Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 518–539

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions☆,☆☆

Chi-Chun Lee[a,*], Athanasios Katsamanis[a], Matthew P. Black[a], Brian R. Baucom[b], Andrew Christensen[c], Panayiotis G. Georgiou[a], Shrikanth S. Narayanan[a,b]

[a] *Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, 3710 McClintock Ave., Los Angeles, CA 90089, USA*
[b] *Department of Psychology, University of Southern California (USC), 3620 McClintock Ave., Los Angeles, CA 90089, USA*
[c] *Department of Psychology, University of California, Los Angeles (UCLA), 1285 Framz Hall, Los Angeles, CA 90095, USA*

Received 27 October 2011; received in revised form 20 June 2012; accepted 22 June 2012
Available online 6 July 2012

Abstract

In human–human interactions, *entrainment* is a naturally occurring phenomenon that happens when interlocutors mutually adapt their behaviors through the course of an interaction. This mutual behavioral dependency has been at the center of psychological studies of human communication for decades. Quantitative descriptors of the degree of entrainment can provide psychologists an objective method to advance studies of human communication including in mental health domains. However, the subtle nature of the entrainment phenomenon makes it challenging for computing such an effect based on just human annotations. In this paper, we propose an unsupervised signal-derived approach within a principal component analysis framework for quantifying one aspect of entrainment in communication, namely, *vocal entrainment*. The proposed approach to quantify the degree of vocal entrainment involves measuring the similarity of specific vocal characteristics between the interlocutors in a dialog. These quantitative descriptors were analyzed using two psychology-inspired hypothesis tests to not only establish that these signal-derived measures carry meaningful information in interpersonal communication but also offer statistical evidence into aspects of behavioral dependency and associated affective states in marital conflictual interactions. Finally, affect recognition experiments were performed with the proposed vocal entrainment descriptors as features using a large database of real distressed married couples' interactions. An accuracy of 62.56% in differentiating between positive and negative affect was obtained using these entrainment measures with Factorial Hidden Markov Models lending further support that entrainment is an active component underlying affective processes in interactions.

Keywords: Entrainment; Interaction synchrony; Principal component analysis (PCA); Couple therapy; Behavioral signal processing (BSP); Factorial Hidden Markov Model; Affect recognition; Dyadic interaction

☆ This paper has been recommended for acceptance by Björn Schuller.
☆☆ The SAIL homepage is http://sail.usc.edu.
[*] Corresponding author. Tel.: +1 213 740 6432; fax: +1 213 740 4651.
E-mail addresses: chiclee@usc.edu, jeremylee.cc@gmail.com (C.-C. Lee).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.06.006

1. Introduction

Various psychological studies of interpersonal communication (e.g., Andersen and Andersen, 1984; Watt and VanLearn, 1996; Burgoon et al., 1995) conceptualize dyadic human–human interaction as an adaptive and interactive process. This process occurs spontaneously in the progression of human interactions serving multiple purposes including achieving efficiency, communicating interest and involvement in the interaction, and increasing mutual understanding through behavioral and affective mechanisms. This mutual coordination of behaviors both in timing and expressive forms between interlocutors is a phenomenon variously referred to as *entrainment*, accommodation, or interaction synchrony. A systematic and quantitative framework for assessing and tracking this notion of behavioral dependency between interlocutors in a conversation is essential in characterizing the overall quality and dynamic flow of human communication in general.

Moreover, numerous psychological theories of intimate relationships, such as couples' interactions, consider behavioral dependence to be a defining and core element of the theory. Support for this theoretical notion comes from a very large body of psychological and communication studies linking various forms of behaviorally dependent couples' interaction to individual well-being and relationship outcomes, e.g., divorce and domestic violence (Murphy and O'Farrell, 1997; Johnson and Jacob, 2000; Romano et al., 1992; Cordova et al., 1994; Jacobson et al., 1994). The quantitative study of the entrainment phenomenon, thus, becomes especially important because of not only its crucial role in analyzing human communication in general but also its utility in providing insights into the study of various mental distress and well-being conditions. In this paper, we consider the relation between the entrainment measured by our proposed computational framework and affective processes in distressed couple interactions.

The entrainment phenomenon has been extensively studied for the past twenty years in the psychology literature. While this body of work has offered many insights into human interaction dynamics, methods for assessing and quantifying the degree of behavioral entrainment have received little attention. Except for a few notable studies (Gottman et al., 2002; Boker and Laurenceau, 2006) in modeling couples' interactions, the computational techniques for quantifying entrainment have been largely based on log-linear models of highly reductionistic, categorical manual observation coding of behaviors. The manual observation coding of behaviors is often time-consuming and error-prone due to the subjective nature of the coding process (Kerig and Baucom, 2004).

Technological advances in capturing human behaviors with increasing ecological validity and mathematical capabilities to quantify interdependent processes have enabled new computational approaches mitigating some of the issues and limitations in the subjective observational coding process. Advances in speech processing and recognition as well as in image processing and computer vision have allowed engineers to understand aspects of not only intent but also human emotions and social signals in verbal and non-verbal form with objective signal processing methodologies. Notably, there has been tremendous progress in social signal processing (SSP) (Vinciarelli et al., 2009) wherein signal processing and machine learning methodologies have been developed and applied to understand complex human social behaviors, such as mutual gaze, head nods, and affective dynamics, in interpersonal interactions with potential applications geared toward natural human–machine interfaces. All these developments have enabled and facilitated the emerging field of behavioral signal processing (BSP) (Black et al., 2010; Lee et al., 2010; Rozgic et al., 2010). BSP requires acquiring behavioral data in an ecologically valid manner across laboratories to real-world settings, extracting and analyzing behavioral cues from measured data streams, developing models offering predictive and decision-making support to the appropriate (human) experts, and aiding in the solution of real world problems. One broad domain of BSP application relates to research and practice in mental health and well-being, e.g., married couples' therapy, addiction behaviors, depression, children with autism spectrum disorders. This paper offers an example of BSP in analyzing marital couple therapy data and providing feedbacks to human experts in family studies. Several previous works have also demonstrated the effectiveness of BSP in modeling distressed couples' behaviors with various types of automatically derived signals (Rozgić et al., 2011; Black et al., 2011, in press; Gibson et al., 2011; Katsamanis et al., 2011b; Georgiou et al., 2011).

In this work, our aim is to introduce a novel computational framework to quantify the degree of one specific channel of entrainment, *vocal entrainment*, using acoustic signals. The subtle nature of the vocal entrainment phenomenon makes its annotation difficult for human experts, which in turn hinders its computational measurement based on conventional supervised machine learning techniques. Relatively few studies in engineering have attempted to capture and quantify the degree of these subtle dependent behaviors through computing measures directly on the automatically extracted observable cues. Some key related studies focusing on quantifying entrainment in various communicative channels include the following: the investigation of mutual entrainment in vocal activity rhymes (McGarva and Warner, 2003); the analysis of high frequency word usage entrainment (Nenkova et al., 2008); the computation of entrainment of body movements (Richardson et al., 2005); the demonstration of phonetic convergence in conversation settings (Pardo, 2006); the existence of linguistic style similarity (Niederhoffer and Pennebaker, 2002).

The existence of vocal entrainment is well-established in psychology (Gregory et al., 1993, 1997; Gregory and Webster, 1996; Gregory and Hoyt, 1982; Chartrand and Bargh, 1999; Bernieri et al., 1988) and also has been demonstrated in engineering works (Levitan and Hirschberg, 2011; Lee et al., 2010). The schemes for quantifying the degree of prosodic entrainment for most of the studies rely on classical synchrony measures (e.g., Pearson correlation) on functionals of separate streams of acoustic features (e.g., mean pitch value per turn), computed across a speaker turn change. This approach of using classical synchrony measures has been widely adopted across a variety of research domains, e.g., econometrics, neuroscience, and physical coupled system studies. Through these various research works, there is a long list of classical synchrony measures and their variants available to quantify the interdependency between two simultaneously measured time series. An excellent review article by Dauwels et al., summarizes these measures for quantifying synchrony in electroencephalography (EEG) time series signals (Dauwels et al., 2010). These classical synchrony measures can be roughly categorized into the following types: linear correlation, nonlinear correlation, phase coherence, state-based synchrony, and information theoretic measures; they are all widely used and varyingly effective depending on the domain of studies.

However, there exist limitations in applying such a quantification approach to the study of vocal entrainment – mainly due to the complex nature of human–human conversations. Human conversation has a turn-taking structure, which challenges the requirement of simultaneously measured time series of certain similar behaviors, notably of vocal activity (visual behavior can co-occur, and be measured, although often one speaker tends to be holding the floor at any given time). Furthermore, the analysis window length for each time series, e.g., length of each speaking turn, varies across time (progressing through the dialog) and across variables (interlocutors in the dialog). Empirical evidence from psychological studies has also shown that multiple acoustic feature streams, often measured with classical synchrony measures, carry information about the entrainment process. These classical synchrony measures are often not directly applicable on multivariate set of acoustic features, e.g., those based on pitch, energy, and speech rate.

The proposed computational framework is a bottom-up approach utilizing automatically derived acoustic features to compute vocal entrainment levels. The formulation can be intuitively thought of as “computing how much people speak/sound like each other as they engage in conversation” captured by acoustic cues. We make a distinction between this intuition and the notion of “similarity in word usage” which is a different type of entrainment (i.e., lexical entrainment). Instead of computing synchrony measures on separate time series of acoustic features between interlocutors, we quantify the degree of vocal entrainment as the similarity between interlocutors’ vocal characteristic representation spaces. The vocal characteristic space is constructed based on a set of parametrized raw acoustic feature streams using principal component analysis (PCA).

We first introduced this notion of quantifying vocal entrainment in the framework of PCA with a single metric in our previous work (Lee et al., 2011b). There, we focused only on the directionality aspect of the entrainment process, e.g., how much speaker A in a dyad entrains toward speaker B and vice versa. The measure that we devised was computed based on the preserved variance as we project one set of acoustic parameters onto the PCA space of another. The derived measures were useful when applied to affect state recognition (Lee et al., 2011a,b). The method, however, suffers from robustness issues when the lengths of turns are significantly different; projecting a much longer-length turn onto a PCA space of shorter-length turns would result in a bias of “preserving more variance” as longer-length turns inherently tend to have larger variations.

We have extended our previous work in two folds by (1) introducing the use of symmetric similarity measures and (2) improving the similarity metric computational framework. The symmetric similarity values are computed based on angles between principal components (Krzanowski, 1979; Johannesmeyer, 1999) as a direct measure of similarity between two separate PCA spaces. This process results in values describing similarity that are symmetric, meaning that they have the same value for each interlocutor.

We propose to measure the degree of the directional entrainment by retaining the idea of projecting one interlocutor’s acoustic parameters in the PCA space of the other interlocutor. Then, instead of measuring the variance preserved, we compute Kullback–Leibler divergence as a metric of similarity, inspired by the work on quantifying similarity between datasets (Otey and Parthasarathy, 2005). Our proposed entrainment measures can be categorized into two types: symmetric and directional entrainment measures. The resulting measures from the proposed scheme consist of eight vocal entrainment values in total.

