through linear autoregressive models for time series prediction [32]. Given time-varying signals X(t) and Y(t), if predicting Y with previous values of X reduces the residual signal energy compared to using Y alone, then X is said to Granger-cause Y. Three Granger features are extracted [33]: the strength of Granger-causality from the child to the psychologist (and vice-versa), taken as the logarithm of the F-statistic; and the child's causal flow, the difference of the out-degree of Granger causal-

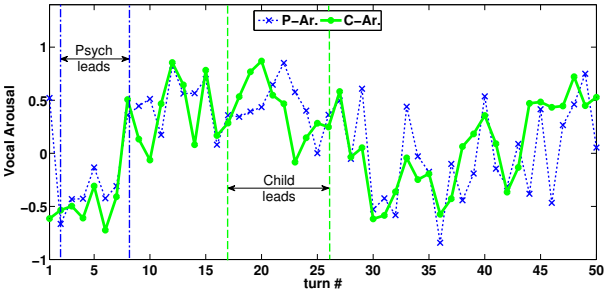

Figure 1: *Example vocal arousal streams from child (C) and psychologist (P) with highlighted regions of synchrony.*

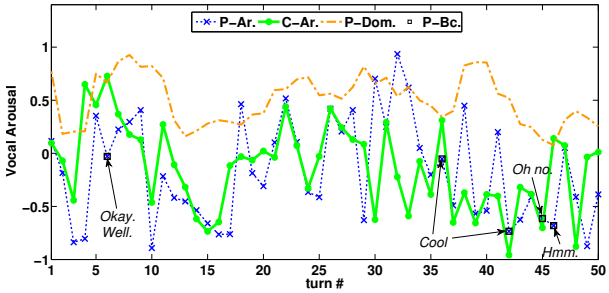

Figure 2: *Example vocal arousal streams for child (C) and psychologist (P) with dominance (Dom) and backchannels (Bc).*

ity minus the in-degree. Features were extracted within overlapping windows. The number of lags for Granger analysis was decided per window through Bayesian Information Criterion. Given the limited length of these sessions, Granger-causality magnitudes were included in computations whether or not they reached statistical significance. All computations are made using the GCCA toolbox [34]. A similar approach was employed to model dominance effects with non-verbal behaviors in [35].

2.4. A Conversational Model of Vocal Arousal

Vocal arousal is an expressed signal that is not only reflective of internal arousal, but also social and conversational factors. Vocal arousal evolves as a function of the interaction; in this data, the coupling is quite significant (Section 3.1). Since a person's vocal arousal depends on their conversational partner's vocal arousal, this is one component of the model we propose in Figure 3; corresponding signals are shown in Figure 2.

The vocal arousal for a particular turn is certainly dependent upon the spoken content. For this purpose, we include dialogue acts in our model of vocal arousal. For instance, compared to acknowledgments, backchannels are expected to be of lower volume and pitch movement; this is because backchannels are spoken with the intention of not overtaking the floor, whereas acknowledgments assert an opinion [36]. Since these features are positive correlates of perceived arousal [9], we may also expect vocal arousal to be lower for backchannels.

The evolution of vocal arousal contours is additionally contingent on the style of the conversation (e.g., interview or casual conversation). We restrict our focus to a speaker's conversational dominance (power or control). Dominance is a global factor of an interaction, but it varies locally as well. For example, Wollmer et al. (2012) investigated the temporal modeling of dominance using acoustic features [37]. Acoustic features like vocal intensity influence perceptions of dominance [38]. We include temporal dominance in our vocal arousal model such that any dependency may be captured and quantified.

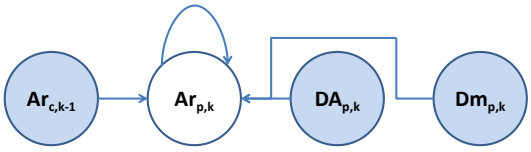

Figure 3: *Conversational model of vocal arousal (psychologist's view shown). Note: p - psychologist; c - child; Ar - vocal arousal; DA - dialogue act; Dm - dominance; bold indicates a vector ending at turn k.*

2.4.1. ASD Severity Classification with Arousal Model

We implement the proposed model as a vocal arousal sequence prediction model using linear-chain conditional random fields (CRFs). The vocal arousal of speaker A is predicted by the previously mentioned social and conversational factors: speaker B's vocal arousal at the previous turn (plus Δ and $\Delta\Delta$), speaker A's current dialog act, and speaker A's current dominance. Since dialogue act annotations are not available for the USC CARE Corpus, we restrict the dialogue act set to *backchannels* and *others*. Backchannels are defined as turns which are at least one-third composed of words from the set listed in Table 2; this set overlaps with the most common backchannels in the Switchboard corpus [39]. We accept that false positives will occur.

