Interpreting Ambiguous Emotional Expressions

Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh,
Carlos Busso, Sungbok Lee, Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory (SAIL)
University of Southern California
University Park, Los Angeles, California, USA 90089
{mower|metallin|chiclee|kazemzad|busso|sungbokl}@usc.edu, shri@sipi.usc.edu

Abstract

Emotion expression is a complex process involving dependencies based on time, speaker, context, mood, personality, and culture. Emotion classification algorithms designed for real-world application must be able to interpret the emotional content of an utterance or dialog given the modulations resulting from these and other dependencies. Algorithmic development often rests on the assumption that the input emotions are uniformly recognized by a pool of evaluators. However, this style of consistent prototypical emotion expression often does not exist outside of a laboratory environment. This paper presents methods for interpreting the emotional content of non-prototypical utterances. These methods include modeling across multiple time-scales and modeling interaction dynamics between interlocutors. This paper recommends classifying emotions based on emotional profiles, or soft-labels, of emotion expression rather than relying on just raw acoustic features or categorical hard labels. Emotion expression is both interactive and dynamic. Consequently, to accurately recognize emotional content, these aspects must be incorporated during algorithmic design to improve classification performance.

1. Introduction

Positive reception of human-computer and human-robot interaction (HCI/HRI) technologies hinge on the ability of the system to accurately recognize emotions expressed by human users [12, 26]. While current categorical emotional classification techniques can achieve good performance for speech with well-defined emotional content, human interaction involves a complex range of mixed emotional manifestations. It is necessary to develop emotion classification schemes that can handle this variation and make sense out of perceived emotional information.

In natural human communication, emotions do not follow a static mold. They vary temporally with speech [6], are expressed and perceived over multiple modalities [13, 25, 34], may be inherently ambiguous [3, 14, 15], or may have emotional connotations resulting from other emotional utterances within a dialog [20]. A classification scheme designed to recognize only the subset of emotional utterances consisting of well-defined emotions will not be able to handle the natural variability in human emotional expression.

Conventionally, when training emotion recognition classifiers, researchers utilize emotional expressions that are rated consistently, by a set of human evaluators. These expressions are referred to as prototypical emotion expressions. This process ensures that the models capture the emotionally-relevant modulations. However, while analyzing natural human interactions, including in an online HCI or HRI application, one cannot expect that every human utterance will contain clear emotional content. Consequently, techniques must be developed to handle, model, and utilize these emotionally ambiguous, or non-prototypical, utterances within the context of HCI or HRI.

This paper presents three techniques to process and interpret non-prototypical emotional utterances: emotional interpolation, emotional profiling, and utterance-level hard labeling. *Emotional interpolation* is a technique in which the emotional label of a dialog is determined by interpolating between salient, prototypical utterances. The input features to this interpolation process are emotional profiles. *Emotional profiling* is a soft labeling technique in which the emotional content of an utterance is expressed in terms of the probability that the utterance is assigned to any of k-emotion classes. Emotional interpolation and profiling can be merged to detect changes in the dialog-level emotional state given both prototypical and non-prototypical emotion expressions and to form robust models of user emotion production and perception. However, in certain applications individual utterances must be emotionally categorized. This paper also presents methods to determine the *hard label* of an utterance using dialog-level modeling, fuzzy logic, and

978-1-4244-4799-2/09/$25.00

by comparing the benefits of utilizing individual-specific and generalized emotion models.

This paper presents analyses of the classification techniques that can be utilized to interpret non-prototypical emotion utterances. We show previously developed solutions and suggest directions for the development of future algorithms. We believe that the integration of these solutions will result in a more robust and human-like performance in automated emotional classification.

The remainder of this paper will discuss the methods to interpret prototypical and non-prototypical emotions. Section 2 will discuss the definitions and design relevance of prototypical and non-prototypical emotion expressions. Section 3 will discuss dialog modeling techniques used to interpret emotional content. Section 4 will discuss methods to assign hard labels to ambiguous emotional utterances. Finally, Section 5 will discuss conclusions.

2. Prototypical vs. non-prototypical emotions

2.1. Definitions

Emotional expressions and emotional classes exist on a spectrum ranging from prototypical expressions to non-prototypical expressions. Prototypical emotion expressions are utterances that are consistently recognized by a set of human evaluators. Examples of prototypical emotion expressions include hot anger or glee. Non-prototypical emotion expressions are expressions that are not consistently recognized. Non-prototypical emotion expressions occur naturally within human emotional communication.

