INTERSPEECH 2009 BRIGHTON

Modeling Mutual Influence of Interlocutor Emotion States in Dyadic Spoken Interactions

Chi-Chun Lee, Carlos Busso, Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA 90089, USA
chiclee@usc.edu, busso@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

In dyadic human interactions, mutual influence - a person's influence on the interacting partner's behaviors - is shown to be important and could be incorporated into the modeling framework in characterizing, and automatically recognizing the participants' states. We propose a Dynamic Bayesian Network (DBN) to explicitly model the conditional dependency between two interacting partners' emotion states in a dialog using data from the IEMOCAP corpus of expressive dyadic spoken interactions. Also, we focus on automatically computing the Valence-Activation emotion attributes to obtain a continuous characterization of the participants' emotion flow. Our proposed DBN models the temporal dynamics of the emotion states as well as the mutual influence between speakers in a dialog. With speech based features, the proposed network improves classification accuracy by 3.67% absolute and 7.12% relative over the Gaussian Mixture Model (GMM) baseline on isolated turn-by-turn emotion classification.

Index Terms: emotion recognition, mutual influence, Dynamic Bayesian Network, dyadic interaction

1. Introduction

In dyadic (two person) human-human conversation, the interactions between the two participants have shown to exhibit varying degrees and patterns of mutual influence along several aspects such as talking style/prosody, gestural behavior, engagement level, emotion, and many other types of user states [1]. This mutual influence guides the dynamic flow of the conversation and often plays an important role in shaping the overall tone of the interaction. In fact, we can view a dyadic conversation as two interacting dynamical state systems such that the evolution of a speaker's user state depends not only on its own history but also the interacting partner's history. This modeling will not only allow us to capture interactants' user states more reliably, but it could also provide a higher level description of the interaction details, such as talking in-sync, avoidance, or arguing.

The increasing sophistication of automatic meeting and dialog analysis due to the advances in audio-visual technologies, modeling emotion evolution has since become an important aspect of dialog modeling. Emotion evolution is related to people's perception on the overall tones of interaction, and it can also be used to identify salient portions in a conversation. Further, if we can better model the mutual influence during interaction, we could bring insights into designing communication strategy for a human-machine interaction agent to promote efficient communication. In this paper, we propose and implement a model describing the evolution of emotion states of the two participants engaged in dyadic dialogs by incorporating the idea of mutual influence during interaction.

Emotion can be represented by three-dimensional attributes as presented in [2]: (V)Valence: positive - negative, (A)Activation: aroused - calm, (D)Dominance: strong - weak, with each attribute associated with a numerical value indicating the level of expression. In our model, we focus on Activation and Valence dimension only. This dimensional representation offers a general description of the emotion, and it provides a natural way for describing dynamic emotion evolution in a dialog since not all utterances in a dialog can be easily labeled as a specific categorical emotion. Our approach contrasts with most of the previous emotion classification schemes that have primarily focused on utterance level recognition of categorical labels [3] or emotion attributes [4]. Others, such as proposed in [5] have used features that encode contextual information to perform emotion recognition. However, most of these works have neither considered decoding dynamic emotions through the dialog, nor have they incorporated the mutual influence exhibited between interactants in their models.

Because of its ability to model conditional dependency between variables within and across time, we utilize the Dynamic Bayesian Network (DBN) framework to model the mutual influence and temporal dependency of speakers' emotional states in a conversation. The experiment of this paper used the IEMOCAP database [6] since it provides a rich corpus of expressive dyadic spoken interaction. Also, detailed annotation of emotion is available for every utterance in the corpus. We hypothesize that by including cross speaker dependency and modeling the temporal dynamics of the emotion states in a dialog, we can obtain better emotion recognition performance and bring improved insights into mutual influence behaviors in dyadic interaction.

The paper is organized as follows. Our research methodology is described in Section 2. The experimental results and discussion are presented in Section 3. Conclusion and future work are given in Section 4.