We analyze these entrainment measures on a database, referred to here as the Couple Therapy corpus, of real distressed married couples going through problem-solving spoken interactions as part of their participation in a randomized clinical trial of couple therapy. The corpus not only provides rich data for human communication studies but also represents an important realm for potentially beneficial contributions by behavioral signal processing. A recent review of more than three decades of marital interaction research indicates the importance of behavioral dependency for marriages (Eldridge and Baucom, 2010) and the proposed measures can provide a means for quantifying such constructs.

In this work, we carry out the analyses of the proposed computational measures of entrainment in three steps:

i. **Verification**: verifying that the proposed signal-derived measures capture psychologically valid notions of entrainment.
ii. **Analysis**: analyzing the relationship between the vocal entrainment phenomenon and the interacting spouses' affective states.
iii. **Application**: applying vocal entrainment measures as features in an affective state recognition task.

The analysis carried out in the verification step is important in order to verify that such a signal-derived analytic is appropriate for characterizing and capturing the notion of inherent behavioral dependencies conceptualized in psychological studies of interpersonal interactions. The assumption is that if there exists a natural cohesiveness in human–human conversations, the proposed entrainment measures should be expected to result in higher values when computed in a dialog between an in-conversation dyad compared to randomly generated dialogs between not-in-conversation dyads. The results of this evaluation indicate that the proposed measures are indeed higher in real conversation compared to artificial conversations, making their use a viable approach to quantitatively describe the phenomenon of vocal entrainment.

The second evaluation focuses on exploring the usefulness of this computational tool in providing psychologically significant insights about the relationship between affective states and the varying degree of vocal entrainment for the spouses in the couples' interactions. Results from our analysis indicate that most of the vocal entrainment values show significantly higher value in interactions where the spouse was behaviorally coded as having high positive affect compared to high negative affect. This analysis provides some of the first empirical evidence that vocal entrainment offers an indication of a positive interacting process during couple interactions. It is also consistent with other psychological studies documenting the positive effects of entrainment in other interaction contexts (Kimura and Daibo, 2006; Verhofstadt et al., 2008).

The third analysis is to demonstrate that these measures can also be utilized as features in human behavior classification tasks. Affect recognition of the spouse (positive affect vs. negative affect) is used as an exemplary application with the proposed analytics as features. In our previous work (Lee et al., 2011a), we performed the same affect recognition task with the same database using a multiple instance learning framework. We utilize a temporal modeling technique (Factorial Hidden Markov Model) in this work for this purpose, and we obtain a classification accuracy of 62.86%, which is a 8.93% absolute (16.56% relative) improvement over our previous result. This result lends further support to the observation that the behavioral dependencies underlying affective processes are reflected in the proposed measures. It should be noted that this experiment is not the primary focus of the paper but is presented to show the potential of the entrainment measures to reflect behavioral dependencies in affective dynamics.

The rest of the paper is organized as follows: Section 2 describes the Couples Therapy database; Section 3 describes our PCA-based vocal entrainment quantification scheme; Section 4 presents the two approaches in analyzing the signal-derived entrainment measures; Section 5 describes the affective state classification framework and experimental results; and Section 6 presents conclusions and ideas for future works.

2. The Couple Therapy corpus

The Couple Therapy corpus originated from a collaborative project between the psychology departments of the University of California, Los Angeles and the University of Washington (Christensen et al., 2004). This collaborative

Table 1
The complete list of 32 behavioral codes: 19 from SSIRS (Jones and Christensen, 1998) and 13 from CIRS (Heavey et al., 2002).

Social Support Interaction Rating System (SSIRS)	Couples Interaction Rating System (CIRS)
Global positive affect, global negative affect, use of humor, sadness, anger/frustration, belligerence/domineering, contempt/disgust, tension/anxiety, defensiveness, affection, satisfaction, solicits partner's suggestions, instrumental support offered, emotional support offered, submissive or dominant, topic is a relationship issue, topic is a personal issue, discussion about husband, discussion about wife	Acceptance of other, blame, responsibility for self, solicits partner's perspective, states external origins, discussion, clearly defines problem, offers solutions, negotiates, makes agreements, pressures for change, withdraws, avoidance

project resulted in the largest longitudinal, randomized, behaviorally based couple therapy clinical trial to date. A total of 134 seriously and chronically distressed couples participated in the study, and they received up to 26 couple therapy sessions over the course of a year. As part of their participation in the study, each couple engaged in problem-solving interactions where one of the spouses picked one distinct topic related to a serious problem in their relationship to discuss, and they tried to resolve it. Each topic of a problem-solving interaction lasted about 10 min. Each 10 min interaction was audio–video recorded for observation analysis, and each spouse was coded separately by trained human annotators.

The Couple Therapy corpus consists of audio–video recordings, manual transcriptions, and behavioral codings of each couples' problem-solving interactions. The interactions that we consider were recorded at three different points in time: pre-therapy, the 26-week assessment, and the two-year post-therapy assessment. The recorded audio–video data includes a split-screen video and a single channel far-field audio recorded from the video camera microphone. The recording conditions, e.g., microphone and camera positions, background noise level, and lighting conditions, varied from session to session. Manual word transcriptions were carried out to aid the analysis of couples' language use. The resulting word-level transcriptions were chronological, and the speaker identity was explicitly labeled in the transcript. The transcriptions, however, did not have explicit timing information on speakers' turn-taking.

For each interaction session, multiple evaluators (ranging from 2 to 12 evaluators) rated each spouse with 32 different behavioral codes based on two established coding manuals, Social Support Interaction Rating System (SSIRS) (Jones and Christensen, 1998) and Couples Interaction Rating System (CIRS) (Heavey et al., 2002). The SSIRS consists of 19 codes assessing the emotional content and the topic of the conversations corresponding to four different categories: affect, dominance/submission, features of the interaction, and topic definition. The CIRS consists of 13 codes which were specifically designed for coding problem-solving discussions. Table 1 lists the complete list of the 32 codes for SSIRS and CIRS. Each code was evaluated on an integer scale from 1 (none/not at all) to 9 (a lot). All evaluators went through a training process to standardize the coding process. They were instructed to make their judgments after observing the whole interaction session. For each problem-solving interaction, each spouse was rated with one global value for each of the 32 behavioral codes. Each spouse selected a topic that he/she wanted to discuss in the problem-solving interactions, and then the other spouse selected a different topic for the discussion in another interaction session. The original study aimed at recording 804 sessions (134 couples × 3 points in time × 2 topics per couple). After eliminating sessions where either codes were missing or spouses withdrew from the study, the remaining 569 problem-solving interactions constitute the Couple Therapy corpus, totaling 95.8 h of data with 117 unique couples.

While it is desirable to utilize the entire corpus, not all of the sessions in the Couple Therapy corpus are suitable for automatic analysis due to the varying noise conditions across different sessions resulting in unreliable estimates of acoustic features. We had to identify a subset of sessions out of the original 569 sessions, denoted here as $\textbf{Dataset}_{\text{qual}}$, that were deemed to be of suitable audio quality for the analyses considered. The subset consists of 372 sessions in total (the detailed selection criterion and preprocessing steps are described in Section 2.1). The dataset, $\textbf{Dataset}_{\text{qual}}$, of 372 sessions was used in verifying the validity of the proposed vocal entrainment measures (Section 4.1). To carry out the analysis of unambiguously marked affective states of the spouses (Sections 4.2 and 5), we selected a subset out of the 372 sessions, denoted here as $\textbf{Dataset}_{\text{emo}}$, based on the extremes of the "Global Positive" and "Global Negative" ratings of each spouse. Details of this selection are described in Section 2.2.

2.1. ***Dataset***$_{qual}$: *preprocessing*

The first pre-processing step that we carried out was to identify a subset of the 569 sessions with sufficient quality that could be robustly analyzed with automatically derived acoustic features. This was done with two criteria: average signal-to-noise ratio (SNR) estimation based on voice activity detection (VAD) (Ghosh et al., 2010) and speech-text alignment algorithm using SailAlign (Katsamanis et al., 2011a). VAD was designed to detect non-speech segments larger than 300 ms. SNR was then estimated as follows:

$$\mathrm{SNR}\,(\mathrm{dB}) = 10\log_{10}\frac{(1/|i \in S|)\sum_{i\in S}A_i^2}{(1/|i \notin S|)\sum_{i\notin S}A_i^2}, \tag{1}$$

where $\{A_i\} \in S$ is the set of amplitudes resulting from the VAD-detected speech regions and $\{A_i\} \notin S$ is the set of amplitude outputs in the non-speech regions, again based on the VAD. We empirically chose 5 dB SNR as the cutoff for determining which sessions to include for automatic analysis. This procedure and the chosen SNR criterion eliminated 154 sessions from the current study. This resulted in a total of 415 sessions out of the 569 sessions based on SNR criterion. As is common in dyadic conversation studies, the spoken analysis unit adopted is the speaking turn. The Couple Therapy corpus does not contain explicit timing of each speaking turn. Instead of manually segmenting the speaking turn for each spouse for all sessions, we used a "hybrid" manual/automatic speaker segmentation, given the availability of manual word-level transcriptions. We implemented a recursive Automatic Speech Recognition (ASR)-based procedure to align the transcription with the audio data using an open source tool, SailAlign.[1] As a result of this speech-text alignment, we obtained timing information on each alignment along with approximate speaking turn segmentation. We used these turn estimates (referred to as *speaking turns*, or just *turns* in the rest of the paper) as an approximation of the actual speaking turns for each spouse in each interaction session. Due to the nature of the alignment process and the non-ideal nature of audio quality, not every word in the transcription could be reliably aligned. We further eliminated sessions where the algorithm failed to align 55% or more of the words in the transcripts. The 55% cut-off eliminated another 43 sessions out of 415 sessions. The percentage was chosen as a trade-off between retaining a greater number of sessions for the analyses and eliminating those sessions that were of poor audio quality and had poor speaker segmentation results. This resulted in a final dataset, **Dataset**$_{\mathrm{qual}}$, of 372 interaction sessions for the current study totaling 62.8 h of data with 104 unique couples; the same dataset was used in past studies on the same corpus (Black et al., in press).

2.2. ***Dataset***$_{emo}$: *positive vs. negative affect*

There are numerous psychology studies (Jacobson et al., 1994; Verhofstadt et al., 2008; Gottman, 1993) describing and indicating various degrees of relations between affective states and behavioral dependencies in couples' interactions. In this work, we investigate the relation between the proposed measures, that reflect a type of behavior dependency (vocal entrainment), and the affective states. In particular, we study the relationship between spouses' affective states and their associated degree of vocal entrainment. For this purpose, we defined two emotional classes, *positive* and *negative*, of each spouse, with respect to the rating of the two behavioral codes, namely "Global Positive" and "Global Negative", derived from the SSIRS coding manuals. The mean inter-evaluator agreements, computed using intraclass correlation (Shrout and Fleiss, 1979), of "Global Positive" and "Global Negative" are 0.831 and 0.867 respectively, indicating a reasonably high agreement between evaluators for these two behavioral codes. The following are the coding instructions quoted directly from the SSIRS manual for both codes:

> "[**Global Positive** An overall rating of the positive affect the target spouse showed during the interaction. Examples of positive behavior include overt expressions of warmth, support, acceptance, affection, positive negotiation, and compromise. Positivity can also be expressed through facial and bodily expressions, such as smiling and looking happy, talking easily, looking comfortable and relaxed, and showing interest in the conversation.]"
> "[**Global Negative** An overall rating of the negative affect the target spouse shows during the interaction. Examples of negative behavior include overt expressions of rejection, defensiveness, blaming, and anger. It can also include

[1] http://sail.usc.edu/software/SailAlign/.