Certain emotion classes are inherently ambiguous resulting from their broadly accepted emotional characteristics. For example, the emotion class of frustration has the potential to overlap with classes ranging from anger, to neutrality, to sadness [3]. Expressions of frustration can vary widely with respect to the level of activation in the voice of the speaker. Other classes, with more well-defined characteristics still may contain shades of non-prototypical expressions. Expressions of anger may range from annoyance to rage; love can be defined as a combination of joy and acceptance [14]. Individual evaluators may perceive these mixed presentations differently based on context, mood, culture, and/or personality. This leads to the assignment of these mixed-emotion utterances as non-prototypical emotions.

Non-prototypical emotions also result from mismatches between emotions expressed using the available modalities (e.g., the face and the voice). These mismatches may occur naturally in communication. Humans can convey inconsistent emotional messages through different modalities (e.g., sarcasm) [23]. This behavior has attracted the interest of psychology researchers [22]. These mismatches may also be created artificially to study perception [25]. The emotional evaluations resulting from these mismatches are different than those of either of the modalities individually, indicating that emotion perception is naturally multi-modal. Thus, non-prototypical emotions may also result from incomplete stimuli presentation (e.g., audio-only).

Non-prototypical emotions occur in human communication resulting from the natural variation in the strength of emotion expression over the course of a dialog. For example, in an angry dialog, not every utterance is expressed as unmistakable hot anger. Instead, during the course of a dialog, humans transition through flavors of anger (anger may range from annoyance to rage [14]), expressing the overall emotional meaning through the emotional context of dialog [9]. As a result, single utterances may not be individually recognizable as a single semantic emotional label [20].

The multi-faceted level of emotion expression characteristics does not indicate that only a subset of the expressions should be reconciled. Instead, it suggests that techniques must be developed in order to better comprehend the continuum of human emotional behavior.

2.2. Expression in data

The properties of prototypical and non-prototypical expressions are demonstrated using the Interactive Emotional Dyadic Motion Capture database (IEMOCAP database), collected at USC [3]. This database consists of approximately 12 hours of audiovisual data from five mixed gender pairs of actors (ten actors total: five male and five female).

The database collection utilized emotion elicitation techniques rooted in the core of acting training, including the use of scripts and improvisation of hypothetical scenarios. This contrasts with conventional approaches in which the subjects are asked to read sentences displaying specific stereotypical emotions. These techniques provide a viable research methodology for studying human emotions [7, 8, 16]. Instead of recording isolated sentences, the IEMOCAP contains entire dialogs (approximately five minutes long), manually segmented into utterances. The actors conveyed different emotions within a dialog, as dictated by the course of the improvisations or the scripts. The database contains both utterances with clear emotional content and utterances with mixed of emotions, similar to the emotional nuances found in real life scenarios. There are a total of 10,039 utterances within the database, 5,255 from the scripted sessions and 4,784 from the improvisational sessions. The average duration of an utterance is 4.5 seconds and there are an average of 11.4 words per utterance. For more details about this database, the readers are referred to [3].

The data were annotated sequentially by twelve naïve evaluators , who tagged overlapping subsets of the data. The categorical labels in the IEMOCAP database included: angry, happy, sad, neutral, frustrated, surprised, fearful, excited, disgusted, and other. This paper considers the sub-

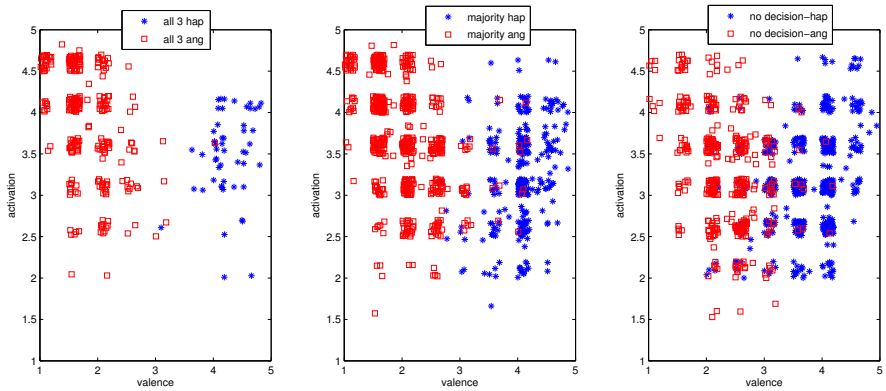

Figure 1. Valence and activation plots of angry and happy sentences. The leftmost graph presents prototypical angry and happy utterances, the center graph presents non-prototypical majority-voted utterances assigned either an angry or a happy label, the rightmost graph presents non-prototypical non-majority-vote utterances where angry or happy labels were selected by a single evaluator.