2. Research Methodology

2.1. Database and Annotation

2.1.1. IEMOCAP Database

We use the IEMOCAP database [6] for the present study. The database was collected for the purpose of studying expressive dyadic interaction from a multimodal perspective. The designing of the database assumed that by exploiting dyadic interactions between actors, a more natural and richer emotional dis-

Table 1: *Emotion Label Clustering ($k = 5$)*

Emotion Cluster	Number of *Turns*	Cluster Centroid (V,A)
Class 1	1254	(2.19 , 3.29)
Class 2	1954	(3.15 , 3.14)
Class 3	2027	(4.06 , 2.21)
Class 4	1092	(1.89 , 2.25)
Class 5	2016	(3.84 , 3.55)

play would be elicited than in speech read by a single subject [7]. This data allows us to investigate our hypothesis about the mutual influence between speakers during spoken interaction. The database was motion captured and audio recorded in five dyadic sessions with 10 subjects, where each session consists of a different pair of male-female actors both acting out scripted plays and engaging in spontaneous dialogs. The analysis in this paper utilizes the recorded speech data from both subjects in every dialog available with speech transcriptions and emotional annotations. Three human annotations on categorical emotion labels, such as happy, sad, neutral, angry, etc, and two human evaluation of the three emotion attributes (Valence, Activation, Dominance) are available for every utterance in the database. Each dimension is labeled on a scale of 1 to 5 indicating different levels of expressiveness.

The database was originally manually segmented into *utterances*. But, to ensure that we have both speakers' acoustic information for a given analysis window in our dynamic modeling, we define a *turn change*, T , as one analysis window. Each T consist of two *turns*. Each *turn* is defined as the portion of speech belonging to a single speaker before he/she finishes speaking, and may consists of multiple original segmented *utterances*. Figure 1 shows an example that explains our definition.

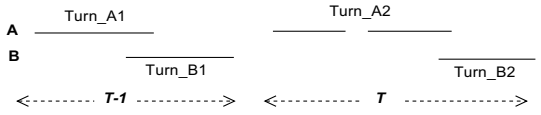

Figure 1: Example of Analysis Windows.

The example has two speakers, A and B, and a total of two analysis windows, T-1 and T, segmented. Speaker A is defined as the first person to speak in a dialog, and is always the starting point of any analysis window. A speaker can speak multiple *utterances* in a given *turn* as shown in Figure 1 of *Turn_A2*. Two *turns* - one from each speaker, denotes a *turn change*, which is defined as our one analysis window. Annotators were asked to provide a label for every *utterance* in the database. Since our basic unit is a *turn*, an emotional label is given to every *turn* as described in the following section.

2.1.2. Emotion Annotation

In this work, we focused on the Valence-Activation dimensions of emotion representation, since the combination of these two dimensions can be intuitively thought as corresponding to marking most of the conventional categorical emotions [4]. The dimension values for each *turn* is obtained by averaging the two annotated values. In order to further reduce the number of emotional states values, $5^2 = 25$, we cluster these two dimensions' values. Based on our empirical observation, we decided to group these two dimension values into five clusters using the K-Means clustering algorithm. Figure 2 shows our clustering output. Although this averaging may create quantization noise, from Figure 2, we can see that this process does provide reasonably interpretable clusters. For example, cluster 3 represented

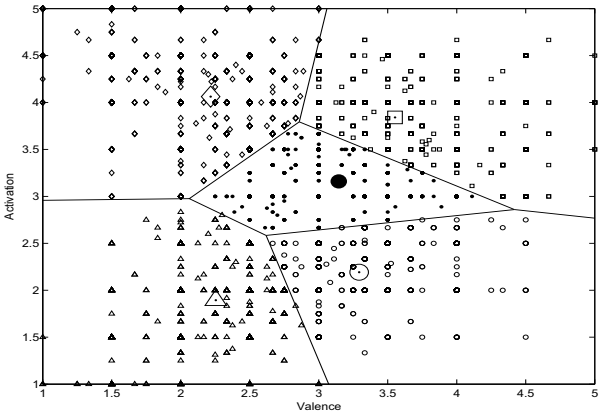

Figure 2: K-Means Clustering Output of Valence-Activation.

by diamond-shaped markers could be thought as corresponding to *angry* because of its concentration on lower values of valence with higher level of activation; in fact, about 70% of all angry utterances in the database where at least 2 annotators agree on, reside in cluster 3. Cluster 2 represented by point-shaped markers are centered at about the mid-range of Valence-Activation levels could be thought as *neutral* emotion, and about 51% of the neutral utterances of the database reside in this cluster. There are a total of 5 pairs of subjects in 151 dialogs consisting of 8343 *turns* used in this paper. A table showing the distribution of *turns* for each emotion cluster and its clustering centroid is given in Table 1.

