Affective State Recognition in Married Couples' Interactions Using PCA-Based Vocal Entrainment Measures with Multiple Instance Learning

Chi-Chun Lee[1], Athanasios Katsamanis[1], Matthew P. Black[1], Brian R. Baucom[2], Panayiotis G. Georgiou[1], and Shrikanth S. Narayanan[1,2]

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

Abstract. Recently there has been an increase in efforts in Behavioral Signal Processing (BSP), that aims to bring quantitative analysis using signal processing techniques in the domain of observational coding. Currently observational coding in fields such as psychology is based on subjective expert coding of abstract human interaction dynamics. In this work, we use a Multiple Instance Learning (MIL) framework, a saliency-based prediction model, with a signal-driven vocal entrainment measure as the feature to predict the affective state of a spouse in problem solving interactions. We generate 18 MIL classifiers to capture the variable-length *saliency* of vocal entrainment, and a cross-validation scheme with maximum accuracy and mutual information as the metric to select the *best* performing classifier for each testing couple. This method obtains a recognition accuracy of 53.93%, a 2.14% (4.13% relative) improvement over baseline model using Support Vector Machine. Furthermore, this MIL-based framework has potential for identifying meaningful regions of interest for further detailed analysis of married couples interactions.

Keywords: multiple instance learning, vocal entrainment, couple's therapy, behavioral signal processing (BSP), affective recognition.

1 Introduction

There has been an increasing effort in bridging the manual observation coding of human behaviors done in various mental health applications [10], such as couple therapy and autism spectrum disorder diagnosis, with the automatic annotation of abstract human behaviors/states using signal processing techniques, such as emotion recognition, using low level behavioral cues [8,5,13]. Manual observation coding done by the domain experts provides a subjective and detailed analysis of human-human behaviors/interactions in terms of various annotated attributes of interest, while automatic annotation using cues derived directly from observed signals provides a objective and quantitative analysis. Previous works [1,7,12]

S. D´Mello et al. (Eds.): ACII 2011, Part II, LNCS 6975, pp. 31–41, 2011.

have shown the effectiveness of applying machine learning techniques in predicting various behavioral ratings in married couples engaging in problem-solving interactions, e.g., blame, negativity, positivity, approach, avoidance, etc, using different variants of directly observed low level behavioral descriptors. There are two main goals in this present work. The first goal is to apply a suitable machine learning technique using signal-derived measures of vocal entrainment calculated at the speaking turn-level to predict the overall session-level codes of negativity/positivity of a spouse; result of this investigation could provide evidence of engineering utility of such a signal-derived entrainment measure. Second, since entrainment has been shown [3] to play a crucial role in analyzing marital communications, multiple instance learning (MIL) presents itself as an appealing framework because of its ability to identify *saliency* as it performs predictions. This potential advantage when combined with features that carry meaningful insights into the study of marital communication, can offer an opportunity to perform detailed analysis on these salient regions generated through MIL.

Multiple instance learning (MIL) is a widely used machine learning technique that has shown its effectiveness in pattern recognition applications such as drug activity estimation and image retrieval [11,15]. MIL is different from the traditional supervised learning technique, where an instance is associated with a label. In MIL, the label is associated with a bag that consists of multiple instances, and there is no explicit label given to each instance as training. This framework is appealing in working with session-level behavioral code prediction while the individual speaking turn annotation is unavailable. Furthermore, MIL can often be solved in a general way by using Diverse Density introduced in Maron and Lozano-Perez's [11] work. This involves in finding a concept point in the feature space that is close to at least one instance from every positive bag and far away from instances in negative bags. This concept point can be viewed as the *salient* points for a given bag label, and is used to identify the *salient* instances of a bag. This formulation of MIL along with the usage of vocal entrainment features can help identify the meaningful *salient* vocal entrainment measures to offer interpretable insights into regions of interest of couples' interactions while performing prediction on spouse's affective state.