Dominance is perceived through several cues. Prosody is already used in vocal arousal, so we rely on another feature. Total speaking length has been shown effective in dominance prediction [38]. We devise a temporal dominance feature based on turn length. For a dyadic conversation, we define speaker A's dominance at turn *k* as the ratio of the length of speaker A's previous turns to the total length of speaker A's and speaker B's previous turns; we use 3 previous turns with a decaying weight function of [*0.15, 0.30, 0.55*] for turns [*k-2, k-1, k*].

Table 2: *List of words defined as a backchannel in our corpus.*

List of backchannel words
'mm', 'hmm', 'mm-hmm', 'uh', 'huh', 'uh-huh', 'um', 'ah', 'oh', 'okay', 'yeah', 'yes', 'cool', 'nice', 'alright', 'I see', 'my goodness', 'oh no'

In Section 3.2, we perform binary classification of ASD severity with CRFs through a leave-one-session-out approach. We define *high-severity* ASD as an ASD severity of 7 or higher. This division produces an approximately equal distribution of 14 *high-severity* and 15 *low-severity* sessions. CRF models are trained with the HCRF toolbox [40]. CRFs are first trained for each class to predict speaker A's vocal arousal sequence given a set of features. Then maximum likelihood classification is performed by selecting the model that produces the highest likelihood for the observed arousal sequence of the test session. Rather than computing the likelihood of the exact sequence, we define the sequence likelihood as the product of the likelihood of each observation in the sequence. Vocal arousal is quantized into three equally-balanced categories per session (high, medium, and low). Dominance (Dm) is quantized into three categories as well: Dm$<=1/3$, $1/3<$Dm$<=2/3$, and Dm$>2/3$.

3. Results and Discussion

In Section 3.1 the various measures of arousal synchrony are investigated in relation to ASD symptom severity. Then in Section 3.2, we model the arousal dynamics conditioned on other relevant social and conversational features; the content captured by this model is evaluated through a classification task.

3.1. Vocal Arousal Synchrony

The social-affective exchange between child and psychologist is expected to differ if the child has social-communicative difficulties. In this section, the synchrony between vocal arousal signals is related to the severity of the child's ASD symptoms.

First, we consider the strength of coupling between the signals, calculated as the maximum absolute cross-correlation value. In order to capture different styles of interaction (e.g., synchrony/asynchrony or varying linear predictive coefficients with Granger analysis), the sessions are windowed. Window sizes (W_s) of 25-70 turns are used with a step-size of $W_s/2$. Within each window, the lag that maximizes absolute correlation is chosen; then the median coupling and lag are computed across all windows in a session. Coupling magnitude is medium in the sessions: $mean$=0.41 ($stdv.$=0.12) for W_s=25. However, no relation is observed between this coupling magnitude and ASD severity ($p>0.05$) for any window size.

Next we ask the question, 'Who leads the affective exchange and how does it relate to ASD severity?' This topic is first examined using cross-correlation with peak picking. Referring to Figure 4, a significant positive correlation frequently occurs between peak-lag and ASD severity (4 times $p<0.05$, twice $p<0.10$, and 4 times $p>0.10$). This indicates that as ASD severity increases, the child tends to lead the arousal coordination more, as illustrated in Figure 5. For very low ASD severity, the psychologist leads the interaction; but for very high ASD severity, the child often leads. However, it is uncertain whether this occurs because the child is less influenced by the psychologist, or because the psychologist is more attuned to the child.