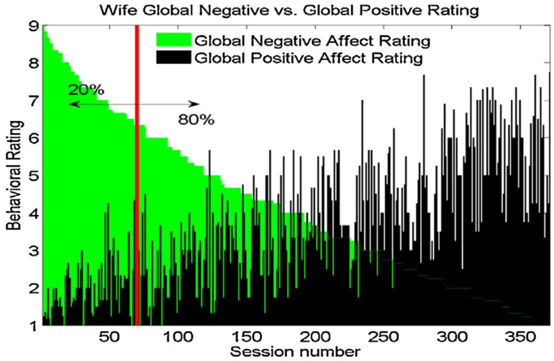


Fig. 1. ‘Global Negative Rating’ for wife in the Couple Therapy corpus in a descending order with the corresponding “Global Positive Affect” rating in the same session.

facial and bodily expressions of negativity such as scowling, crying, crossing arms, turning away from the spouse, or showing a lack of interest in the conversation. Also factor in degree of negativity based on severity (e.g., a higher score for contempt than apathy).]”

Since the coding manual for the two behavioral codes instructs annotators to treat each code as independent of each other, a high rating of “Global Negative” does not guarantee a low rating of “Global Positive” (see Fig. 1). In order to mitigate the ambiguity in defining *positive* affect and *negative* affect, the specific subset of database that we used in this work comes from the extreme ratings of these affect codings. We chose the ratings for which any one of the spouses was rated in the top 20% for either of the codes (high rating of global positive and high rating of global negative) to serve as the prototypical (unambiguous) *positive* affect and *negative* affect of a spouse; this subset of data is denoted here as **Dataset**$_{\text{emo}}$. The 20% threshold was inspired from previous behavioral studies (Jurafsky et al., 2009; Ranganath et al., 2009) and was also used in our previous research work as a starting point to study the extreme behaviors in couples' interactions (Black et al., in press).

The spouses that we defined as having *positive* affect state had a mean rating score of 7.00, on the “Global Positive”, which was much higher than its mean rating score, 2.15, on the “Global Negative”. The spouses that we defined as having *negative* affect state had a mean rating of 6.25 on the “Global Negative” and a much lower rating, 2.08, on the “Global Positive”. None of the spouses in this dataset had a mean evaluator score of both “Global Positive” and “Global Negative” to appear in the top 20% (see Fig. 2). This dataset, **Dataset**$_{\text{emo}}$, was used to perform our analysis in Section 4.2 and affect recognition in Section 5. The resulting dataset consists of interaction sessions from 81 unique couples with 280 ratings: 140 high-positive ratings (70 of husbands, 70 of wives) and 140 high-negative ratings (70 of husbands, 70 of wives).

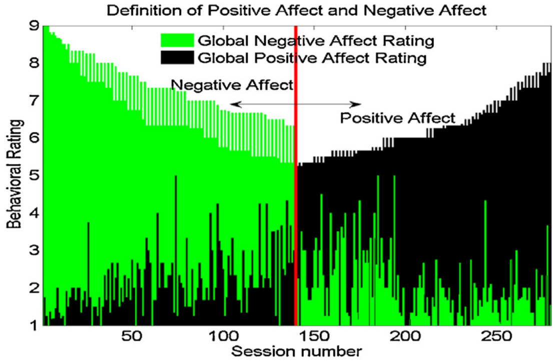


Fig. 2. **Dataset**$_{\text{emo}}$: bar plot of the original ratings of “Global Positive Affect” and “Global Negative Affect” codes for the sessions that we consider belonging in the positive affect and negative affect categories respectively.

2.3. Audio feature extraction

After the pre-processing steps, we extracted various speech-related features from the 372 sessions detailed in our previous work (Black et al., in press). We utilized the following subset of the acoustic features, namely mel-frequency cepstrum coefficients (MFCCs), pitch (f_0), intensity (int), and speech rate, in this work. The 15 MFCCs were computed using 25 ms windows and 10 ms shift with the OpenSMILE toolbox (Eyben et al., 2010). MFCCs were normalized using cepstral mean subtraction as follows:

$$\mathrm{MFCC}_n[i] = \mathrm{MFCC}[i] - \mu_{\mathrm{MFCC}[i]}, \quad i = 0, \ldots, 14, \tag{2}$$

where the $\mu_{\mathrm{MFCC}[i]}$ values correspond to the mean MFCC of the ith coefficient of the speaker across the whole session.

Fundamental frequency and intensity were both extracted using an autocorrelation-based method as implemented in the Praat toolbox (Boersma, 2001). Intensity frame values at each frame, n, were normalized in the following way:

$$\mathrm{int}_n = \frac{\mathrm{int}}{\mu_{\mathrm{int}}}, \tag{3}$$

where the μ_{int} values correspond to the mean intensity of speech during the active speaker regions, computed across the whole session.

We further carried out several postprocessing procedures to ensure that the raw pitch extraction was reasonably accurate. We attempted to mitigate pitch doubling and halving by passing the raw pitch signals through an algorithm detecting large differences in f_0 values in consecutive frames. The pitch values were forced to be zeros at regions where the VAD algorithm detects a non-speech portion. We interpolated over unvoiced regions with duration less than 300 ms using piecewise-cubic Hermite interpolation. Finally, a median filter of length five was applied to eliminate spurious noise. f_0 values were normalized as follows:

$$\overline{f}_{0_{\log}} = \log_2\left(\frac{f_0}{\mu_{f_0}}\right) \tag{4}$$

where μ_{f_0} values were computed across the whole session using the speaker segmentation results.

Finally, we computed the mean syllabic speaking rate for each aligned word directly from the automatic word alignment results with the help of a syllabified pronunciation dictionary.[2]

3. Signal-derived vocal entrainment quantification

Our proposed signal-derived vocal entrainment quantification is based on the core idea of computing similarity measures between the vocal characteristic spaces (represented by the corresponding PCA spaces) of interlocutors. The framework computes vocal entrainment values at the level of speaking turns for each interlocutor in the interaction. It involves two steps. The first is to obtain an adequate set of acoustic feature parameters to represent the speaking characteristics. The second is to represent these acoustic parameters in the PCA space based on which we compute various similarity measures. In this section, we will first describe four general similarity measures, given two PCA representations on two sets of time series observations. Then we will discuss the parametrization of the acoustic features to serve as descriptors of vocal characteristics, and, lastly, we will describe how to apply the method to extract a total of eight features indicating the degree of vocal entrainment for each spouse in couples' interactions.

3.1. PCA-based similarity measures

Principal component analysis (PCA) is a well-known statistical method for analyzing multivariate time series. PCA performs an orthogonal transformation of a set of observation variables onto a set of uncorrelated variables called principal components. The first component accounts for the maximum variance of the observed data, and each

[2] http://www.haskins.yale.edu/tada_download/index.php.

succeeding component explains the highest possible variance with the constraint that it be orthogonal to the preceding component. The mathematical formulation of PCA follows:

$$\mathrm{Y}^{\mathrm{T}} = \mathrm{X}^{\mathrm{T}}\mathbf{W} = \mathbf{\Sigma}\mathrm{V}^{\mathrm{T}} \tag{5}$$

where $\mathbf{X}$ is the zero-mean data matrix, $\mathbf{W}$ is the matrix of eigenvectors of $\mathbf{XX}^T$, $\mathbf{Y}$ is the representation of $\mathbf{X}$ after PCA, $\mathbf{V}$ is the matrix of eigenvectors of $\mathbf{X}^T\mathbf{X}$, and $\mathbf{\Sigma}$ is a diagonal matrix containing values of variance associated with each principal component.

Assume we are given two sets of multivariate time series observations (e.g., from two individuals in a dyadic interaction), $\mathbf{X}_1$ and $\mathbf{X}_2$, each comprising the same n time series signals but that can be of different lengths. We can then respectively compute the two sets of principal components, $\mathbf{W}_1$ and $\mathbf{W}_2$, and the two associated diagonal variance matrices, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. We propose two types of similarity measures based on these representations:

- **Symmetric**: similarity between the two PCA representations, $\mathbf{W}_1$ and $\mathbf{W}_2$.
- **Directional**: similarity when representing one set of observations, e.g., $\mathbf{X}_1$, in the other PCA space, e.g., $\mathbf{W}_2$.

3.1.1. Symmetric similarity measures

From (5), PCA is essentially a process of rotating the original data matrix to a new coordinate system with the optimization criterion of maximizing explained variances. The general procedure for computing symmetric similarity measures with PCA is listed below:

1. Obtain principal components for each time series separately:

$$\begin{aligned}\mathrm{Y}_{\mathbf{1}} &= \mathrm{X}_{\mathbf{1}}{}^{\mathrm{T}}\mathrm{W}_{\mathbf{1}}\\ \mathrm{Y}_{\mathbf{2}} &= \mathrm{X}_{\mathbf{2}}{}^{\mathrm{T}}\mathrm{W}_{\mathbf{2}}\end{aligned}.$$

2. Retain k components of each time series,

$$k = \max(k_1, k_2)$$

$k_1 < n$, $k_2 < n$, each explaining a fixed fraction (95% here) of variance.
3. Compute measures of similarity based on angles between the k reduced set of components (Eqs. (6) and (7)).

The first similarity value is proposed in the work of Krzanowski (1979):

$$ssim_u(\mathrm{X}_{\mathbf{1}}, \mathrm{X}_{\mathbf{2}}) = trace(\mathrm{W}_{\mathbf{1}\mathrm{L}}^{\mathrm{T}}\mathrm{W}_{\mathbf{2}\mathrm{L}}\mathrm{W}_{\mathbf{2}\mathrm{L}}^{\mathrm{T}}\mathrm{W}_{\mathbf{1}\mathrm{L}}) = \sum_{i=1}^{k}\sum_{j=1}^{k}\cos^2(\theta_{ij}), \tag{6}$$

where θ_{ij} is the angle between the ith principal component of $\mathbf{X}_1$ and jth principal component of $\mathbf{X}_2$. $\mathbf{W}_{1L}$ and $\mathbf{W}_{2L}$ contain the reduced number of principal components, i.e., k components. Consequently, $ssim_u(\mathbf{X}_1, \mathbf{X}_2)$ ranges between 1 and k.

Another similarity measure proposed by Johannesmeyer (1999) is an extension to the previous measure by weighting the angles with their corresponding variance. The measure in (6) can be thought of as an unweighted symmetric measure and the following is its weighted symmetric counterpart:

$$ssim_w(\mathrm{X}_{\mathbf{1}}, \mathrm{X}_{\mathbf{2}}) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(\lambda_{X_{1,i}}\lambda_{X_{2,j}}\cos^2(\theta_{ij}))}{\sum_{i=1}^{k}\lambda_{X_{1,i}}\lambda_{X_{2,i}}}, \tag{7}$$

where $\lambda_{X_{1,i}}, \lambda_{X_{2,j}}$ are the diagonal elements of $\mathbf{\Sigma_1}, \mathbf{\Sigma}_2$.

