Quantification of Prosodic Entrainment in Affective Spontaneous Spoken Interactions of Married Couples

Chi-Chun Lee[1], *Matthew Black*[1], *Athanasios Katsamanis*[1], *Adam Lammert*[1]
Brian Baucom[2], *Andrew Christensen*[3], *Panayiotis G. Georgiou*[1], *Shrikanth Narayanan*[1,2]

[1]Signal Analysis and Interpretation Laboratory (SAIL), Los Angeles, CA, USA
[2]Department of Psychology, University of Southern California, Los Angeles, CA, USA
[3]Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA
http://sail.usc.edu[1], baucom@usc.edu[2], christensen@psych.ucla.edu[3]

Abstract

Interaction synchrony among interlocutors happens naturally as people adapt their speaking style gradually to promote efficient communication. In this work, we quantify one aspect of interaction synchrony - prosodic entrainment, specifically pitch and energy, in married couples' problem-solving interactions using speech signal-derived measures. Statistical testings demonstrate that some of these measures capture useful information; they show higher values in interactions with couples having high positive attitude compared to high negative attitude. Further, by using quantized entrainment measures employed with statistical symbol sequence matching in a maximum likelihood framework, we obtained 76% accuracy in predicting positive affect vs. negative affect.

Index Terms: interaction synchrony, prosody, entrainment, behavior signal processing, affect, couple therapy

1. Introduction

Interaction synchrony in human-human conversations occurs naturally and spontaneously as a result of the coordination of movements between interacting individuals in both timing and forms during interpersonal communication [1]. Multi-person interactions can often be viewed as coupled interactive dynamical systems. This phenomenon, also known as *entrainment*, is based closely on the principles of self-organization, where humans adapt gradually to each others behaviors to promote efficient communication. Quantification of the degrees of entrainment can happen at different levels, e.g., from examining the overall interaction session to turn-wise synchronization of multimodal behavioral cues. This type of analysis has been of interest in both psychology, and more recently in, engineering research. In this paper, we utilize several turn-wise *latent entrainment/synchronization* measures derived from prosodic cues, specifically pitch and energy, to analyze prosodic entrainment of real married couples engaging in problem-solving spoken interactions. These latent entrainment measures are speech signal-derived behavioral measures. We believe that they can provide a gateway to not only perform quantitative analysis that can strengthen the qualitative aspect of interaction synchrony, but also demonstrate the ability of signal processing tools in capturing such an effect in real and spontaneous dyadic interactions. These entrainment measures can be applied to the automatic annotation of the overall attitude of the interacting partners. They can be further integrated with other streams of features to identify salient portions e.g., patterns of positive affect, agreement, blame, etc., in the couples' interaction.

Several previous studies have considered entrainment from different analysis viewpoints and by using various behavioral cues; for example, McGarva investigated the mutual entrainment in vocal activity rhymes [2], Nenkova analyzed the high frequency word usage entrainment [3], Richardson analyzed the entrainment of body movements [4], and Pardo showed the phonetic convergence in conversation settings [5]. While many investigations have been carried out, there seems to be a lack of agreement on which measure or set of measures can best describe the entrainment phenomenon. Relatively few works have shown the effectiveness of predicting experts' behavioral annotations, such as from a therapist or a psychologist, for describing the interaction or the participants' overall attitude using these entrainment measures. In this paper, we analyze the prosodic synchrony using a variety of turn-level measures derived from couples' interactions. We analyze the prosodic entrainment phenomenon and use these measures to perform automatic annotation on one specific attribute describing the overall husband or wife's attitude during interactions - *positive* vs. *negative*. Works in psychology [1, 6] have posited that the degree of synchrony is higher in a *positive* interaction than in a *negative* one. This work provides a quantitative verification that the same trends hold with these various signal-driven prosodic entrainment measures and with real couples' interactions. It also offers a modeling technique to perform session-level automatic annotation using such features. Such annotations are predominantly derived manually in the state of the art family studies practice and research [7].

Traditional supervised pattern recognition techniques are not well-suited for recognizing the synchronization phenomenon. It is difficult to treat it as a classification or regression task because entrainment is generally difficult for human evaluators to label, especially at turn level, due to its inherently subtle nature. Hence, in this paper, we take the approach of using multiple measures computed at turn level, in particular Pearson correlation, mutual information, and coherence, to quantify such an effect. These measures are computed for stylized pitch and energy at longer time frames, for the entrainment effect is often only apparent at a longer time scale. Two specific experiments were conducted and reported in this work. In the first one, we hypothesize that these signal-driven measures have an overall larger value indicating higher interaction synchrony in positive interactions than in negative interactions to agree with what is commonly known from psychology studies. The second is to use these measures in a Markov sequence modeling framework to show the effectiveness of recognizing couples' attitude.