In this work, we focus our recognition task using a vocal entrainment measure on sessions in which the spouse was rated with *high positive* or *high negative* affect. We utilize signal-derived vocal entrainment measures [9] as the only describing feature per instance in a MIL framework. Furthermore, our formulation allows a salient *instance* span not only a single speaking turn but multiple turns. We have applied a leave-one-couple out cross validation within training along with maximum accuracy and mutual information criteria to select the best performing classifier out of multiple MIL-based classifiers trained with different features lengths of *instances* for each training fold. The proposed method obtains a 53.93% accuracy, which is 3.93% (7.86% relative) over chance and 2.14% (4.13% relative) over a baseline Support Vector Machine (SVM) based classifier. This MIL-based classification scheme provides a method for performing classification through identification of a *best* MIL classifier, each classifier is learned

with variable-length salient instances of vocal entrainment. The method is able to obtain a fair recognition accuracy with vocal entrainment as the *only* features, and the classification result identifies interpretable saliency.

The paper is organized as follows: section 2 describes the database and research methodology. Experiment setup and results are discussed in section 3, and conclusions are in section 4.

2 Research Methodology

2.1 Database

The data that we are using was collected as part of the largest longitudinal, randomized control trial of psychotherapy for severely and stably distressed couples [2]. A total of 569 sessions consisting of 117 unique real married couples engaging in ten-minute problem solving interactions, in which they discussed a problem in their marriage. The corpus consists of manual word transcriptions, split-screen videos, and a single channel far-field audio of varying quality across sessions. Furthermore, both spouses were evaluated with 33 session-level codes using two standard manual codings, the Social Support Interaction Rating System (SSIRS) and the Couples Interaction Rating System (CIRS), with multiple trained evaluators. Details of this database are described in the previous work [1]. The audio data is automatically aligned with the word-level transcripts using the system *SailAlign* developed at SAIL [6]. *SailAlign* takes in the manual word transcriptions and automatically performs iterative word alignment to the whole session of audio data, and as a result, pseudo-turns are generated with speaker identifications (3 categories: husband, wife, unknown). These automatically aligned pseudo-turns are used as *speaking turns* in our research work. After this automatic process, 372 sessions out of 569 sessions are considered as *good* sessions to be used to conduct research because they meet the criteria of 5 dB signal-to-noise ratio (SNR) and at least 55% of the words within each session are successfully aligned.

The focus of this work is to utilize a quantification measure of vocal entrainment to predict extreme affective state (positive & negative), based on "Global Positive" and "Global Negative" code from SSIRS, of each spouse in married couples' interactions. Instead of working with the original 9-point scale rating, we transform it into a binary classification problem in which we take top 20% of positive and negative rated wife and husband. This results in a total of 140 (70 ratings on husband, 70 ratings on wife) sessions of *high positive* and 140 (70 ratings on husband, 70 ratings on wife) sessions of *high negative*. This accounts for a total of 280 interaction sessions with 81 unique couples, and it is the classification dataset of interest for this paper.

2.2 Multiple Instance Learning

In this section, we describe a brief summary of the multiple instance learning (MIL) in the context of predicting spouses' session-level affective states. MIL

is a learning algorithm to handle a classification situation where a label is associated with a bag consisting of multiple unlabeled instances rather than the more common scenario of associating a label with every training instance. The traditional formulation of MIL in a binary classification task (positive (1) vs. negative (0)) is that, a bag is labeled as positive if at least one instance in that bag is positive, and the bag is negative if only all instances are labeled as negative. A general way to solve the MIL problem is with the use of maximization of *Diverse Density* (DD) function [11] with respect to a concept point t – a point in the feature space that is close to at least one instance from every positive bag and far away from instances in negative bags. The maximization of the diverse density function ($DD(t)$) is defined as follows, where B_i^+, B_i^- denote the positive, negative bag, B_{ij}^+, B_{ij}^- denote the j^{th} instance in bag i, and assuming the bag label are indexed as logical 0 and 1.