Emotional Category	Non-prototypical (no consensus)	Non-prototypical (consensus)	All 3 agree
Anger	802	604	497
Happiness / excitement	2095	1189	441
Neutrality	1623	1296	388
Sadness	616	618	465
Frustration	1383	1280	562

Table 1. Number of sentences in the IEMOCAP database from each emotion class and each prototypical and non-prototypical type. Note that the utterances may be tagged with more than one non-prototypical non-majority-vote label.

set of emotions tagged by at least one evaluator as: angry, happy, sad, neutral, frustrated, or excited. Table 1 presents the number of available sentences for each emotional category. The dimensional properties were rated on scales of 1–5 across the dimensions of valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs. aggressive). Each utterance was categorically evaluated by at least three evaluators and dimensionally evaluated by at least two evaluators. The final dimensional rating was an average (round up) of the evaluators' ratings.

Utterances in the IEMOCAP database with full evaluator agreement are defined as prototypical emotions (three out of three evaluators agree on the label). Non-prototypical emotions are defined as emotions with inconsistent evaluator agreement. The non-prototypical emotions are broken into two classes: non-prototypical emotions with majority-vote consensus (two out of three evaluators agree on the label) and without majority-vote consensus (one out of three evaluators selected a label).

Prototypical and non-prototypical emotions have different perceptual effects on evaluators. Emotional perceptual variance results from the natural range in human expression of categorical emotions (e.g., anger can vary from cold expression to hot rage). This natural variance can lead to perceptual confusion between categorical emotion classes as the differences between their manifestations decrease. In this paper the prototypical and non-prototypical labels are defined with respect to evaluator categorical labeling agreement (e.g., angry, happy, etc.). However, the differences between prototypical and non-prototypical emotions can be better visualized using valence-activation plots to understand the perceptual effects of the designations.

Previous research has demonstrated that categorical emotions can be plotted as positions with the valence-activation space [10, 17, 25, 28]. For example, the emotions of happiness and anger are differentiated primarily by the valence in the signal. One would expect that prototypical angry and happy emotions would occupy disparate regions in the valence-activation space. Figure 1 presents the valence-activation dimensional evaluations (plus a small amount of uniform noise for visualization purposes) of the IEMOCAP utterances labeled by evaluators as either angry or happy. In Figure 1, the leftmost plot represents prototypical angry and happy utterances. In this plot, the two categorical emotion classes clearly occupy different regions of the valence-activation space. When non-prototypical majority-voted emotion expressions are considered (center plot in Figure 1) it is clear that the dimensional perceptions of the emotional expressions are no longer as strongly differentiated. This trend is also witnessed in non-prototypical non-majority-vote utterances (rightmost figure of Figure 1). These non-prototypical expressions exhibit a large amount of dimensional perceptual overlap. These graphs demonstrate that non-prototypical emotions have more overlap in the valence-activation domain, suggesting that they are less perceptually differentiable than prototypical emotions.

2.3. Classification differences between prototypical and non-prototypical emotions

Non-prototypical emotions have increased perceptual variance with respect to prototypical emotions. Classification techniques designed to recognize the emotion

Original	Classified emotion				
	Ang	Hap	Neu	Sad	Fru
Ang	**69.16**	7.49	4.85	5.29	13.22
Hap	25.93	**19.21**	14.58	20.60	19.68
Neu	1.46	6.43	**35.09**	41.23	15.79
Sad	0.45	3.37	8.31	**82.25**	5.62
Fru	18.70	13.36	15.46	22.52	**29.96**

Table 2. HMM recognition results for training & testing on prototypical emotions (accuracy = 47.34%)

Original	Classified emotion				
	Ang	Hap	Neu	Sad	Fru
Ang	**53.45**	8.74	9.41	10.08	18.32
Hap	24.46	**18.86**	17.09	24.07	15.52
Neu	9.19	11.72	**31.85**	34.65	12.60
Sad	4.05	4.75	12.85	**71.13**	7.22
Fru	26.58	12.71	17.87	20.89	**21.96**

Table 3. HMM recognition results for training & testing on non-prototypical emotions (accuracy = 34.28%)

expressed in prototypical utterances do not perform as well when recognizing non-prototypical emotion expression. This is demonstrated using five Hidden Markov Model (HMM) [27] training-testing scenarios: 1) training and testing on prototypical emotions, 2) training and testing on non-prototypical majority-vote emotions, 3) training and testing on prototypical and non-prototypical majority-vote emotions, 4) training on prototypical and testing on non-prototypical majority-vote emotions, and 5) training on non-prototypical majority-vote and testing on prototypical emotions. No testing was performed on the non-prototypical non-majority-vote emotions due to the difficulty in establishing an appropriate ground truth.