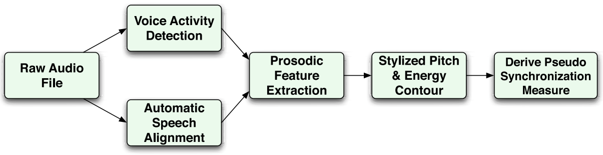
Voice Activity Detection
Raw Audio File
Prosodic Feature Extraction
Stylized Pitch & Energy Contour
Derive Pseudo Synchronization Measure
Automatic Speech Alignment

Figure 1: Block Diagram of Steps from Audio File to *Latent Entrainment* Measures

In summary, our work indicates that the majority of pitch-related entrainment measures are statistically significantly higher during positive interactions compared to negative interactions, while energy-based entrainment measures are not as informative in that regard. Also, we obtain a 76% accuracy in identifying husband and wife's positive vs. negative attitude in their interactions using just these entrainment measures with simple sequence modeling.

The paper is organized as follows: our research methodology is described in Section 2, experiment results and discussion are presented in Section 3, and conclusion and future work are given in Section 4.

2. Research Methodology

2.1. Database

2.1.1. Overview of Corpus

The database that we are using was collected as part of a longitudinal study at the University of California, Los Angeles and the University of Washington [8]. 134 seriously and chronically distressed married couples received couples therapy for one year, and as part of the study, the couples participated in sessions where they discussed a problem in their relationship with no therapist or research staff present. There is a total of 574 sessions of interaction recorded at different time points in the therapy session. The database consist of audio-video data - split-screen video and a single channel of far-field audio. The quality of recording conditions varies from session to session. We also have access to word transcriptions for each session. A more detailed overview of the corpus can be found in [9].

In addition to audio-video data, both spouses were evaluated with 33 session-level codes using two separate coding schemes that were designed for this type of interaction. The Social Support Interaction Rating System (SSIRS) measures both the emotional features of the interaction as well as the topic of conversation. Its 20 codes are broken into 4 categories: Affectivity, Features of the Interaction, Topic Definition,and Dominance/Submission [10]. The Couples Interaction Rating System (CIRS) was designed to be more geared towards conversations involving a problem in the relationship [11]. Three to four trained evaluators coded each session (one set of 33 codes for *each* spouse) on an integer scale from 1 to 9 with written guidelines.

2.1.2. Preprocessing & Annotation of Interest

First, the database was processed by performing automatic alignment of the sessions into speaker-homogeneous regions, which we refer to as *turns* in this paper. We used an iterative automatic speech recognition technique based on [12], and we were able to segment more than 60% of the sessions' words for 293 of the 574 sessions. The rest of sessions were deemed too noisy to achieve good automatic segmentation result and are ignored from the present automated analysis.

In this work, we focus on analyzing the "Global Positive Affect" and "Global Negative Affect" codes (with inter-evaluator agreements of 0.74 and 0.79, respectively) to test our hypothesis of whether the interaction synchrony is higher in interactions with positive affect vs. negative affect. We analyzed only the sessions that had mean scores (averaging across evaluators) that fell in the top 50 for each of these two codes per spouse in this paper. The data of interest for this work consists of 200 sessions of interactions with 60 unique pairs of couples; 100 sessions of them were rated as highly negative affect while the others were highly positive affect.

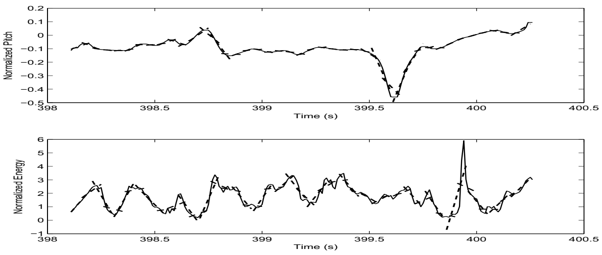

Figure 2: Example of Stylized Pitch and Energy Contour (See Section 2.2.2 for details)

2.2. Latent Entrainment Features

Figure 1 shows a block diagram of the process of computing prosodic entrainment measures and details are described in the following sections.