$$\begin{aligned}
\underset{t}{\operatorname{argmax}}\, DD(t) &= \underset{t}{\operatorname{argmax}} \prod_i P(t|B_i^+) \prod_i P(t|B_i^-) \\
P(t|B_i^+) &= 1 - \prod_j (1 - P(B_{ij}^+)) \\
P(t|B_i^-) &= \prod_j (1 - P(B_{ij}^-)) \\
P(t|B_{ij}^+) &= \exp(-||B_{ij}^+ - t||^2)
\end{aligned}$$

Since there is no closed-form solution for this optimization, a gradient ascent method is often used to find the local maximum of DD function with respect to this concept point, t. In this work, we utilize a method called Expectation-Maximization Diverse Density (EM-DD) [15], in which the knowledge of which instance determines the label of the bag is modeled using a set of hidden variables that are estimated using the Expectation-Maximization framework.

The crucial assumptions of this MIL framework is that the single *most positive* instance in a bag determines whether a bag is positive, meaning if at least one of the instances in the bag has the probability of being positive > 0.5, the bag is labeled as positive. While there are variants in MIL that relax this assumption, this standard formulation is intuitively appealing. This approach, allows us to perform session-level affective code prediction without explicit labels at individual speaking-turns and also identify the most *salient* instance of vocal entrainment that determines the session-level affective code.

2.3 PCA-Based Signal-Derived Vocal Entrainment Measures

In this work, we used a feature that is based on the quantification of vocal entrainment for each spouse at the turn level using Principal Component Analysis (PCA) on vocal features. We have shown that this signal derived feature is indeed a viable quantification method of the often qualitatively-described vocal entrainment phenomenon [9] during interpersonal interactions. In order to compute the

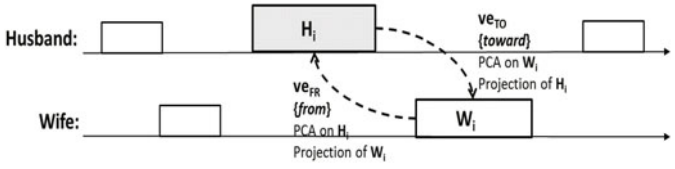

Fig. 1. Example of Computing Two Directions of Vocal Entrainment for Turns H_i

PCA-based vocal entrainment features, there are two major steps: first is to construct the speaking characteristics space, and second is to project the other speaker's vocal features on this constructed vocal characteristic space. There can be two directions (*toward*, *from*) of vocal entrainment in a dyadic interaction; for example, at a husband's turn seen in Figure 1, H_i, he can be entraining toward wife's speech, denoted as *toward*, and wife's speech can also becoming entraining toward this husband's speech segment, denoted as*from*. Figure 1 shows an example of computing the two directions of vocal entrainment, $ve_{\mathrm{TO}}, ve_{\mathrm{FR}}$, for husband's speech at turn H_i in an interaction session. The following is the list of steps in computing the husband's ve_{TO} at turn H_i:

1. Extract appropriate vocal features, X_1, to represent the husband's speaking characteristics at turn H_i.
2. Perform PCA on z-normalized X_1, such that $Y_1^T = D_1 X_1^T$.
3. Predefine a variance level ($v_1 = 0.95$) to select L-subset of basis vectors, D_{1L}.
4. Project the z-normalized vocal features, X_2 extracted from wife's speech at turn W_i, using D_{1L}.
5. Compute the vocal entrainment measure as the ratio of represented variance of X_2, in W_{1L} basis, and the predefined variance level in step 3.

In order to compute the husband's ve_{FR} at turn H_i, we just need to swap X_1 with X_2 of the above steps. The vocal features representing speaking characteristics include polynomial stylization of pitch contour and statistical functionals computed for energy and MFCC per word. Detailed of this data-driven PCA-based vocal entrainment measure can be found another of our paper on analyzing vocal entrainment in marital communication [9].

2.4 MIL Classifiers Setup and Selection

As discussed in Section 2.2, if we treat a *high positive* emotion label as the *positive bag* in the standard MIL framework, a salient vocal entrainment will predict positive emotion, and lack of such salient vocal entrainment will predict negative emotions. It is desirable to retain this framework of understanding saliency especially because the features themselves carry meaningful insights into the analysis of couples' interactions. However, we have made two modifications in an effort to relax the assumption of the standard MIL framework.