The interpretation of these two measures, namely $ssim_u$ and $ssim_w$, is based on the assumption that if two sets of observations are similar to each other, the angles between their corresponding principal components will be closer to zero; hence, the corresponding sum of $\cos^2(\theta_{ij})$, $i = 1, \ldots, k$, will be larger. Note that these two measures are symmetric (i.e., $ssim(X_1, X_2) = ssim(X_2, X_1)$).

3.1.2. Directional similarity measures

The *entrainment* process inherently carries notions of directionality – a given process can be entraining *toward* or getting entrained *from* another interacting process or reflect a combination of both. We propose to quantify each of these directionality aspects in the same PCA framework. The idea is to compute similarity when we represent one time series in the PCA space of another time series.

For each process, $\mathbf{X}_1$, there can be two directions of entrainment. We can compute the degree that it is entraining *toward* the other process, $\mathbf{X}_2$, denoted as $dsim_{to}^{X_1}$, as the similarity between $\mathbf{X}_1$ and $\mathbf{X}_2$ when representing $\mathbf{X}_1$ in the PCA space of $\mathbf{X}_2$. The degree that it is getting entrained *from* another process, $dsim_{fr}^{X_1}$, is computed as the similarity between $\mathbf{X}_1$ and $\mathbf{X}_2$ when representing $\mathbf{X}_2$ in the PCA space of $\mathbf{X}_1$.

We first compute four normalized variance vectors $\{\vec{\boldsymbol{\lambda}}_{1to2}^{n}, \vec{\boldsymbol{\lambda}}_{2}^{n}, \vec{\boldsymbol{\lambda}}_{2to1}^{n}, \vec{\boldsymbol{\lambda}}_{1}^{n}\}$. The first two, namely $\vec{\boldsymbol{\lambda}}_{1to2}^{n}, \vec{\boldsymbol{\lambda}}_{2}^{n}$, are used for computing $dsim_{to}^{X_1}$, while $\vec{\boldsymbol{\lambda}}_{2to1}^{n}, \vec{\boldsymbol{\lambda}}_{1}^{n}$ are used for computing $dsim_{fr}^{X_1}$. Computation proceeds as follows:

- **Compute $\vec{\boldsymbol{\lambda}}_{1to2}^{n}$ and $\vec{\boldsymbol{\lambda}}_{2}^{n}$**
  1. Project $\mathbf{X}_1$ using $\mathbf{W}_2$: $\mathbf{Y}_{1to2} = \mathbf{X}_1{}^T\ \mathbf{W}_2$.
  2. Compute variance vector: $\vec{\boldsymbol{\lambda}}_{1to2} = var(\mathbf{Y}_{1to2})$.
  3. Normalize variance vector: $\vec{\boldsymbol{\lambda}}_{1to2}^{n} = \vec{\boldsymbol{\lambda}}_{1to2}/\sum_i \boldsymbol{\lambda}_{1to2,i}$.
  4. Project $\mathbf{X}_2$ using $\mathbf{W}_2$: $\mathbf{Y}_2 = \mathbf{X}_2{}^T\ \mathbf{W}_2$.
  5. Compute variance vector: $\vec{\boldsymbol{\lambda}}_2 = var(\mathbf{Y}_2)$.
  6. Normalize variance vector: $\vec{\boldsymbol{\lambda}}_2^{n} = \vec{\boldsymbol{\lambda}}_2/\sum_i \boldsymbol{\lambda}_{2,i}$.
- **Compute $\vec{\boldsymbol{\lambda}}_{2to1}^{n}$ and $\vec{\boldsymbol{\lambda}}_{1}^{n}$**
  1. Project $\mathbf{X}_2$ using $\mathbf{W}_1$: $\mathbf{Y}_{2to1} = \mathbf{X}_2{}^T\ \mathbf{W}_1$.
  2. Compute variance vector: $\vec{\boldsymbol{\lambda}}_{2to1} = var(\mathbf{Y}_{2to1})$.
  3. Normalize variance vector: $\vec{\boldsymbol{\lambda}}_{2to1}^{n} = \vec{\boldsymbol{\lambda}}_{2to1}/\sum_i \boldsymbol{\lambda}_{2to1,i}$.
  4. Project $\mathbf{X}_1$ using $\mathbf{W}_1$: $\mathbf{Y}_1 = \mathbf{X}_1{}^T\ \mathbf{W}_1$
  5. Compute variance vector: $\vec{\boldsymbol{\lambda}}_1 = var(\mathbf{Y}_1)$.
  6. Normalize variance vector: $\vec{\boldsymbol{\lambda}}_1^{n} = \vec{\boldsymbol{\lambda}}_1/\sum_i \boldsymbol{\lambda}_{1,i}$.

Each normalized variance vector characterizes the proportion of the variance explained as the time series projections represented in each of the principal components. If we retain all components, they sum to one. We can, then, consider them as random variables, V_2, V_{1to2}, V_1, V_{2to1}, with probability mass distribution based on each element in the normalized variance vectors described in the following:

$$\begin{aligned} P_2 &= P(V_2 = i) = \boldsymbol{\lambda}_{2,i}^{n} \\ P_{1to2} &= P(V_{1to2} = i) = \boldsymbol{\lambda}_{1to2,i}^{n} \end{aligned} \tag{8}$$

$$\begin{aligned} P_1 &= P(V_1 = i) = \boldsymbol{\lambda}_{1i}^{n} \\ P_{2to1} &= P(V_{2to1} = i) = \boldsymbol{\lambda}_{2to1,i}^{n} \end{aligned} \tag{9}$$

The similarity between variance vectors can be thought of as similarity between two probability distributions. We employ the use of symmetric Kullback–Leibler Divergence (KLD) to quantify the difference (hence, the similarity) between two probability distributions. If two sets of observations are more similar to each other, the symmetric KLD will result in a lower numerical value. This method of quantifying similarity is inspired from the work of Otey and Parthasarathy (2005), but Otey's method was used on two completely different datasets and does not possess the same notion as our proposed method, i.e., projection of one time series onto another:

$$dsim_{to}^{X_1} = \frac{1}{2}(D_{KL}(P_2 \| P_{1to2}) + D_{KL}(P_{1to2} \| P_2)) \tag{10}$$

$$dsim_{fr}^{X_1} = \frac{1}{2}(D_{KL}(P_1 \| P_{2to1}) + D_{KL}(P_{2to1} \| P_1)) \tag{11}$$

$$D_{KL}(P\|Q)=\sum_i P(i)\,\log\frac{P(i)}{Q(i)} \tag{12}$$

where $dsim_{to}^{X_1}$, $dsim_{fr}^{X_1}$ represent how much the time series observation, $\mathbf{X}_1$, is entraining *toward*, and getting entrained *from*, its interacting process, $\mathbf{X}_2$, respectively.

The same procedure needs to be carried out to calculate $dsim_{to}^{X_2}$ and $dsim_{fr}^{X_2}$ to represent how much the time series observation, $\mathbf{X}_2$, is entraining *toward*, and getting entrained *from*, its interacting process, i.e., $\mathbf{X}_1$. Note that while this computation would result in the same numerical values for $dsim_{to}^{X_2}$ and $dsim_{fr}^{X_1}$ (also $dsim_{to}^{X_1}$ and $dsim_{fr}^{X_2}$), the underlying interpretation on the directionality of similarity is different. It can be intuitively interpreted as a vector representation composed of two components: direction and magnitude. While the numerical values, i.e., *magnitudes*, are the same, the *directions*, toward vs. from, of the entrainment process are different.

3.2. Representative vocal features

Vocal entrainment can be intuitively thought of as a phenomenon that represents "how people *sound alike* when they speak to each other". In order to quantify the degree of entrainment using the method proposed in Section 3.1, we need to capture this speaking style with acoustic vocal features. We utilized the four acoustic feature streams, pitch, intensity, speech rate, and MFCCs, as described in Section 2.3. Prosodic cues, e.g., pitch, intensity and speech rate, can often be used to describe more explicit speaking style characteristics, e.g., intonation patterns, loudness, and rate of speaking. MFCCs, on the other hand, capture general spectral properties, which apart from speaker identity (Reynolds, 1994) and phonemic content may also reflect characteristics of the articulation relating to the speaker's emotional state (Kwon et al., 2003).

We carried out further parametrization of these features because the extracted 10 ms frame-by-frame values are too detailed, considering the long-term nature of entrainment. To adequately capture the inherent dynamic variations of the acoustic features in characterizing speaking styles at the turn level, we performed the parametrization of raw acoustic features at the word level using statistical functionals and contour fitting methods. Both the contour-based and statistical functional methods of analysis are common for pitch and intensity. Contour-based methods can capture the temporal variation while statistical functional methods are used to describe the overall statistical properties. We decided to parametrize these two feature streams using two methods. We used least-squares to fit a third-order polynomial to pitch values (Eq. (13)) and a first-order polynomial (Eq. (14)) to intensity values at the word level. This method of polynomial-based parametrization is the same as in our previous work on analyzing entrainment of individual prosodic feature streams (Lee et al., 2010):

$$\overline{f}_{0_{\log}}(t)=\alpha_3t^3+\alpha_2t^2+\alpha_1t+\alpha_0 \tag{13}$$

$$\mathrm{int}_n(t)=\beta_1t+\beta_0 \tag{14}$$

To further obtain information on the statistical properties, we computed mean (μf_{0w}, $\mu\mathrm{int}_w$) and variance ($\sigma^2 f_{0w}$, $\sigma^2\mathrm{int}_w$) of both pitch and intensity at the word level. We only used α_3, α_2, α_1 for pitch values and β_1 for intensity values to characterize the pattern of pitch and intensity dynamics; intercept terms under the least square contour fitting method can be thought of as capturing approximately the same information as the mean values. The speech rate feature is a one-dimensional feature and is based on the average syllable rate. We computed mean and variance for 13 MFCCs, resulting in 26 MFCC-related parameters per word. The following is the final list of parameters of acoustic features calculated for each word:

- Pitch parameters (5): $[\alpha_1, \alpha_2, \alpha_3, \mu f_{0w}, \sigma^2 f_{0w}]$.
- Intensity parameters (3): $[\beta_1, \mu\mathrm{int}_w, \sigma^2\mathrm{int}_w]$.
- Speech rate (1): $[\mathrm{sylb}_\mu]$.
- MFCCs (26): $[\mu\mathrm{MFCC}_w[i], \sigma^2\mathrm{MFCC}_w[i]]$ ($i=0, \ldots, 12$)

This parametrization resulted in a 35-dimensional vocal characteristic parameter vector derived from raw acoustic low level descriptors per word. Vocal quality features (e.g., shimmer, jitter, and harmonic-to-noise ratio) also convey information about vocal characteristics; however, they are computed based on the estimated fundamental frequencies.