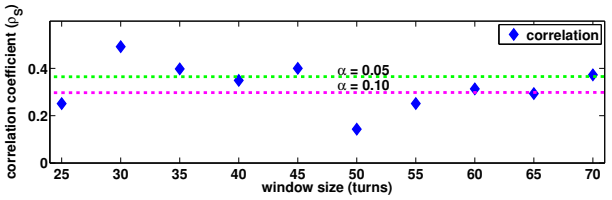

Figure 4: *Correlation between lag (positive when child leads) and ASD severity vs. window length.*

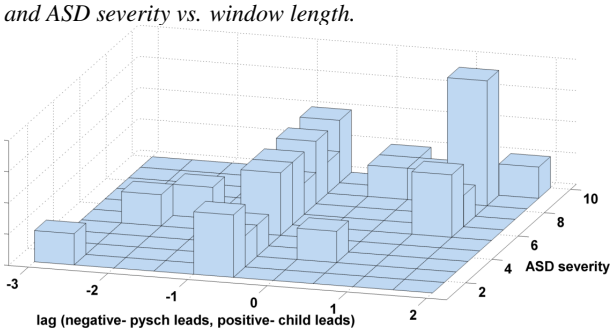

Figure 5: *Lag vs. ASD severity for* W_s=30. ρ_S=0.49 (p<0.01)

In order to gain insight into the causal structure of the interaction relative to ASD severity, we utilize Granger causality analysis. The results (displayed in Figure 6) vary somewhat according to window size. The child's influence on the psychologist ($F_{C->P}$) is not statistically significantly related to the child's ASD severity (p>0.05) for any window size. But the psychologist's influence on the child ($F_{P->C}$) is found to decrease with higher ASD severity for W_s=35 (p<0.05). Also, the child's causal flow ($cf(S)$) is significantly more outward given higher ASD severity for W_s=45 (p<0.05). These observations suggest that a child with higher ASD severity is less influenced by the psychologist's vocal arousal. No significant relation to ASD severity is found for the psychologist's attunement to the child's behavior. While these results cohere with the cross-correlation analysis, they should be interpreted cautiously since few parameter settings showed significance.

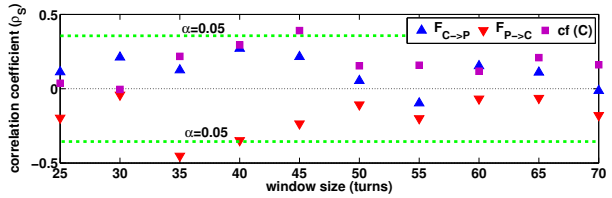

Figure 6: *Correlation between Granger causality parameters and ASD severity vs. window length.* $F_{C->P}$ *is the magnitude of interaction for the child G-causing the psychologist's behavior, and vice-versa for* $F_{P->C}$. *cf(C) is the child's causal flow.*

3.2. Classification with Conversational Model

The relationships between a speaker's vocal arousal and other social (i.e., partner's vocal arousal) and conversational (i.e., backchannels and dominance) factors may capture social-communicative patterns associated with ASD severity. In this section, we perform classification of high and low ASD severity groups using CRF vocal arousal predictive models.

First, we validate that the model captures statistical relations between vocal arousal and the proposed features by examining perplexity. Perplexity is calculated with whole sequence prediction via leave-one-session-out cross-validation. Since the vocal arousal is evenly distributed among three levels, maximum perplexity is $log_2(3)$=1.59. We find a small decrease in perplexity when using backchannels or dominance as features. The largest gain comes from including the partner's vocal arousal; recall that arousal coupling is rather strong in this database. Thus, the model captures information about the relation between feature streams. The remaining question is whether information is discriminative.

Table 3: *Model perplexity as a function of feature input.*

feature	*baseline*	*partner*	*backch.*	*dom.*	*all*
perplexity psych	1.59	1.49	1.55	1.57	1.46
perplexity child	1.59	1.48	1.57	1.58	1.47

The results in Table 4 indicate that the model can discriminate between high and low ASD severity. Predicting the psychologist's or child's vocal arousal with the preceding vocal arousal of the other speaker achieves above 50% unweighted average recall (UAR), but this is not significantly above chance. The same is observed with the other features used in predicting the child's vocal arousal. Fusion decreases performance, likely due to insufficient data size. Interestingly, the relationships of the psychologist's vocal arousal with backchannels (80% UAR) and dominance (79% UAR) discriminate ASD severity.