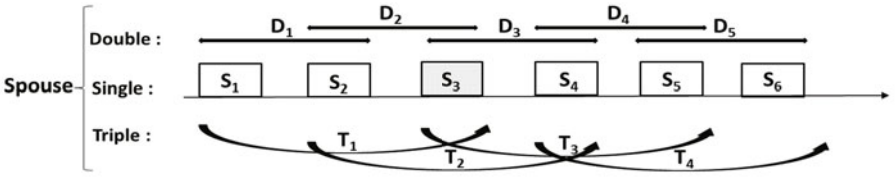

Fig. 2. A Diagram of MIL Classifier Setup

- While a salient instance can be the deciding factor of predicting *high positive* emotion, at the same time, conversely, a salient instance of *high negative* emotion may be the most informative predictor of *high negative* emotion. Hence, for a given fixed set of features, we train two MIL classifiers with bag labels reversed (0 becomes 1, and 1 becomes 0).
- While a *single* turn window of vocal entrainment features can be used to predict the affective state, multiple-length windows of vocal entrainment can be more informative predictors. Hence, we train three MIL classifiers, *single, double, triple* using entrainment features with an instance span of up to three turns.

Figure 2 shows a schematic diagram of the classifiers set up described above: for predicting a spouse's affective rating (a bag), three different turn lengths (defining an instance) of features will be used (shown as *single, double triple*), and within each instance three different types of vocal entrainment features will be used ($\{toward\}$, $\{from\}$, $\{toward\ \&\ from\}$). For each of this MIL classifier, we swap the bag labels as described above. This in total generate $(3 \times 3 \times 2 = 18)$ 18 classifiers. We select the one best performing MIL classifier based on maximum accuracy and mutual information through cross validation within training. As an example of how this framework can be used to locate saliency is referred in Figure 2. Assuming for a testing couple, through this MIL classifier setup and selection, the best performing classifier is based on *single* vocal entrainment features. The selected classifier predicts the spouse as having *high positive* emotion code because $\mathbf{S}_3$ is a salience instance determining the overall bag label of *positive emotion* of the spouse. This saliency, $\mathbf{S}_3$, turn in the session can then be treated as a region of interest for further analyzing for entrainment occurrence.

3 Experiment Setups and Results

Two different experiments were set up to evaluate the effectiveness of the MIL classification framework described in Section 2 in recognizing each spouse's affective state (*high positive* vs. *high negative*). In all of the experiments, the evaluation scheme was based on leave-one-couple-out cross validation (81 folds). The selection of the *best* classifier to be used in performing recognition for each testing fold was based on the leave-one-couple-out cross validation (80 folds) within training. The evaluation metric is the percentage (%) of accurately recognized spouse' affective ratings. EM-DD was trained using the MILL [14] software.

- **Experiment I :** Evaluate the performance of the proposed MIL classification scheme compared to different methods of selecting the *best* classifiers out of the 18 MIL classifiers.
- **Experiment II:** Evaluate the performance of the proposed MIL classification scheme compared to using different lengths of turns as features for each instance.

3.1 Experiment I Setup

The main assumption underlying Experiment I is that since there can be a large variability between each couple as they interact in their own norm, a different classifier is required to perform the recognition of the affective state of each spouse. As described in Section 2.3, 18 different classifiers were trained, and the proposed method of classifier selection was based on maximum accuracy with mutual information as tie breaker. In this experiment, we compared the proposed method with the following classifier selection technique:

- *Baseline*: SVM-based classifier using nine statistical functionals (mean, variance, range, maximum, minimum, 25% quantile, 75% quantile, interquartile range, median) computed per session of one specific type of entrainment measures (*toward*).
- *Voting*: Majority voting on 18 classifiers' results to assign the session label.
- *Same* MIL classifier for entire training folds: pick the classifier that occurs the most (out of 81 folds), in which it obtains the maximum accuracy in each fold. (Denoted in Table 1 by ONE_{maxA}).
- *Different* MIL classifiers for each training fold: pick the classifier according to our selection criterion (Denoted in Table 1 by Proposed).