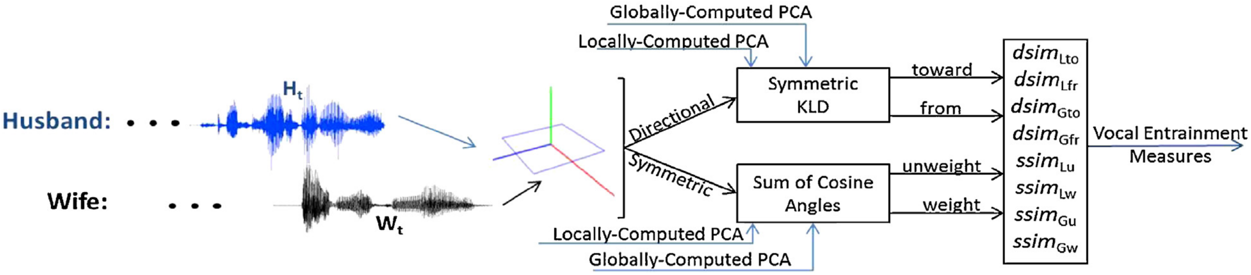


Fig. 3. Example of computing measures quantifying vocal entrainment for turns H$_t$ in a dialog.

Since it is a challenging task to robustly estimate these features in noisy conditions, we did not include them in this present work.

3.3. Vocal entrainment measures in dialogs

Section 3.1 describes a general framework to compute similarity between two multivariate time series using PCA, and Section 3.2 describes the acoustic parameters used to represent vocal characteristics. We will describe in this section the complete process of quantifying vocal entrainment in human–human conversation.

There are two variants to each of the similarity values proposed in Section 3.1 depending on the manner in which the PCA vocal characteristic space is computed. Since PCA is meant to represent vocal characteristics, we computed PCA for each speaker both at each speaking turn level (*locally computed PCA*) and at each talker level (*globally computed PCA*). The locally computed vocal characteristic space was specified by performing PCA for every single turn, and, if the turn did not have at least 35 words, it was merged with the nearby turns to ensure a unique representation of PCA given that the dimension of acoustic parameters is 35. The globally computed vocal characteristic space was specified by first aggregating all the turns of a single subject from all the sessions he/she participated in to perform PCA. The locally computed PCA captures the moment-by-moment changes in the vocal characteristics of an individual speaker, and the globally computed PCA captures an individual's overall vocal properties. In the case of the locally computed PCA, the computation procedure listed in Section 3.1 can be directly implemented resulting in four vocal entrainment values for each speaker at each speaking turn, denoted as $dsim_{L_{to}}$, $dsim_{L_{fr}}$, $ssim_{L_u}$, $ssim_{L_w}$.

In the case of globally computed PCA, we need to substitute the locally computed turn-level PCA components with the globally computed subject-level PCA. All of the various projections and variance vector computations remain the same using the *turn-level* acoustic parameters. Because the projections were done using turn-level acoustic parameters, the resulting computation could be interpreted as finding similarities of the local representation derived from the global vocal characteristics between interlocutors. This method gives four additional vocal entrainment values, denoted as $dsim_{G_{to}}$, $dsim_{G_{fr}}$, $ssim_{G_u}$, $ssim_{G_w}$.

The complete procedure of computing vocal entrainment is illustrated in the example depicted in Fig. 3. For each speaker (husband, wife) and each of their speaking turns (H_t, W_t) in the Couple Therapy database, we compute *eight* similarity measures (Section 3.1) between (H_t and W_t) using z-normalized acoustic parameters (Section 3.2) as multivariate time series observations with two different ways of computing PCAs as mentioned above. These values serve as quantitative descriptors of vocal entrainment for the speaker (husband, wife) at that moment. Table 2 summarizes the eight entrainment measures with the associated computation methods.

4. Statistical analysis of vocal entrainment

Section 3 describes a framework to measure vocal entrainment using a completely signal-derived and unsupervised method. Since these measures are computed directly on the raw acoustic cues, we devised a statistical hypothesis test to investigate whether these signal-derived measures are capable of capturing the existence of the natural cohesiveness in human-to-human conversations. We set this up as a verification scheme to establish the validity of the proposed computational method using the dataset, **Dataset**$_{\text{qual}}$ (Section 2.1). The procedure and the result of this evaluation scheme are described in Section 4.1.

Table 2
Summarization of methods for computing the proposed vocal entrainment measures.

	Symmetric		Directional		PCA type	
	Unweight	Weight	Toward	From	Global	Local
$ssim_{G_u}$	√				√	
$ssim_{G_w}$		√			√	
$ssim_{L_u}$	√					√
$ssim_{L_w}$		√				√
$dsim_{G_{to}}$			√		√	
$dsim_{G_{fr}}$				√	√	
$dsim_{L_{to}}$			√			√
$dsim_{L_{fr}}$				√		√

After establishing the validity of the proposed measures in quantifying the subtle phenomenon of vocal entrainment, we carried out a second analysis that focuses on the relationship between vocal entrainment and the affective states. This analysis was built upon existing psychology literature describing the positive processes between interacting dyads under other types of interaction contexts. In this work, we pursue the hypothesis that vocal entrainment reflects behavior dependencies underlying the affective interactions of severely distressed couples. This analysis was carried out using the dataset **Dataset**$_{\text{emo}}$ (Section 2.2).

We employed two different statistical testing techniques. We first performed the commonly used Student's *t*-Test given the large number of samples in our database. The histograms show that directional measures are skewed slightly to the left and symmetric measures are skewed slightly to the right. We took the square root of symmetric measures and the square of directional measures to transform the histograms to be more *normal* before carrying out Student's *t*-Test. Two different tests of normality were carried out on the transformed variables: the Kolmogorov–Smirnov test and the Shapiro–Wilk normality test. Sometimes, the two tests indicated different results on the normality test for different entrainment measures. We also include Mann–Whitney's *U*-Test, a non-parametric version of Student's *t*-Test, in the result table to offer a more thorough and complete analysis.

4.1. Analysis I: verifying signal-derived vocal entrainment measures

The design of this analysis was based on the psychological knowledge that interlocutors exert mutual influences (Andersen and Andersen, 1984; Watt and VanLearn, 1996; Burgoon et al., 1995) on each other's behavior as they engage in conversations; we refer to this well-known intuitive nature of dialogs in this context as *natural cohesiveness*. To establish the validity of this unsupervised computational framework, we analyzed whether the proposed vocal entrainment measures capture the existence of this natural cohesiveness between interlocutors. We first computed the eight vocal entrainment measures for each of the spouses in the dataset, **Dataset**$_{\text{qual}}$, at every speaking turn in every interaction session. Then, we compared the mean value of each of the eight measures to the same set of the vocal entrainment measures computed on "*randomly generated*" dialogs. The following one-sided hypothesis testing was carried out for each individual entrainment measure.

$$H_o : \mu_{entrain_{\text{dialog}}} = \mu_{entrain_{\text{rand}}}$$
$$H_a : \mu_{entrain_{\text{dialog}}} > \mu_{entrain_{\text{rand}}}$$

The hypothesis states that the measures computed in dialogs where spouses were engaging in *real* conversations were expected to have higher degrees of entrainment compared to measures computed in artificially generated dialogs from two randomly selected spouses that were not interacting. Note that since measures $dsim_{L_{to}}$, $dsim_{L_{fr}}$, $dsim_{G_{to}}$, $dsim_{G_{fr}}$ were computed based on KLD, we expect lower numerical values indicating a higher level of entrainment (similarity). We followed these steps to generate "artificial dialogs" for the hypothesis testing:

Table 3
Analyzing vocal entrainment measures for natural cohesiveness (1000 runs): percentage of rejecting H_o at $\alpha = 0.05$ level.

Measures	Student's *t*-Test	Mann–Whitney's *U*-Test
$ssim_{G_u}$	100.00%	100.00%
$ssim_{G_w}$	100.00%	100.00%
$ssim_{L_u}$	100.00%	99.70%
$ssim_{L_w}$	100.00%	100.00%
$dsim_{L_{to}}$	100.00%	100.00%
$dsim_{L_{fr}}$	100.00%	100.00%
$dsim_{G_{to}}$	100.00%	100.00%
$dsim_{G_{fr}}$	100.00%	100.00%

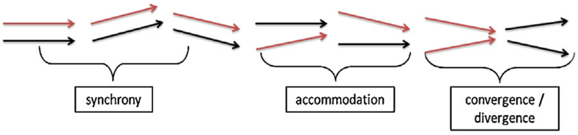


Fig. 4. Conceptualization of the dynamic interplay of the directionality of influences in dyadic interactions.

1. For a given subject in the Couple Therapy corpus in each of his/her sessions, randomly select another subject from another session in the database with the constraint that these two subjects are not a couple and this randomly selected subject is of opposite gender.
2. Gather these two “randomly selected” non-interacting spouses' speaking turns to form an *artificial* dialog.
3. Compute *eight* (Section 3.3) entrainment values for this subject in this artificial dialog.
4. Repeat steps 1–3 for every subject in the corpus.
5. Repeat step 4 for 1000 times.

The purpose of step 5 listed above is to generate a large “artificial dataset” consisting of “randomly assembled artificial” dialogs.

Table 3 shows the percentage of times that each measure passed the hypothesis test (out of 1000 runs) indicating a statistically significant higher degree of entrainment captured by that specific quantitative descriptor. We observe that both the symmetric and the directional measures of vocal entrainment almost always indicate a statistically significant higher degree of vocal entrainment in *real* conversations compared to *artificial* conversations establishing the evidence that the proposed computation is a viable method in quantifying natural cohesiveness in interpersonal conversations. While the Couple Therapy corpus' audio recording conditions varied from session to session, the evidence of improved robustness of the audio feature extraction (Section 2.1) is also clear in the result of this hypothesis test. This result provides one validation that our vocal entrainment measures computed with signal processing techniques using audio-only features carry meaningful information about the nature of the interaction.

Another point to make is that this test of the “natural cohesiveness” in the dialog is conceptually non-directional. A psychology study on interpersonal communication (Burgoon et al., 1995) describes the following phenomenon in a dyadic human–human interaction; for a given attribute of interest, e.g., engagement level, when the direction of influence is introduced (not concentrating solely on the absolute degree of influence between dyads), the dynamic interplay between these influences can be roughly described in terms of three possible categories (Fig. 4) depending on who exerts a stronger force of influence. It has been conceptualized that the evolution of these dynamic interplays characterizes the essential *flow* of the dialog.

While our proposed signal-derived directional entrainment quantification measures carry this notion of dynamic interplay between influences of the dyad, it is challenging to systematically validate the psychological significance of directional measures in the context of this comparison between *real* dialogs and *artificial* dialogs. It is encouraging, however, to see that our proposed vocal entrainment measures demonstrate their efficacy in capturing the natural cohesiveness expected to occur in spontaneous human–human conversations.

Table 4
Analyzing the vocal entrainment measures for affective interactions (*positive* affect vs. *negative* affect): the one-sided *p*-value is presented.