The HMMs were trained and tested using the data from all 10 speakers from the emotion classes of angry, happy, excited, sad, neutral, and frustrated. The classes of happiness and excitement were merged due to data sparsity. The utterances were classified at the phoneme-level using three-state, eight-mixture models implemented using the Hidden Markov Model Toolkit (HTK) [35]. The phonemes were clustered a priori into seven broad phonetic categories consisting of: front vowels, back/mid vowels, diphthong, liquid, nasal, stop consonants, and fricatives. A detailed mapping can be found in [4]. The acoustic feature vector consisted of the first thirteen Mel Filter Bank Coefficients (MFB), their deltas, and acceleration. MFBs have been demonstrated to contain more emotional information than Mel Frequency Cepstral Coefficients (MFCC) [4]. The utterances were tested using forced alignment to known transcripts. The models estimated the emotion of a given phoneme, rather than the emotion and the phoneme. The final class label was assigned using majority-voting. The accuracy is reported for each of the five training-testing scenarios. The results were validated using speaker-dependent five-fold cross-validation.

Original	Classified emotion				
	Ang	Hap	Neu	Sad	Fru
Ang	**55.33**	12.04	6.89	12.78	12.96
Hap	22.37	**21.87**	16.48	28.95	10.34
Neu	6.25	14.46	**28.20**	42.53	8.57
Sad	2.38	6.00	9.43	**78.29**	3.90
Fru	21.34	17.38	16.94	28.69	**15.65**

Table 4. HMM recognition results for training and testing on prototypical and non-prototypical emotions (accuracy = 35.06%)

Original	Classified emotion				
	Ang	Hap	Neu	Sad	Fru
Ang	**38.01**	10.73	11.51	17.67	22.08
Hap	17.18	**17.96**	19.07	31.10	14.69
Neu	5.75	11.13	**27.93**	41.37	13.82
Sad	1.98	4.30	16.20	**71.74**	5.79
Fru	17.31	12.67	20.61	29.90	**19.51**

Table 5. HMM recognition results for training on prototypical and testing on non-prototypical emotions (accuracy = 30.04%)

Original	Classified emotion				
	Ang	Hap	Neu	Sad	Fru
Ang	**73.57**	12.33	1.98	5.73	6.39
Hap	32.64	**21.06**	11.11	27.08	8.10
Neu	8.19	15.20	**25.44**	46.20	4.97
Sad	2.92	6.97	7.64	**77.75**	4.72
Fru	31.11	17.56	12.98	25.00	**13.36**

Table 6. HMM recognition results for training on non-prototypical and testing on prototypical emotions (accuracy = 42.24%)

The first training-testing scenario (prototypical data only) represents the common laboratory practice of training only on well-recognized emotional utterances. In this scenario the accuracy is 47.34% (Table 2). In the second scenario, the HMMs were trained and tested on non-prototypical majority-vote utterances. The classification accuracy decreases to 34.28% (Table 3). In the third scenario, the HMMs were trained and tested on prototypical and non-prototypical majority-vote utterances. The overall accuracy is similar at 35.06% (Table 4). In the fourth scenario, the models were trained on prototypical utterances and tested on non-prototypical emotions majority-vote utterances. This scenario represents the disconnect between creating models based on well-recognized emotions and testing based less constrained expressions. In this scenario the accuracy further decreases to 30.04% (Table 5). In the final scenario, the HMMs were trained on non-prototypical majority-vote utterances and tested on prototypical utterances. The accuracy decreases of this classification method is 42.24% (Table 6). This result is lower than the first scenario of training and testing on prototypical utterances, only. It should be noted that the size of the training set for the final scenario is greater than that of the first scenario, although the performance is still lower. The trends in the results of the five classification scenarios demonstrate the importance of recognizing the prototypical or non-prototypical nature of an emotional utterance. The results suggest that

it may be detrimental to classify non-prototypical utterances using only models created for prototypical emotional speech. This has also been demonstrated in [32].