3.2 Experiment I Result

A summary of the result of Experiment I is in Table 1. Several observations can be made with the results from Experiment I. First is that the absolute recognition rate is not very high, which may have been due to the fact that there is essentially only one type of feature used in this work. It is promising to see that our proposed method based on MIL framework achieves a 3.93% absolute (7.86% relative) improvement over chance and a 2.14% (4.13% relative) improvement over the

Table 1. Summary of Experiment I Result

Classifier	Accuracy (%)
Chance	50.00%
Baseline SVM	51.79%
Majority Voting	44.64%
ONE_{maxA}	50.71%
Proposed	**53.93%**

baseline model although significance testing using difference of proportion does not show the result significantly higher than the baseline at the $\alpha = 0.05$ level.

The results indicate that these signal-derived vocal entrainment measures possess some discriminative power in recognizing the spouse's affective state. Second, the baseline model used nine commonly-used statistical functionals computed at the session level with a SVM (radial basis function kernel), which serves as baseline of *not* using salient instances for prediction. While the comparison is not perfectly fair without optimizing SVM parameters, it is still encouraging that we obtain a better recognition accuracy through this saliency-based classification scheme. Furthermore, the majority voting over 18 classifiers obtains a performance less than chance indicates that a single optimized classifier per interacting couple is essential in achieving better accuracy.

3.3 Experiment II Setup

In this experiment, we examined the idea that a *salient* instance can occur at longer intervals, here measured in number of consecutive turns. The maximum number of turns considered for training of the MIL classifier was three, determined empirically. We compare the the three sets of classifiers (*Single*, *Double*, *Triple* as shown in Figure 2 described in Section 2.4 with the proposed combined set of classifiers.

- *Single*: Includes classifiers trained on vocal entrainment features only for one turn for a given rated spouse.
- *Double*: Includes classifiers trained on vocal entrainment features only for two turns for a given rated spouse.
- *Triple*: Includes classifiers trained on vocal entrainment features only for three turns for a given rated spouse.
- *Proposed*: Includes all 18 classifiers

3.4 Experiment II Result

Table 2 shows the accuracies of different sets of classifiers.

There are a few points to be noted by observing Table 2. When we considered the *Single*, *Double*, *Triple*-only classifier set, none of the classifier sets by itself is

Table 2. *Summary of Experiment II Results.*

Classifier	Accuracy (%)
Chance	50.00%
Baseline SVM	51.79%
Single	47.50%
Double	45.36%
Triple	47.86%
Proposed	**53.93%**

able to obtain a recognition rate better than chance. However, by considering all three sets of classifiers to perform classifier selection, we achieve the best performance. This bears implication that *salient* instances do span multiple turns, and that they differ from couple to couple. In fact, out of 81 folds cross-validation, 38 folds (46.91%) used *Single* classifier set, 31 folds (38.27%) used *Double* classifier set, and 12 folds (14.81%) used *Triple* classifier set. Experiment II shows that it is beneficial to consider multiple turns of vocal entrainment in this MIL framework to perform the binary classification on affective states of the spouses when working with disjoints set of couples in training and testing.

From Table 1 and Table 2, it is understood that the PCA-based vocal entrainment feature of itself may not be optimal to provide the discriminant power in recognizing affective state. In fact, the study [4] shows that higher-level of entrainment may not actually always correspond to the positive outcome of couples interaction, and it indeed is a complex relationship between entrainment and emotions. It is, however, still encouraging to observe that through the use of vocal entrainment measures with this proposed MIL framework, we are able to retain the original concept of identifying saliency within an interaction and also, through a couple of intuitive modifications, obtain an improved recognition accuracy over the baseline model. Further analysis on what specific roles or any temporal characteristic of these identified *salient* vocal entrainment episodes play in terms of affective states of the married couples will be very important.