Measures	Student's *t*-Test	Mann–Whitney's *U*-Test
$ssim_{L_u}$	**0.047**	**0.047**
$ssim_{L_w}$	**0.012**	**0.003**
$ssim_{G_u}$	0.073	**0.045**[*]
$ssim_{G_w}$	**0.005**	0.406
$dsim_{L_{to}}$	0.098	0.078
$dsim_{L_{fr}}$	**0.002**	**0.002**
$dsim_{G_{to}}$	**0.016**	**0.004**
$dsim_{G_{fr}}$	**<0.0001**	**<0.0001**

[*] $H_a : \mu_{entrain_{\text{pos}}} < \mu_{entrain_{\text{neg}}}$.

4.2. Analysis II: analyzing vocal entrainment in affective interactions

In this section, we demonstrate the potential of utilizing the proposed computational framework in discovering insights of psychological significance about distressed couples' interactions. There is extensive psychology literature studying the nature of the affective states of distressed married couples in conflicts. One very important finding from the large body of work on negative conflict processes in distressed couples is that behavioral rigidity is characteristic of relationship dysfunction (Eldridge et al., 2007). In other words, very dissatisfied couples tend to be increasingly negative and only negative during conflict, and this "reinforcing" of negative behaviors between the spouses is common and problematic. While most studies have concentrated on negative processes of the distressed married couples, positive processes have received surprisingly little attention in the psychology literature in general and in work with distressed couples specifically.

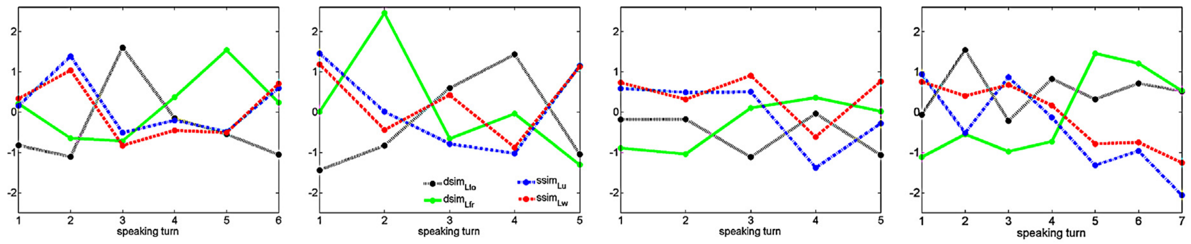
It remains unknown what processes are associated with greater flexibility, such as increased levels of positivity, during couples' conflict. We hypothesize that entrainment is likely to be one such process because it is a precursor to empathy (Verhofstadt et al., 2008). Numerous theoretical models of relationship functioning suggest that empathy plays a crucial role in helping couples successfully negotiate and resolve conflict (Baucom and Atkins, in press), perhaps by allowing them to be more flexible and express greater positive emotion during conflict. Based on this idea, the following one-sided hypothesis testing was carried out on the dataset, **Dataset**$_{\text{emo}}$ (Section 2.2):

$$H_o : \mu_{entrain_{\text{pos}}} = \mu_{entrain_{\text{neg}}}$$
$$H_a : \mu_{entrain_{\text{pos}}} > \mu_{entrain_{\text{neg}}}$$

This hypothesis states that each of the vocal entrainment measures would result in a higher degree of similarity for spouses in sessions rated with *positive* affect compared to spouses in sessions rated with *negative* affect.

Table 4 shows the result of the hypothesis testing; numbers in bold are statistically significant at the 5% level. We observe that most of the measures (especially, directional measures) indicate a statistically significant higher degree of vocal entrainment for spouses rated with positive affect. Another observation we can make about the symmetric measures is that those calculated based on the globally computed PCA may lead to different results of the two statistical tests. For example, the *p*-values for Student's *t*-Test and Mann–Whitney's *U*-Test differ a lot for $ssim_{G_w}$ and $ssim_{G_u}$. Given the assumptions of these two tests, this inconsistency may be attributed to the fact that the tested values are highly concentrated and have very little variance. This property is not really captured correctly by the parametric test. These two measures, $ssim_{G_w}$ and $ssim_{G_u}$, apparently do not differ much between the two affective states we are interested in. However, the overall degrees of vocal entrainment computed by the directional measures are all statistically significantly higher (except for $dsim_{L_{to}}$, which did not meet the 5% level). This suggests that although all of the entrainment measures are based on metrics of similarity, the directional measures carry somewhat distinct information from the symmetric measures.

While most psychological studies have shown that the "behavioral dependency" of the negative process occurs often during distressed married couples' interactions, we have also demonstrated an initial empirical evidence that there can be a higher degree of vocal entrainment for spouses rated with a positive affective state when we consider the directional influences of the vocal entrainment phenomenon. More detailed analyses are needed of the entrainment



(a) Positive session 1 (b) Positive session 2 (c) Negative session 1 (d) Negative session 2

Fig. 5. Examples of vocal entrainment measures: $dsim_{L_{to}}$, $dsim_{L_{fr}}$, $ssim_{L_u}$, $ssim_{L_w}$, are computed for one couple in different affective interactions; (a) and (b) correspond to positive affect. (c) and (d) correspond to negative affect. (a) Positive session 1. (b) Positive session 2. (c) Negative session 1. (d) Negative session 2.

process measured by symmetric measures to strengthen the theories about positive processes in a couple's conflict. Furthermore, although there has been previous work (Pasch et al., 1997) analyzing the notion of *directionality* of influence in affective marital interactions, the knowledge of the dynamics between the behavior interplay remains limited. The analyses we presented here can be viewed as a first attempt to bring insights into the positive process of distressed couples' interactions through the use of the proposed computational framework.

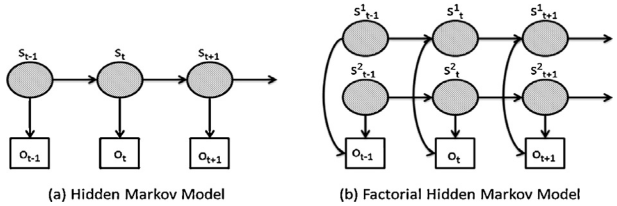
In summary, we tested two psychology-inspired hypotheses. The purpose was to demonstrate that the proposed signal-derived entrainment measures, based only on acoustic vocal properties, can capture vocal entrainment between interlocutors. These tests, however, only examined the overall pattern of entrainment and not the dynamic flow of the interplay between interlocutors' mutual influences. This was due to the challenges in grounding the results within the complex nature of these pattern variations. It is, nevertheless, encouraging to see that these computational measures are capable of numerically describing several important aspects of human interactions. The analysis that we present in this section serves as an example of how this computational framework can be a potential viable method in helping psychologists to quantitatively study entrainment in interpersonal communication and more importantly in various mental health applications, e.g., distressed marital interactions, where the knowledge remains limited.

5. Analysis III: applying vocal entrainment measures in affect recognition

Section 4 presents two analyses examining the validity and usefulness of the proposed computational framework of vocal entrainment. Herein we consider an application to investigate whether the entrainment measures can be useful in affective behavioral code classification. The purpose of the affect recognition experiment is to serve as an exemplary case application to examine the predictive power of these audio signal-derived entrainment measures. The dataset used in this section is the **Dataset**$_{\text{emo}}$ (Section 2.2). We first discuss briefly the statistical modeling framework, Factorial Hidden Markov Model, that we used to perform the affect recognition. Then we describe the classification setup, and finally we present the results and discussions.

5.1. Classification framework

The entrainment phenomenon is a complex temporal evolution of interplay between the directions of influences (see conceptualization shown in Fig. 4) of the interlocutors. As an example illustrating the complex dynamics of our proposed vocal entrainment values, we show four vocal entrainment measures for a specific couple under two different affective states (two sessions per emotion class) in Fig. 5. There is not an easily observable pattern of evolution on each individual entrainment measure for the two emotion classes throughout the dialog. Each measure seems to indicate a slightly different degree of vocal entrainment at the same time point. In order to better model and capture this complex dynamic in an interaction, we employ temporal statistical modeling techniques. The entrainment measures (our observed features) were computed at each speaking turn and the affective state rating was assigned at the interaction session level.



(a) Hidden Markov Model (b) Factorial Hidden Markov Model

Fig. 6. Dynamic Bayesian Network representation of (a) HMM and (b) FHMM.

5.1.1. *Factorial Hidden Markov Model*

Factorial Hidden Markov Model (FHMM, Ghahramani and Jordan, 1997) is a generalization and extension of the Hidden Markov Model (HMM). Given a time series, $\mathbf{O} = \{O_t : t = 1, \ldots, T\}$, HMMs describe that the observation at each time index is probabilistically generated from one of K hidden discrete states. The probability that an observation sequence, $\mathbf{O}$, is generated from a particular HMM model, λ_i, is expressed as follows:

$$p(\mathbf{O}|\lambda_i) = \sum_{S} \pi(S_1) p(O_1|S_1) \prod_{t=2}^{T} p(S_t|S_{t-1}) p(O_t|S_t) \tag{15}$$

where

- $\mathbf{O}$ = sequence of observation vectors $\{O_t : t = 1 \ldots T\}$;
- S = sequence of discrete states $\{S_t : t = 1 \ldots T\}$;
- $p(S_t|S_{t-1})$ = transition probability from state S_{t-1} to S_t;
- $\pi(S_1)$ = initial state probability;
- $p(O_t|S_t)$ = probability of observation vector, O_t, given the state, S_t;
- K = number of states in the model, i.e., $S_t \in \{1, \ldots, K\}$.

HMMs can be easily represented in a directed acyclic graphical structure, i.e., Dynamic Bayesian Network (DBN). The DBN representation of HMMs is shown in Fig. 6a, where shaded nodes are hidden states. As seen in Fig. 6a, an HMM is essentially modeling a single hidden process generating a set of observable features probabilistically. Hence, a natural extension of this framework is introducing *layers* of hidden processes consisting of multiple hidden variables (Ghahramani and Jordan, 1997). Now instead of single state variable, S_t, we obtain a new state variable, $S_t^{(M)}$:

$$S_t^{(M)} = S_t^1, S_t^2, \ldots, S_t^M$$

where M represents the number of layers, and each S_t^m can take on K_m number of states (we simplify it by having each S_t^m take the same number of possible states, K). Ghahramani introduced one intuitive constraint placed on the state transitions by considering that each state variable evolves based on its own previous dynamics and is *a priori* decoupled from other state variables. This means that each state transition probability distribution can be expressed as follows:

$$p(S_t^{(M)}|S_{t-1}^{(M)}) = \prod_{m=1}^{M} p(S_t^m|S_{t-1}^m) \tag{16}$$

An example DBN representation of FHMM with two layers is shown in Fig. 6b. The transition matrix for all the state variables can be parametrized by M distinct $K \times K$ matrices. While the state transitions are not coupled together, they are coupled at the observation node. One simple form of such a dependency is linear Gaussian, i.e., assuming that the continuous observation O_t is a Gaussian random vector ($N \times 1$ dimension) whose mean is a linear function of the states. We can write the observation probability as shown below:

$$p(O_t|S_t^m) = |C|^{-1/2}(2\pi)^{-N/2} \exp\left\{-\frac{1}{2}(O_t - \mu_t)' C^{-1} (O_t - \mu_t)\right\}$$

where

$$\mu_t = \sum_{m=1}^{M} W^m S_t^m.$$

Each W^m matrix is an $N \times K$ matrix with each column indicating the contribution to the means for each setting of S_t^m. C is the covariance matrix.