2.2.1. Prosodic Feature Extraction

Two different prosodic features, pitch and energy, were extracted in order to compute several *latent entrainment* measures described in the next section. Voice activity detection was first applied to the audio in order to exclude the non-speech regions. Then the pitch contour was extracted using the Praat Toolbox [13]. The resulting raw pitch signal was passed through an algorithm that attempted to fix instances of pitch halving or doubling. The pitch signal was then median filtered, and linearly interpolated over unvoiced region. Pitch values were normalized speaker-by-speaker on a logarithmic scale, $\log(f0/f0_\mu)/\log(2)$, where $f0_\mu$ is the mean pitch value per speaker. Energy was extracted using the Praat Toolbox, with normalization done speaker-by-speaker using E/E_μ, where E_μ is the mean energy value per speaker.

2.2.2. Stylized Pitch and Energy Contour

The raw frame-by-frame pitch and energy values themselves are too detailed and last over too short a time window (10ms in our case) for the purpose of observing prosodic entrainment. In order to obtain trends of pitch and energy trajectories, we fitted the contours with a linear line ($Y = \alpha X + B$) computed every 100ms with 50ms overlap. The slope (α) and intercept (B) were calculated using the method of least squares. Figure 2 shows an example of stylized pitch and energy contour over an automatically aligned turn. Hence, we have two parameters for every 100ms to describe the evolution of prosodic cues instead of 100 raw values. In this work, we focus only on the slope, which encodes information describing the intonation and the raise or fall of energy values. We hypothesize the changes of slope values across time captures aspects of speaking style, and the co-variation of this parameter between speakers can help us identify prosodic entrainment.

2.2.3. Entrainment Measures

Stylizing the pitch and energy contours generated two one-dimensional feature vectors consisting of several frames of α's for pitch and energy, respectively, at every automatically-aligned speaker turn. We then computed prosodic entrainment

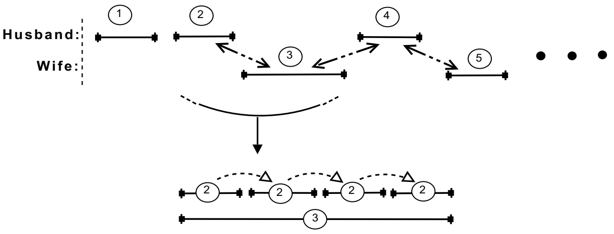

Figure 3: Example of Entrainment Measures Time Frame

based on three main methods: square of Pearson correlation, mutual information, and mean of spectral coherence across turn on this sequence of α's to estimate the level of synchronization. *Correlation coefficient* computes the degree of linear dependency between two random variables; taking the square is to measure an unsigned overall strength of dependency. *Mutual information* is an information theoretic quantity to measure mutual dependence between two random variables. *Spectral coherence* is commonly used to estimate the causality between a system's input and output relationship; taking the mean of coherence across all frequency bands is used as an overall measure of similarity between two time series in this work.

The computation of entrainment was done for every speaker turn change. Figure 3 shows an example of the time frame that we extracted the entrainment measures; we computed entrainment measures between turn 2 and turn 3, turn 3 and turn 4, turn 4 and turn 5, and so on. Furthermore, since the turn length varies, we can not directly perform computation on two sequences of different lengths. An example is shown in Figure 3 where turn 2 is much shorter than turn 3. Our approach was to use a sliding window of the shorter sequence onto the longer sequence to compute square of correlation coefficient and mutual information multiple times. Following that, we computed the first, mean, and maximum across multiple segments of square coefficient and mutual information as an approach to overcome the issue with different length sequences. Coherence was adjusted by zero-padding the shorter sequence to match the length of the longer sequence. Zero-padding in the computation of coherence does not alter the result much since it does not change the frequency components of the signals. While more advanced techniques can be used to eliminate this problem, as a preliminary study, we wanted to examine whether the adopted measures can quantify prosodic entrainment. In total, there are 14 features describing prosodic entrainment between the couple for every turn change. A summary list of the entrainment features is provided below with the symbols used to denote them in the rest of the paper.