4 Conclusion and Future Work

Experts in psychology often derive a set of subjective attributes to annotate specific behavior patterns to understand human interaction dynamics; this process of behavioral coding is an important human analytical instrument. Machine pattern recognition in behavioral signal processing aims to learn from a set of objective signals/cues in order to predict abstract human mental states. In this work, we focus on a specific method in bridging between these two aspects – human and machine analysis of behavior patterns– using a signal-derived vocal entrainment measure to represent an attribute of interest. We apply a multiple instance learning method to learn the salient portion of this attribute in predicting the session-level affective rating of each spouse in a dataset of married couple interactions. A classification scheme based on MIL is used in this paper to utilize salient instances of vocal entrainment measures of variable turn lengths to obtain the best performing classification accuracy. As described in Section 3, through the usage of this MIL framework, we achieve an overall 53.93% recognition rate.

This work has many possible future directions. The first is to gain further improvements through inclusion of other interpretable and meaningful attributes with signal-derived approximations, such as arousal level, expected interacting behaviors and possible deviation from the normal behaviors. This combined with the proposed MIL framework will be essential to provide the region of interest in an interaction and highlight the various attributes that experts in psychology can concentrate to perform their analysis. Furthermore, this MIL framework

selects *one* best classifier for each testing couple from the assumption that salient instances happen at variable turn lengths. This can be further relaxed to devise a technique to incorporate such notion in the MIL framework itself instead of performing classifier selection. Pin-pointing *salient* instances, especially in the framework of predicting using meaningful attributes describing an interaction, is essential in the detailed study of deeper understanding of various observation-driven codes that describe human interaction dynamics.

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

References

1. Black, M.P., Katsamanis, A., Lee, C.C., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S.: Automatic classification of married couples' behavior using audio features. In: Proceedings of Interspeech (2010)
2. Christensen, A., Atkins, D., Berns, S., Wheeler, J., Baucom, D.H., Simpson, L.: Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. J. of Consulting and Clinical Psychology 72, 176–191 (2004)
3. Eldridge, K., Baucom, B.: Couples and consequences of the demand-withdraw interaction pattern. In: Positive Pathways for Couples and Families: Meeting the Challenges of Relationships. Wiley-Blackwell (in press)
4. Gottman, J.M.: 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(1), 6–15 (1993)
5. Grimm, M., Kroschel, K., Mower, E., Narayanan, S.: Primitives-based evaluation and estimation of emotions in speech. Speech Communication 49(10-11), 787–800 (2007)
6. Katsamanis, A., Black, M.P., Georgiou, P.G., Goldstein, L., Narayanan, S.S.: SailAlign: Robust long speech-text alignment. In: Very-Large-Scale Phonetics Workshop (January 2011)
7. Lee, C.C., Black, M.P., Katsamanis, A., Lammert, A.C., Baucom, B.R., Christensen, A., Georgiou, P.G., Narayanan, S.S.: Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proceedings of Interspeech (2010)
8. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13(2), 293–303 (2005)
9. Lee, C.C., Katsamanis, A., Black, M.P., Baucom, B.R., Georgiou, P.G., Narayanan, S.S.: An analysis of PCA-based vocal entrainment measures in married couples' affective spoken interactions. In: Proceedings of Interspeech (2011)
10. Margolin, G., Oliver, P., Gordis, E., O'Hearn, H., Medina, A., Ghosh, C., Morland, L.: The nuts and bolts of behavioral observation of marital and family interaction. Clinical Child and Family Psychology Review 1(4), 195–213 (1998)
11. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 570–576 (1998)

12. Rozgić, V., Xiao, B., Katsamanis, A., Baucom, B., Georgiou, P.G., Narayanan, S.S.: Estimation of ordinal approach-avoidance labels in dyadic interactions: Ordinal logistic regression approach. In: ICASSP (2011)
13. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L.: The relevance of feature type for automatic classification of emotional user states: Low level descriptors and functionals. In: Proceedings of Interspeech (2007)
14. Yang, J.: MILL: A multiple instance learning library, `http://www.cs.cmu.edu/~juny/MILL/index.html`
15. Zhang, Q., Goldman, S.: EM-DD: An improved multiple-instance learning technique. In: Advances in Neural Information Processing Systems, vol. 2, pp. 1073–1080 (2002)