Table 4: *Classification performance in UAR. Note:* ***bold*** *indicates significance above chance at* p<0.01.

feature	*chance*	*partner*	*backch.*	*dom.*	*all*
predict psych	50%	58%	**80%**	**79%**	**75%**
predict child	50%	55%	59%	62%	48%

Conditional probability tables may support interpretation. For backchannels, the largest difference appears to be the probability of low vocal arousal given the current state is a backchannel, or p($Ar_{p,k}$=$low|BC_{p,k}$=$true$). This probability is higher for the high ASD severity group (0.57 vs. 0.49). This result could occur if the psychologist is more cautious of taking over the floor during backchannels for the children with higher ASD severity. For dominance, there is some difference between groups in the conditional probabilities p($Ar_{p,k}|Dm_k$) for high and low dominance, but the interpretation is less certain. We do note that the psychologist is more dominant (mean of Dm_p) for the high ASD severity group (p<0.05). Still, the proposed vocal arousal model captures relations between vocal arousal and relevant feature streams that is informative of ASD severity. There is again the intriguing finding that the psychologist's behavior alone is informative; this result mirrors findings that the psychologist adjusts prosody, turn-taking, and language behavior relative to the child's social-communicative difficulties [24].

4. Conclusions and Future Work

We investigated how the social-affective interaction between child and psychologist varies according to ASD severity. Vocal arousal is demonstrated to be a useful automatic measure for affective synchrony studies. The findings reveal that the child with higher ASD severity is less responsive to the psychologist, and thus appears to lead the affective exchange. We also proposed a vocal arousal model that incorporates social and conversational influences. The model discriminated between sessions involving children with high and low ASD severity, using only the psychologist's behavior.

Future work will seek to refine the conversational model and evaluate on a larger conversational database. We believe that affective-dynamic analysis can provide key insights and advancements toward one of the major goals of Behavioral Signal Processing [41], providing tools that support clinicians.

5. Acknowledgments

This work was supported by funds from NSF and NIH. D.B. is an Achievement Rewards for College Scientists Scholar.

6. References

[1] M. S. Yik, J. A. Russell, and L. F. Barrett, "Structure of self-reported current affect: Integration and beyond." *Journal of personality and social psychology*, vol. 77, no. 3, p. 600, 1999.

[2] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[3] K. P. Leith and R. F. Baumeister, "Why do bad moods increase self-defeating behavior? emotion, risk tasking, and self-regulation." *Journal of personality and social psychology*, vol. 71, no. 6, p. 1250, 1996.

[4] W. Kroeber-Riel, "Activation research: Psychobiological approaches in consumer research," *Journal of Consumer Research*, pp. 240–250, 1979.

[5] K. Lambourne and P. Tomporowski, "The effect of exercise-induced arousal on cognitive task performance: a meta-regression analysis," *Brain research*, vol. 1341, pp. 12–24, 2010.

[6] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current directions in psychological science*, vol. 8, no. 2, pp. 53–57, 1999.

[7] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–1.

[8] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE TASLP*, vol. 13, no. 2, pp. 293–302, 2005.

[9] P. Juslin and K. Scherer, *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford: Oxford University Press., 2005, ch. 3. Vocal Expression of Affect, pp. 65–135.

[10] C. Lee, A. Katsamanis, M. Black, B. Baucom, P. Georgiou, and S. Narayanan, "An Analysis of PCA-based Vocal Entrainment Measures in Married Couples' Affective Spoken Interactions," in *Proceedings of Interspeech*, 2011.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[12] D. Bone, C.-C. Lee, and S. Narayanan, "A robust unsupervised arousal rating framework using prosody with cross-corpora evaluation," in *Proc. Interspeech*, 2012a.

[13] A. W. Harrist and R. M. Waugh, "Dyadic synchrony: Its structure and function in childrens development," *Developmental Review*, vol. 22, no. 4, pp. 555–592, 2002.

[14] F. J. Bernieri, J. S. Reznick, and R. Rosenthal, "Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions." *Journal of personality and social psychology*, vol. 54, no. 2, p. 243, 1988.

[15] R. Feldman, "On the origins of background emotions: from affect synchrony to symbolic expression." *Emotion*, vol. 7, no. 3, p. 601, 2007.