- Square of correlation coefficient (3) : r_{n1}, r_{μ}, r_{max}
- Mutual Information (3): $mi_{n1}, mi_{\mu}, mi_{max}$
- Mean Coherence (1): c_{μ}

2.3. Statistical Sequence Modeling Based Classification

We utilized a statistical Markov modeling technique to perform classification in this work. This is akin to the popular n-gram technique used to model word sequence statistics in language processing [14]. The first step was to quantize the continuous entrainment measures into discrete states using k-means clustering. Then two probabilistic n-gram symbol "language" models were built on these discrete states for the class labels *high positive* and *high negative*. This approach assumes that there exists some form of progression of these quantized entrainment states that is different in the two classes of interaction. The classifier scored the whole session using these two competing sequence models, and selected the class using maximum likelihood as shown in Equation 1. Assume $\mathcal{C}$ represents the class label and $\mathbf{X}$, sequence of input discrete states.

$$\hat{C} = \arg\max_{C \in \mathcal{C}} \{P_C(\mathbf{X} \mid C)\} \quad (1)$$

where $P_C(\mathbf{X} \mid C)$ is the score of $\mathbf{X}$ from the sequence model of label C.

There are several issues with the above approach. The total space of possible symbol items that a person can choose from is a vast set $\mathcal{S}$. Hence, it is most likely that the training data will be insufficient to express all the variability to be seen in human interactions. Further, the data from a single session is either so small that it does not represent the same variability of the training data, or it may not be a subset of the training data. These issues prompt us to interpolate:

$$\mathcal{C}' = (1-\lambda)\mathcal{C} + \lambda\mathcal{S} \quad (2)$$

In practice $\mathcal{S} = \epsilon(\text{BM}) + (1-\epsilon)\mathcal{C}_{all}$, where $\mathcal{C}_{all} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \ldots \cup \mathcal{C}_N$, and BM is a background model built from available generic data.

3. Experiment Setup and Results

Two different experiments were set up in this work:

- Hypothesis testing: measured entrainment is higher in positive interaction compared to negative interaction.
- Classification task: using entrainment features with symbol sequence modeling on recognizing couples having high rating in global positive vs. high rating in global negative attitude.

3.1. Hypothesis Testing

We hypothesize that our proposed entrainment measure is higher-valued in sessions when the speaker is rated with high-positive affect compared with high-negative affect. Student's *t-test* was performed on mean values of all 14 features computed per session, and the histogram of each feature was examined to make sure the features were distributed approximately normal. Table 1 shows a summary of the statistical testing results. The results indicate that out of all the 14 features, six entrainment features from pitch and two entrainment features from energy show significantly higher value for positive interaction.

The statistical testing result suggests that the entrainment in intonation can be well captured by these signal-driven synchronization measures. They show statistically significantly higher values, which agrees with what previous psychology studies suggest for people undergoing emotional interactions. However, the same phenomenon does not hold in most of the energy based measures. Two of the energy-based measures show the trend as pitch-based measures, while the rest seem to suggest the opposite. Several factors may be contributing to this result; the slope of energy contours may not be an adequate measure of entrainment in energy values. Another reason may be the fact that couples show more synchronization in the energy values in negative interactions (for example, by both yelling). Further analysis is required. The statistical testing, nevertheless, shows the encouraging outcome that several intonation-related entrainment features can indeed capture and quantify aspects of interaction synchrony.

3.2. Classification

In this section, we show that these signal-driven entrainment measures also provide a discriminant ability. We utilize the idea of symbol sequence modeling to dynamically model the flow of

Table 1: *Summary of Statistical Testing*

	Pitch Entrainment													
	r_{n1}		r_{μ}		r_{max}		mi_{n1}		mi_{μ}		mi_{max}		c_{μ}	
	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg
Mean	0.1061	0.0863	0.1001	0.0879	0.2013	0.2171	0.0519	0.0283	0.0487	0.0268	0.1111	0.1163	0.2294	0.1818
p-value	<**0.0001**		<**0.0001**		0.99		<**0.0001**		<**0.0001**		<**0.0001**		<**0.0001**	
	Energy Entrainment													
	r_{n1}		r_{μ}		r_{max}		mi_{n1}		mi_{μ}		mi_{max}		c_{μ}	
	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg
Mean	0.0690	0.0716	0.0646	0.0707	0.1348	0.1829	0.0284	0.0237	0.0190	0.0223	0.0680	0.1069	0.0284	0.0237
p-value	0.8153		0.99		≈1.0		**0.0035**		0.99		≈1.0		**0.0001**	

quantized prosodic entrainment measures of both speakers as discussed in Section 2; the SRILM toolbox [14] was used to build the models. Two different sets of features were tried: one with a feature vector of 28 dimension per analysis window - each of the 14 features describing the prosodic entrainment of one spouse. The other feature set used only the ones that passed the statistical testing given in Table 1. The k-means clustering was utilized to quantize the continuous entrainment values into discrete states, with $k = 9$ and $k = 7$ respectively, determined empirically. The accuracy is the number of correctly-recognized session labels divided by the total number of sessions. Leave-one-couple out cross-validation was used to test the model.