2.2. Dynamic Bayesian Network Model

Dynamic Bayesian Network (DBN) is a statistical graphical modeling framework, where each node in a network is a random variable and the connecting arrows represent the conditional dependency between random variables. Since we want to capture the time dependency and mutual influence between speakers' emotion states, we propose to use the DBN structure shown in Figure 3.

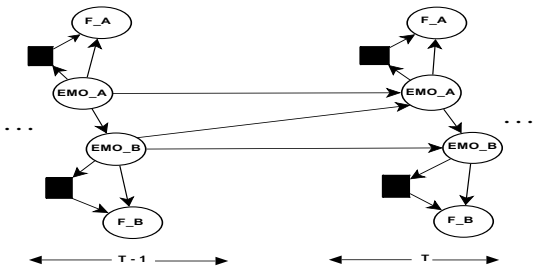

Figure 3: Proposed Dynamic Bayesian Network Structure.

In Figure 3, the EMO_A and EMO_B nodes represent the emotional class label for speakers A, B in the dialog, and the F_A and F_B nodes represent the respective observed acoustic information modeled by Mixture of Gaussian Distribution; the black rectangle represents the hidden mixture weights for the GMM. The proposed network tries to model two aspects of emotion evolution in an interaction. One is the time dependency of the emotion evolution, where a person's emotion state is conditionally dependent on his/her previous emotion state modeled as a *first order Markov process*. Second, the model incorporates the mutual influence between the two speakers in the dyadic interaction, where one speaker's emotion state is affected by the interacting partner's emotion. The joint probability of emotion

states E_{Bt} and E_{At} and feature vectors Y_{Bt} and Y_{At} for a dialog under this model can be factored as shown in Equation 1.

$$P(\{E_{At}, Y_{At}\}, \{E_{Bt}, Y_{Bt}\}) = \quad (1)$$
$$P(E_{A1})P(Y_{A1}|E_{A1})P(E_{B1}|E_{A1})P(Y_{B1}|E_{B1}) \times$$
$$\prod_{t=2}^{T} P(E_{Bt}|E_{Bt-1})P(E_{Bt}|E_{At})P(Y_{Bt}|E_{Bt}) \times$$
$$\prod_{t=2}^{T} P(E_{At}|E_{At-1})P(E_{At}|E_{Bt-1})P(Y_{At}|E_{At})$$

3. Experimental Results and Discussion

3.1. Feature Extraction

We focused on acoustic cues for the modeling study in this paper. All features except speech rate were extracted using the Praat Toolkit [8], while speech rate was estimated as the number of phonemes per second obtained from ASR forced alignment output detailed in [6] . The following is the list of extracted features at the *turn* level as previously defined.

- F0 Frequency: Mean, Standard Deviation, Minimum, Maximum, 25% Quantile, 75% Quantile, Range, InterQuantile Range, Median, Kurtosis, Skewness
- Harmonic to Noise Ratio (HNR): Mean, Standard Deviation, Minimum, Maximum, 25% Quantile, 75% Quantile, Range, InterQuantile Range, Median, Kurtosis, Skewness
- Intensity/Energy: Mean, Standard Deviation, Minimum, Maximum, 25% Quantile, 75% Quantile, Range, InterQuantile Range, Median, Kurtosis, Skewness
- Speech Rate: Mean, Maximum, Minimum
- 13 MFCC Coefficients: Mean, Standard Deviation
- 27 Mel Frequency Bank Filter Output: Mean, Standard Deviation

This resulted in a 116-dimension feature vector. Furthermore, feature normalization was obtained by performing *z-normalization* on the feature vectors with respect to each individual speaker's neutral utterances. The rationale behind this normalization was that while individuals may express emotions differently, by normalizing with respect to neutral utterances, speaker-dependent emotional modulation should be more comparable across speakers.