[16] R. Feldman, R. Magori-Cohen, G. Galili, M. Singer, and Y. Louzoun, "Mother and infant coordinate heart rhythms through episodes of interaction synchrony," *Infant Behavior and Development*, vol. 34, no. 4, pp. 569–577, 2011.

[17] D. S. Messinger, M. H. Mahoor, S.-M. Chow, and J. F. Cohn, "Automated measurement of facial expression in infant–mother interaction: A pilot study," *Infancy*, vol. 14, no. 3, pp. 285–305, 2009.

[18] 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*, vol. 28, no. 2, pp. 518–539, 2014.

[19] J. Baio, "Prevalence of autism spectrum disorders: Autism and developmental disabilities monitoring network, 14 sites, united states, 2008. morbidity and mortality weekly report. surveillance summaries. volume 61, number 3." *Centers for Disease Control and Prevention*, 2012.

[20] *Diagnostic and Statistical Manual of Mental Disorder, Ed. 4 text revision*, American Psychiatric Assoc., Washington D.C., 2000.

[21] C. Lord and R. M. Jones, "Annual research review: Re-thinking the classification of autism spectrum disorders," *Journal of Child Psychology and Psychiatry*, vol. 53, no. 5, pp. 490–509, 2012.

[22] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. Narayanan, "Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist." in *INTERSPEECH*, 2012b, pp. 1043–1046.

[23] ——, "The Psychologist as an Interlocutor in Autism Spectrum Disorder Assessment: Insights from a Study of Spontaneous Prosody," *Journal of Speech, Language, and Hearing Research*, 2014, (in Press).

[24] D. Bone, C.-C. Lee, T. Chaspari, M. Black, M. Williams, S. Lee, P. Levitt, and S. Narayanan, "Acoustic-prosodic, turn-taking, and language cues in child-psychologist interactions for varying social demand," in *Interspeech*, 2013.

[25] C. Lord, S. Risi, L. Lambrecht, E. Cook, B. Leventhal, P. DiLavore, A. Pickles, and M. Rutter, "The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. of Autism and Dev. Dis.*, vol. 30, pp. 205–223, 2000.

[26] K. Gotham, S. Risi, A. Pickles, and C. Lord, "The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity," *Journal of Autism and Developmental Disorders*, vol. 37, no. 4, pp. 613–627, 2007.

[27] K. Gotham, A. Pickles, and C. Lord, "Standardizing ados scores for a measure of severity in autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 39, no. 5, pp. 693–705, 2009.

[28] M. P. Black, D. Bone, M. E. Williams, P. Gorrindo, P. Levitt, and S. S. Narayanan, "The USC CARE Corpus: Child-Psychologist Interactions of Children with Autism Spectrum Disorders," in *Proceedings of Interspeech*, 2011.

[29] J. Miller and A. Smith, "Salt transcription manual: Guidelines for transcribing free speech samples," *Unpublished paper, Language Analysis Lab, University of Wisconsin-Madison*, 1983.

[30] S. M. Boker, J. L. Rotondo, M. Xu, and K. King, "Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series." *Psychological Methods*, vol. 7, no. 3, p. 338, 2002.

[31] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[32] J. D. Hamilton, *Time series analysis*. Princeton university press Princeton, 1994, vol. 2.

[33] A. K. Seth, "Causal connectivity of evolved neural networks during behavior," *Network: Computation in Neural Systems*, vol. 16, no. 1, pp. 35–54, 2005.

[34] ——, "A matlab toolbox for granger causal connectivity analysis," *Journal of neuroscience methods*, vol. 186, no. 2, pp. 262–273, 2010.

[35] K. Kalimeri, B. Lepri, O. Aran, D. B. Jayagopi, D. Gatica-Perez, and F. Pianesi, "Modeling dominance effects on nonverbal behaviors using granger causality," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 23–26.

[36] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[37] M. Wöllmer, F. Eyben, B. W. Schuller, and G. Rigoll, "Temporal and situational context modeling for improved dominance recognition in meetings." in *INTERSPEECH*, 2012.

[38] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 501–513, 2009.

[39] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[40] L. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[41] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. PP, no. 99, pp. 1–31, 2013.