Table 2: *Summary of Classification Task*

Model	Accuracy (%)
Chance	50.00%
Full Set Features	71.00%
Reduced Set Features	**76.00%**

Table 2 shows a summary of the classification accuracy. We obtained the best classification accuracy at 76% - a 26% absolute improvement over chance. Using the reduced feature set the performance while not directly comparable, is similar to what is shown in [9]. The result is encouraging in two different aspects. First, we show that by using just 14 prosodic entrainment features with dynamic modeling, we can perform session-level attribute classification accurately. Second, since a dynamical model was used in this framework, an extension of this can be used to identify salient portions of the interaction and integrate with other cues to further improve prediction accuracy.

4. Conclusions and Future Work

In human interactions, people often exert mutual influence on each other's behavior. This phenomenon is often referred as interaction synchrony or entrainment. In this work, we extract several signal-driven prosodic (pitch and energy) entrainment measures to quantify this effect. A statistical testing in Section 3 shows that many of these features do behave as indicated in psychology studies, and these features also can be good predictors of session-level annotation on positive vs. negative affect as shown in Table 2.

There are two major future directions. The first is to extend the framework to examine other behavioral cues' entrainment phenomenon. This can include multimodal behaviors and more sophisticated methods in examining the *similarity*, meaning in-sync behaviors, between several time series evolving at different time scales. The second is to integrate this into generating turn-level profiles along with other verbal (lexical, syntactic, discourse, etc.) and nonverbal (facial expressions, body posture, etc.) cues to obtain a further improved classifier for predicting perceptual judgments of married couples' interaction. Quantification of these types of entrainment phenomena provides a mathematical approach for studying interpersonal communication, and it can further be seen as a bridge between psychologists and engineers to help bring objective insight into human interaction.

5. Acknowledgments

This research was supported in part by funds from the National Science Foundation, the Viterbi Research Innovation Fund, and the U.S Army.

6. References

[1] M. Kimura and I. Daibo, "Interactional synchrony in conversations about emotional episodes: A measurement by 'the between-participants pseudosynchrony experimental paradigm'," *Journal of Nonverbal Behavior*, vol. 30, pp. 115–126, 2006.

[2] A. R. McGarva and R. M. Warner, "Attraction and social coordination: Mutual entrainment of vocal activity rhymes," *Journal of Psycholinguistic Research*, vol. 32, no. 3, pp. 335–354, 2003.

[3] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL-08: HLT, Short Papers*, Columbus, Ohio, June 2008, pp. 169–172.

[4] M. J. Richardson, K. L. Marsh, and R. Schmit, "Effects of visual and verbal interaction on unintentional interpersonal coordination," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 1, pp. 62–79, 2005.

[5] J. S. Pardo, "On phonetic convergence during conversational interaction," *Journal of Acoustical Society of America*, vol. 119, pp. 2382–2393, 2006.

[6] R. M. Warner, D. Malloy, K. Schneider, R. Knoth, and B. Wilder, "Rhythmic organization of social interaction and observer ratings of positive affect and involvement," *Journal of Nonverbal Behavior*, vol. 11, no. 2, pp. 57–74, 1987.

[7] G. Margolin, P. Oliver, E. Gordis, H. O'Hearn, A. Medina, C. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical Child and Family Psychology Review*, vol. 1, no. 4, pp. 195–213, 1998.

[8] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. H. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *J. of Consulting and Clinical Psychology*, vol. 72, pp. 176–191, 2004.

[9] M. Black, A. Katsamanis, C.-C. Lee, A. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proceedings of Interspeech*, 2010.

[10] J. Jones and A. Christensen, *Couples interaction study: Social support interaction rating system*, University of California, Los Angeles, 1998. [Online]. Available: http://christensenresearch.psych.ucla.edu/

[11] C. Heavey, D. Gill, and A. Christensen, *Couples interaction rating system 2 (CIRS2)*, University of California, Los Angeles, 2002. [Online]. Available: http://christensenresearch.psych.ucla.edu/

[12] P. Moreno, C. Joerg, J.-M. van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," 1998.

[13] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[14] A. Stolcke, "SRILM: An extensible language modeling toolkit," 2002.