3.2. Experiment Setup

- **Experiment I**: Recognize the 5 emotion classes described in Section 2.1.2
- **Experiment II**: Recognize only the Activation and Valence dimension (each with 3 classes) separately using the same proposed structure

Experiment II was performed to help us identify which of the emotion dimension is likely to be affected by mutual influence in an interaction. Here, each dimension was clustered again into 3 classes (High, Medium, Low) using the K-Means algorithm. Table 2 shows a summary of data distribution and centroid of emotion classes for Experiment II.

For both experiments, forward feature selection was performed with accuracy percentage as the stopping criterion to reduced the number of features. We then analyzed four different structures representing different aspects of emotional state evolution in a dialog. The four different structures considered

Table 2: *Valence & Activation Clustering ($k = 3$)*

	Valence		Activation	
	No. of *Turns*	Centroid	No. of *Turns*	Centroid
Low	2355	2.05	3096	2.21
Medium	3271	3.29	2525	2.97
High	2717	4.18	2722	3.69

are shown in Figure 4. The first structure (1) is our baseline model that does not incorporate any time or mutual influence dependency. Therefore, it recognizes each turn separately with trained GMM model using just the acoustic cues. Structure (2) incorporates time dependency of individual speaker's emotion without mutual influence from the interacting partner. Structure (3) models only the mutual influence between speakers, and Structure (4) is our proposed complete model that combined both time and cross-speaker dependencies.

We tied the GMM parameters of both speakers' observation feature vector for both experiments to maximize the use of training data. Each trained baseline GMM models' parameters was passed onto all three other structures to ensure any changes in classification accuracy is due to the change in emotion dependency structures. The model was implemented and tested using the Bayes Net Toolbox [9]. All experiments were done with 15-fold cross validation, where 140 dialogs were selected at training and about 10 dialogs were used as testing. The numbers of mixture for the GMM was determined empirically to be four. At training, emotional labels and feature vectors were provided to learn the mixture weights and conditional dependency between emotional states using the EM Algorithm with Junction Tree Inference. At testing, the trained network decoded both speakers' emotion labels by computing the most likely path of emotion state evolution throughout the dialog given the sequence of observations.

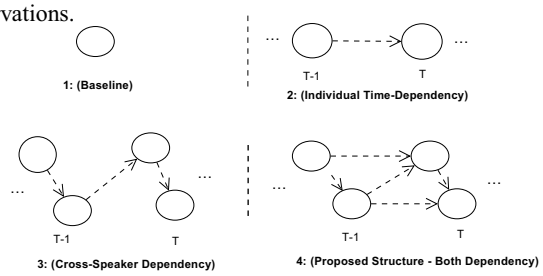

Figure 4: Structures of Emotion States Evolution.

3.3. Results and Discussion

The results of both experiments are summarized in Table 3. The performance measure used is the number of accurately classified *turns* divided by the total numbers of *turns* tested. Two different results are shown for the Experiment II. The *same* column in Table 3 means that the experiment was carried out using the same set features obtained from feature selection output in Experiment I, and the *optimized* column means that forward feature selection was performed on each of the Activation and Valence experiments separately.

In the Experiment I, the results show that it is beneficial to incorporate both time dependency and mutual influence on the emotional state, since both Structure (2) and (3) improve the classification performance. Our proposed DBN model which combined both dependencies obtained an absolute 3.67% increase in accuracy (relative 7.12% improvement) over our baseline model. To see where the improvement comes from, we can examine the results from Experiment II where the classification was performed on Valence-Activation dimension separately.

Table 3: *Summary of Experiment Accuracy Percentage*

DBN Structure	I: 5 - Emotion Classes	II: Activation-Only (3-Class)		II: Valence-Only (3-Class)	
		Same	*Optimized*	*Same*	*Optimized*
Chance	24.29%	37.11%	37.11%	39.21%	39.21%
Baseline - GMM (1)	51.53%	62.30%	**63.45%**	56.59%	59.89%
Time Dependency (2)	52.68%	62.02%	61.92%	59.78%	63.40%
Mutual Influence (3)	53.37%	**62.52%**	62.30%	59.60%	62.67%
Proposed Model (4)	**55.20%**	62.35%	62.49%	**61.26%**	**65.02%**