The use of FHMM in our context of affect recognition, given the vocal entrainment measures as observation, is intuitively appealing. The essence of FHMM is modeling multiple *loosely coupled* hidden processes with the generation of observations depending on all hidden processes. Since the observable feature vectors used in this recognition task were computed based on both of the interlocutors, we satisfy the assumption. We designed the FHMM as having two layers intuitively modeling two interacting processes (husband and wife). Furthermore, the *loosely coupled* nature of FHMM qualitatively corresponds to the *subtle* nature of this entrainment phenomenon.

Since both FHMM and HMM can be represented as DBNs, we implemented them using the Bayes Net Toolbox (BNT) (Murphy, 2001) with the standard junction tree algorithm as the exact inference method; FHMM with two hidden processes is tractable for the junction tree algorithm. Expectation-maximization (EM) was carried out to estimate the model parameters. We used a mixture of Gaussians to model the observations, and this was done by simply adding another discrete node in the construction of the DBNs. The classification rule was based on the standard maximum *a posteriori* probability as shown below:

$$i^* = arg\max_{i} P(\lambda_i|\mathbf{O}) \tag{17}$$

where $i \in \{\text{positiveaffect, negativeaffect}\}$.

5.2. Classification setup

The recognition was a binary classification task classifying each spouse's affective state (positive affect vs. negative affect) in a given interaction session. There are a total of 280 samples with equally sized splits between the two emotion classes.

In addition to the eight *vocal entrainment* measures, we computed five more similarity measures. We denote these five similarity measures as "*self vocal similarity*" quantitative descriptors. We computed them to measure the self similarity of vocal characteristics for a speaker in an interaction. These measures can be interpreted approximately as the degree to which a given speaker's speaking style stays the same (consistent) in the course of the dialog. We computed them using the same PCA framework (Section 3). Since these measures describe the self similarity, instead of using acoustic parameters from the other speaker as the "other interacting process", we used the acoustic parameters of the same speaker from his/her own immediate next speaking turn. Moreover, these measures were computed on two turns from the same subject; therefore, many of the similarity measures using globally computed PCAs were not applicable (resulting in the same values for all turns). We used five out of the eight measures ($dsim^s_{L_{to}}$, $dsim^s_{L_{fr}}$, $ssim^s_{L_u}$, $ssim^s_{L_w}$, $dsim^s_{G_{to}}$).

We trained and evaluated a total of five different models. We trained the *combined* model using feature-level fusion of entrainment measures and self similarity measures; this resulted in a feature vector of length 13. We did not explicitly train an FHMM model using self similarity measures because the computation process of these measures inherently assumes a single hidden process (computed based on a single spouse in the dialog).

We performed model evaluation using leave-one-couple-out cross validation (81 folds of cross validation) with the percentage of accurately classified emotions as our metric. Various parameters, such as the number of states and the number of mixtures in the mixture of Gaussians, were optimized for each testing fold through another fivefold cross validation done on the training dataset only. We chose the parameters for each fold that resulted in the highest classification accuracy from the fivefold cross validation done on the training data. The number of states that were used ranged from three to seven, and the number of mixtures ranged from one to four. We also performed z-normalization on these similarity measures to obtain better numerical properties.

Table 5
Results of binary affective state (positive vs. negative) recognition: overall percentage of accurately classified (%), per emotion class accuracy, and (*dSIM*, *sSIM*) directional and symmetric measures accuracy.

Models	Accuracy: (positive and negative)	*dSIM*: (positive and negative)	*sSIM*: (positive and negative)
Chance	50%	N/A	N/A
HMM (entrainment)	55.75%: (49.30 and 62.14%)	55.71%: (50.00 and 61.43%)	50.00%: (38.57 and 61.43%)
HMM (self similarity)	47.50%: (51.43 and 43.57%)	49.29%: (49.29 and 49.29%)	54.64%: (67.15 and 42.14%)
HMM (combined)	55.72%: (52.86 and 58.57%)	50.00%: (45.00 and 55.00%)	58.57%: (57.86 and 59.29%)
FHMM (entrainment)	**62.86%**: (65.71 and 60.00%)	52.86%: (56.43 and 49.29%)	55.35% (53.57 and 57.14%)
FHMM (combined)	54.29%: (55.00 and 53.57%)	55.36%: (57.86 and 52.86%)	54.64% (57.14 and 52.14%)

The number in bold indicates that the FHMM (entrainment) obtained the highest accuracy among all the models tested. The improvement in the accuracy is statistically significant using one-sided McNemar test.

5.3. *Classification results and discussions*

Table 5 shows the overall affective state classification accuracy results as well as the class-wise accuracy for the five models described above. There are several observations to be made with Table 5. The best performing model is FHMM trained with vocal entrainment features only; it obtained an overall accuracy of 62.86%. We used one-sided McNemar's test for assessing the statistical significance of this classification result, and this model (FHMM with entrainment) outperforms all four other models at $\alpha = 0.05$ level. The quantification method with FHMM improves the affective state recognition accuracy by an absolute of 8.93% (16.56% relative) compared to using multiple instance learning with "variance-preserved" as the only measure of entrainment in the same dataset. It is encouraging to see that the temporal dynamics of these quantitative descriptors of vocal entrainment possess discriminant power in classifying a spouse's affective state.

As noted in the comparison between FHMM and HMM, using only entrainment measures as features, the accuracy of using FHMM is 7.11% absolute (12.75% relative) better than using HMM. This statistically significant improvement in affective recognition emphasizes the importance of adequately capturing the interaction dynamics between interlocutors while using these entrainment measures, which themselves are also derived from both spouses in the interaction. Furthermore, in Section 4.2, we showed that the average session-level entrainment values computed using directional measures are different from symmetrical measures; we included the classification accuracy using models trained separately on directional and symmetrical measures (results are also shown in Table 5). While there exists difference in the classification results between the two types of measures, more detailed future investigation needs to be carried out to understand the relationship between the different dynamics of these two types of measures and each spouse's affective state. However, the result shows that it is beneficial in the FHMM framework to combine both types of measures.

Another interesting point regards the comparison between the use of entrainment measures and self similarity measures (inter vs. intra person modeling comparison). Our result indicates that merely modeling the self similarity of speaking style in the dialog does not carry information on the affective state of an individual. The accuracy of the HMM based on self similarity measures is even below chance, and the combination of these features with entrainment measures is shown to be detrimental to the overall recognition accuracy compared to using only the entrainment measures.

It is possible that this PCA entrainment framework is not appropriate to quantify the "self similarity" or that these measures simply do not carry information about this specific attribute of interest: affective state. However, we hypothesize that we observe this classification result because it is the *interaction*, i.e., the dynamic interplay between spouses, that is at the core of characterizing and shaping the essence of each interlocutor's behaviors and mental states.

In summary, to illustrate a possible application of using entrainment as features to predict behaviors, we performed affective state recognition in married couples' interactions. Using only eight features and by utilizing FHMM, which implicitly modeled the interaction dynamics between the spouses, we obtained an accuracy of 62.86%. It is promising to see that these signal-derived measures not only can be used to quantitatively describe aspects of entrainment between interlocutors, but they can also be incorporated in a statistical modeling framework to carry out behavioral prediction tasks. Further, while we demonstrated through this analysis that there is a relationship between affective state and vocal entrainment, the primary purpose of the experiment was not affect state classification but to explore the role of entrainment in reflecting behavioral dependencies in (affective) interactions. As a result, no effort was made to devise the best performing affect classification setup. Prior work, including ours (Georgiou et al., 2011; Black et al., in press; Katsamanis et al., 2011b), has shown that the overall recognition can be advantageously improved by combining different communication cues. We believe that to achieve higher accuracy in affect recognition, and to extend it to ambiguous emotion classes and other behavioral codes, many more observable cues must also be included, and the proposed vocal entrainment measures can be an essential component of such a system.

6. Conclusion and future work

The degree of interpersonal behavioral dependency is a critical component in understanding human–human communication. It also plays a crucial role for psychologists in their study of various intimate and distressed relationships. In this work, we focused on the phenomenon of entrainment. While the knowledge of this type of behavioral dependency, entrainment, is vast across various domains of human interaction studies, its subtle nature and often qualitative aspect have likely hindered advances in its computational quantification. In this work, we proposed a signal-derived framework for computing numerical values indicating the degree of a specific aspect of entrainment, *vocal entrainment*. The quantification framework is unsupervised, and the idea is centered on computing various similarity measures between PCA-based representations of automatically extracted acoustic parameters of interlocutors engaged in a dialog. We demonstrated in this work that these quantitative descriptors can capture aspects of entrainment and bring insights into distressed married couples' interactions using a well-established corpus of spontaneous affective interactions from real married couples. Furthermore, we obtained an 62.86% accuracy using just these eight entrainment measures in a binary affective state recognition task quantitatively corroborating hypotheses about the relation between entrainment and affective behavior.

There are many future directions in the work of computationally analyzing the phenomenon of entrainment. This work demonstrates the relation between vocal entrainment and affect. Although the two are related, they are not identical, and hence the upper bound of estimation of affect through vocal entrainment is unknown. To improve understanding of the role of vocal entrainment in characterizing human communication, one of the immediate directions is to extend the classification work in the paper to analyze other behavioral codes of interest in this richly annotated corpus of real distressed married couples' interactions. We would like to examine in detail both the predictive power of these vocal entrainment measures for various behavioral attributes and the potential upper bound of classification accuracy using entrainment in the context of couples' interactions. Furthermore, we would also like to analyze the vocal entrainment at the session level for each couple given that the Couple Therapy database is longitudinal in nature. It would be insightful to have a quantitative monitoring of spouses' behaviors through a longer time span.

Another important line of work is related to this broad nature of entrainment. Entrainment includes various aspects beyond vocal similarity, such as lexical entrainment, gestural entrainment, turn-taking entrainment, and mental states entrainment. In fact, this phenomenon spans multiple communicative channels and multiple levels in human communication, and often an interacting effect from all these dependencies between interlocutors characterizes the *felt-sense* or *quality* of a given interaction. Various studies in mental health have pointed out the crucial aspect of this quality of an interaction in understanding different scenarios of interactions. One of our future works is to continuously develop computational frameworks in examining entrainment through other modalities to both see what is the relationship of same-subject, across-modality signaling coherence and also to obtain richer information regarding interlocutors' mirroring behaviors.

Finally, the engineering tools developed in the emerging field of behavioral signal processing (BSP) can benefit psychologists and other domain experts by allowing automated and meaningful computation of behavioral properties both for scientific needs and for translational applications in diagnostics and intervention planning. These tools can generate quantitative descriptors to be used in analyzing different domains of interpersonal communication. In fact, some of these measures have the potential of capturing aspects of interaction that are inherently difficult to annotate even for experts; one example is the dynamic interplay of directional influences between dyads at various levels (acoustic, prosodic, lexical, gestural, etc.). This requires a tight collaboration between psychologists and engineers to develop computational methods that are grounded in psychologically meaningful questions and theory.

Acknowledgment

The work was supported in part by NSF, NIH, and DARPA.