In Experiment II, the first thing to point out is that the classification accuracy using baseline GMM on the Valence and Activation separately shows that by exclusively using speech related features, the classification accuracy is higher with the Activation dimension than with the Valence dimension by absolute 5.71% (relative 10.01%). And this agrees with our knowledge about the discriminative power of acoustic features [10] in each of these dimensions. The second observation is that we improved classification accuracy in the Valence dimension by approximately 5% absolute (relative 8%) over baseline. However, the effect is not as observable with the Activation dimension. It appears that the advantage of this modeling comes primarily in the Valence dimension instead of the Activation dimension. We hypothesize that the mutual influence on interacting partners may be more significant in the Valence dimension. However, further analysis is necessary to verify this claim.

In summary, our proposed model, which captures both time dependency and mutual influence between speakers, was able to improve the overall classification accuracy. In spite of the limited amount of interaction data (151 dialogs with 10 subjects) with potentially noisy emotion classes, it is still encouraging to see that our model is able to capture these effects and improve the recognition results.

4. Conclusions and Future Work

Interpersonal interactions often exhibit mutual influence along different elements of the interlocutor behavior. In this paper, we utilized the Dynamic Bayesian Network (DBN) to model this effect to better capture the flow of emotion in dialogs. In turn, we use the model for performing emotion recognition in the Valence-Activation dimension. As shown in Section 3, it is advantageous to model the dynamics and mutual influence of emotion states in dialog for improving emotion classification.

There are two main limitations with this paper. The first arises because we only had two human annotations on emotion attributes for each utterance. In order to incorporate both annotations to serve as our ground truth, we took the average of two annotations values for every *turn*; this created noise in the emotion labels. We plan on acquiring more annotations in the future to alleviate this problem. The other limitation is that we just relied on speech based features for our modeling; fortunately, the IEMOCAP database has detailed facial and rigid head/hand gesture information as well as transcriptions providing the language information, all of which have been shown useful for emotion modeling, could be incorporated within the model in the future.

Several other future directions can be pursued. One immediate extension is to provide a mapping between decoded Valence-Activation state to some more human interpretable emotion categories, and extend this framework as a first stage processing for inferring higher-level dialog attributes. Further, mutual influence between speakers can happen at multiple levels. In this paper, we examined this effect through recognizing emotion states at the *turn* level. Prior works have shown mutual influence on lexical structure [11] and on predicting task success [12] at the dialog level. We can analyze this effect along such levels using hierarchical structures. Furthermore, we are in the process of obtaining other forms of interaction databases with both natural human interaction and acted interaction. Once we acquire better insights into mutual influence in human interactions, we not only will be able to improve dialog modeling, but may also be able to incorporate such information in the design of robust machine spoken dialog interfaces.

5. Acknowledgements

The paper was supported in part by funds from NSF, Army, and USC Annenberg Fellowship.

6. References

[1] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge University Press, 1995.

[2] K. Roland, "The prosody of authentic emotions," in *Speech Prosody Conference*, 2002, pp. 423–426.

[3] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using gaussian mixture models for face and voice," in *In Proceedings of IEEE International Symposium of Multimedia*, Berkeley, CA, December 2008.

[4] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, "Primitives based estimation and evaluation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, November 2007.

[5] J. Liscombe, G. Riccardi, and D. Hakkani-Tur, "Using context to improve emotion detection in spoken dialog systems," in *Interspeech*, 2005, pp. 1845–1848.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[7] C. Busso and S. Narayanan, "Recording audio-visual emotional database from actors: a closer look," in *Second Intl. Workship on Emotion: Corpora for Research on Emotion and Affect, Int'l conference on Language Resources and Evaluation*, May 2008, pp. 17–22.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.03) [Computer program]," March 2009. [Online]. Available: http://www.praat.org/

[9] K. P. Murphy, "The bayes net toolbox for matlab," *Computing Science and Statistics*, 2001.

[10] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. of the Int'l Conf. on Multimodal Interfaces*, October 2004.

[11] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL-08: HLT*, vol. Companion, no. 169-172, 2008.

[12] D. Reitter and J. D. Moore, "Predicting success in dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.