References

Andersen, P.A., Andersen, J.F., 1984. The exchange of nonverbal intimacy: a critical review of dyadic models. Journal of Nonverbal Behavior 8 (4), 328–349.

Baucom, B., Atkins, D. Polarization in marriage. In: Fine, M., Fincham, F. (Eds.), Family Theories: A Content-Based Approach. Routledge, New York, NY, in press.

Bernieri, F.J., Reznick, J.S., Rosenthal, R., 1988. Synchrony, pseudosynchrony, and dissynchrony: measuring the entrainment process in mother–infant interactions. Journal of Personality and Social Psychology 54 (2), 243–253.

Black, M.P., Georgiou, P.G., Katsamanis, A., Baucom, B.R., Narayanan, S.S., 2011. “You made me do it”: classification of blame in married couples interactions by fusing automatically derived speech and language information. In: Proceedings of Interspeech, pp. 89–92.

Black, M.P., Katsamanis, A., Baucom, B.R., Lee, C.C., Lammert, A.C., Christensen, A., Georgiou, P.G., Narayanan, S.S. Toward automating a human behavioral coding system for married couples’ interactions using speech acoustic features. Speech Communication, in press 2011, http://dx.doi.org/10.1016/j.specom.2011.12.003.

Black, M.P., Katsamanis, A., Lee, C.C., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2010. Automatic classification of married couples’ behavior using audio features. In: Proceedings of Interspeech, pp. 2030–2033.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot International 5, 341–345.

Boker, S.M., Laurenceau, J.P., 2006. Dynamical systems modeling: an application to the regulation of intimacy and disclosure in marriage. In: Models for Intense Longitudinal Data, pp. 195–218.

Burgoon, J.K., Stern, L.A., Dillman, L., 1995. Interpersonal Adaptation: Dyadic Interaction Patterns. Cambridge University Press.

Chartrand, T.L., Bargh, J.A., 1999. The chameleon effect: the perception–behavior link and social interaction. Journal of Personality and Social Psychology 76 (6), 893–910.

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

Cordova, J.V., Jacobson, N.S., Gottman, J.M., Rushe, R., Cox, G., 1994. Negative reciprocity and communication in couples with a violent husband. Journal of Abnormal Psychology 102, 559–564.

Dauwels, J., Vialatte, F., Cichocki, A., 2010. Diagnosis of Alzheimer’s disease from EEG signals: where are we standing? Current Alzheimer’s Research 7, 487–505 (Invited paper).

Eldridge, K., Baucom, B., 2010. Couples and consequences of the demand-withdraw interaction pattern. In: Positive Pathways for Couples and Families: Meeting the Challenges of Relationships. Wiley Blackwell.

Eldridge, K.A., Sevier, M., Jones, J., Atkins, D.C., Christensen, A., 2007. Demand-withdraw communication in severely distressed, moderately distressed, and nondistressed couples: rigidity and polarity during relationship and personal problem discussions. Journal of Family Psychology 21, 218–226.

Eyben, F., Wöllmer, M., Schuller, B., 2010. OpenSMILE – the Munich versatile and fast open-source audio feature extractor. In: ACM Multimedia, Firenze, Italy, pp. 1459–1462.

Georgiou, P., Black, M., Lammert, A., Baucom, B., Narayanan, S., 2011. “That’s aggravating, very aggravating”: is it possible to classify behaviors in couple interactions using automatically derived lexical features? In: Affective Computing and Intelligent Interaction, pp. 87–96.

Ghahramani, Z., Jordan, M.I., 1997. Factorial hidden markov models. Machine Learning 29, 245–274.

Ghosh, P.K., Tsiartas, A., Narayanan, S.S., 2010. Robust voice activity detection using long-term signal variability. IEEE Transactions on Audio, Speech, and Language Processing 19, 600–613.

Gibson, J., Katsamanis, A., Black, M., Narayanan, S., 2011. Automatic identification of salient acoustic instances in couples behavioral interactions using diverse density support vector machines. In: Proceedings of Interspeech, pp. 1561–1564.

Gottman, J., Swanson, C., Swanson, K., 2002. A general system theory of marriage: nonlinear difference equation modeling of marital interaction. Personality and Social Psychology Review 6, 326–340.

Gottman, J.M., 1993. The roles of conflict engagement, escalation, and avoidance in marital interaction: a longitudinal view of five types of couples. Journal of Consulting and Clinical Psychology 61, 6–15.

Gregory, S., Dagan, K., Webster, S., 1997. Evaluating the relation between vocal accommodation in conversational partners’ fundamental frequencies to perceptions of communication quality. Journal of Nonverbal Behavior 21, 23–43.

Gregory, S., Hoyt, B., 1982. Conversation partner mutual adaptation as demonstrated by Fourier series analysis. Journal of Psycholinguistic Research 11, 35–46.

Gregory, S., Webster, S., 1996. A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. Journal of Personality and Social Psychology 70, 1231–1240.

Gregory, S., Webster, S., Huang, G., 1993. Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. Language and Communication 13, 195–217.

Heavey, C., Gill, D., Christensen, A., 2002. Couples Interaction Rating System 2 (CIRS2). University of California, Los Angeles.

Jacobson, N.S., Gottman, J.M., Waltz, J., Rushe, R., Babcock, J., Holtzworth-Munroe, A., 1994. Affect, verbal content, and psychophysiology in the arguments of couples with a violent husband. Journal of Consulting and Clinical Psychology 62, 982–988.

Johnannesmeyer, M., 1999. Abnormal situation analysis using pattern recognition techniques and historical data. Master's Thesis. UCSB, Santa Barbara, CA.

Johnson, S.L., Jacob, T., 2000. Sequential interactions in marital communication of depressed men and women. Journal of Consulting and Clinical Psychology 68, 4–12.

Jones, J., Christensen, A., 1998. Couples Interaction Study: Social Support Interaction Rating System. University of California, Los Angeles.

Jurafsky, D., Ranganath, R., McFarland, D., 2009. Extracting social meaning: identifying interactional style in spoken conversation. In: Human Language Technologies, Boulder, CO, USA, pp. 638–646.

Katsamanis, A., Black, M.P., Georgiou, P.G., Goldstein, L., Narayanan, S.S., 2011a. SailAlign: robust long speech-text alignment. In: Very-Large-Scale Phonetics Workshop.

Katsamanis, A., Gibson, J., Black, M., Narayanan, S., 2011b. Multiple instance learning for classification of human behavior observations. In: Affective Computing and Intelligent Interaction, pp. 145–154.

Kerig, P., Baucom, D.E., 2004. Couple Observational Coding Systems. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Kimura, M., Daibo, I., 2006. Interactional synchrony in conversations about emotional episodes: a measurement by 'the between-participants pseudosynchrony experimental paradigm'. Journal of Nonverbal Behavior 30, 115–126.

Krzanowski, W., 1979. Between-groups comparison of principal components. Journal of the American Statistical Association 74, 703–707.

Kwon, O.W., Kwokleung, C., Hao, J., Lee, T.W., 2003. Emotion recognition by speech signals. In: Proc. EUROSPEECH, pp. 125–128.

Lee, C.C., Black, M.P., Katsamanis, A., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. Proceedings of Interspeech, 793–796.

Lee, C.C., Katsamanis, A., Black, M.P., Baucom, B.R., Georgiou, P.G., Narayanan, S.S., 2011a. Affective state recognition in married couples' interactions using PCA-based vocal entrainment measures with multiple instance learning. In: Affective Computing and Intelligent Interaction, pp. 31–41.

Lee, C.C., Katsamanis, A., Black, M.P., Baucom, B.R., Georgiou, P.G., Narayanan, S.S., 2011b. An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions. In: Proceedings of Interspeech, pp. 3101–3104.

Levitan, R., Hirschberg, J., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Proceedings of Interspeech, pp. 3081–3084.

McGarva, A.R., Warner, R.M., 2003. Attraction and social coordination: mutual entrainment of vocal activity rhymes. Journal of Psycholinguistic Research 32, 335–354.

Murphy, C.M., O'Farrell, T.J., 1997. Couple communication patterns of maritally aggressive and nonaggressive male alcoholics. Journal of Studies on Alcohol 58, 83–90.

Murphy, K.P., 2001. The Bayes net toolbox for Matlab. Computing Science and Statistics.

Nenkova, A., Gravano, A., Hirschberg, J., 2008. High frequency word entrainment in spoken dialogue. In: Proceedings of ACL-08: HLT, Short Papers, Columbus, OH, pp. 169–172.

Niederhoffer, K.G., Pennebaker, J.W., 2002. Linguistic style matching in social interaction. Journal of Language and Social Psychology 21, 337–360.

Otey, M.E., Parthasarathy, S., 2005. A dissimilarity measure for comparing subsets of data: application to multivariate time series. In: Proceedings of ICDM Workshop on Temporal Data Mining, Houston, TX.

Pardo, J.S., 2006. On phonetic convergence during conversational interaction. Journal of Acoustical Society of America 119, 2382–2393.

Pasch, L.A., Bradbury, T.N., Davila, J., 1997. Gender, negative affectivity, and observed social support behavior in marital interaction. Personal Relationships 4, 278–361.

Ranganath, R., Jurafsky, D., McFarland, D., 2009. It's not you, it's me: detecting flirting and its misperception in speed-dates. In: Conference on Empirical Methods in Natural Language Processing, Suntec City, Singapore, pp. 334–342.

Reynolds, D.A., 1994. Experimental evaluation of features for robust speaker identification. IEEE Transactions on Speech and Audio Processing 2, 639–643.

Richardson, M.J., Marsh, K.L., Schmit, R., 2005. Effects of visual and verbal interaction on unintentional interpersonal coordination. Journal of Experimental Psychology: Human Perception and Performance 31, 62–79.

Romano, J.M., Turner, J.A., Friedman, L.S., Bulcroft, R.A., Jensen, M.P., Hops, H., Wright, S.F., 1992. Sequential analysis of chronic pain behaviors and spouse responses. Journal of Consulting and Clinical Psychology 60, 777–782.

Rozgić, V., Xiao, B., Katsamanis, A., Baucom, B., Georgiou, P.G., Narayanan, S.S., 2011. Estimation of ordinal approach-avoidance labels in dyadic interactions: ordinal logistic regression approach. In: ICASSP, pp. 2368–2371.

Rozgic, V., Xiao, V., Katsamanis, A., Baucom, B.R., Georgiou, P.G., Narayanan, S.S., 2010. A new multichannel multimodal dyadic interaction database. In: Proceedings of Interspeech, pp. 1982–1985.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlation: uses in assessing rater reliability. Psychological Bulletin 86, 420–428.

Verhofstadt, L.L., Buysse, A., Ickes, W., Davis, M., Devoldre, I., 2008. Support provision in marriage: the role of emotional similarity and empathic accuracy. Emotion 8, 792–802.

Vinciarelli, A., Pantic, M., Bourlard, H., 2009. Social signal processing: survey of an emerging domain. Image and Vision Computing 27, 1743–1759.

Watt, J.H., VanLearn, C.A. (Eds.), 1996. Dynamic Patterns in Communication Processes. Sage Publication, Inc.