It is important to note that although the performance across the three training-testing scenarios varies, the trends in the results remain similar. In all conditions, anger and sadness were most well recognized, while happiness and frustration remained obfuscated. The high recognition accuracy of the sadness classification may be due in part to the type of information conveyed within the audio channel. Audio information tends to convey activation, while additional modalities, such as video, are needed to convey valence [18]. Audio emotional models can effectively recognize sadness since it can be differentiated from the other four emotion states by its activation. The plots in Figure 1 suggest that anger and happiness should be easily differentiated due to their separation along the valence axis. However, as previously stated, vocalizations primarily provide activation information. The level of activation in both emotions is similar. Thus, both the poor performance of the happiness classification and the confusion that exists between the classes of happiness and anger, are expected. Furthermore, previous work has demonstrated that the happiness expressions in this database are difficult to recognize using only speech information [23]. The trends in the results presented here were also observed in [31].

3. Dialog modeling

3.1. Emotional interpolation and profiling

Emotion recognition in HRI and HCI is hampered by the non-constant nature of human emotion expression. However, humans have developed methods to identify salient stimuli of interest. Component Process Theory [29, 30] states that in the presence of a salient and relevant stimulus, an organism synchronizes its major subsystems (physiological regulation, cognition, monitoring-feeling, motor expression, and motivation) to regulate behavior during, "emotion episodes," while absent such a stimulus, the organism's subsystems function relatively independently [1]. This suggests that humans can and do differentiate between emotionally salient and non-salient utterances. This work proposes modeling this ability in an *emotional interpolation* framework. For example, in a dialog there may be a set of sentences leading an evaluator or interlocutor to perceive the entire dialog as angry. However, not all sentences may be unambiguously recognized. In this case, it may be unnecessary to further classify the emotionally ambiguous utterances. Instead, it may be enough to note that the emotion content of these utterances is not sufficiently different to discount the emotion content of the dialog as a whole.

Emotional interpolation requires that the evaluator or interlocutor be able to identify whether or not an ambiguous, or non-prototypical, utterance contains information that differentiates it significantly from the emotion of the dialog as a whole. Emotional profiling is a computational technique that describes the confidence with which an emotional label is assigned to an utterance. It quantifies the probability that an utterance is assigned to any of k-emotion states such that the sum of the probabilities is equal to one. Here, the soft label is representative of the classification output, rather than the k-evaluator estimates of [33]. The profile of an unambiguous, or prototypical, emotion expression would have an emotional profile approximating a delta function, indicating that the utterance is assigned an emotional label with probability approaching one. The profile of an ambiguous, or non-prototypical, emotion expression would most likely take one of two forms (or a combination of the two), a uniform distribution over all emotions or spikes at one emotion and neutrality. The uniform distribution (i.e., the probability of each emotion approaches $\frac{1}{k}$) representation would indicate that the assignment of the utterance to any emotion class is approximately equally likely. A profile with peaks at a particular emotion and neutrality would indicate that the emotional expression is subtle or subdued.

The profiles of ambiguous emotional utterances allow HCI/HRI technologies to develop an idea of the emotional content and context, without necessitating a noisy hard assignment. In an ambiguous utterance, this allows for the retention of emotional information when it otherwise might be lost in a hard emotion assignment. Furthermore, in an emotional dialog where only a subset of the utterances are emotionally salient, this profile approach provides a computational technique to determine if the emotional tenor of the dialog is altering or staying within the bounds suggested by the emotional labels resulting from salient utterances.

Emotional utterance profiles describe the content of an emotional utterance expressed in terms of the basic emotions present within that utterance. These profiles can be leveraged to create general and user-specific models. These user models enable dynamic emotion tracking within the course of an interaction. Dynamic emotion tracking is the process of tracking the change in the emotion expression of a user over the course of an interaction. This ability would allow HCI/HRI technologies to form better estimates of user state and to create better input-output models between system actions and user state. Emotional profiling techniques can also be used to create user-specific models of emotion perception. Emotion perception, like production and expression, is person dependent. Emotion profiles allow the system to create a detailed estimate of how a user will perceive a presented emotion utterance.

3.2. Computational framework

Dialog modeling of emotional profiles is a layered approach that can be used to model emotional expressions

Time	Content	Recognized Emotion
0 – 75	silence	–
76 – 98	things	sad
99 – 116	just	frustrated
117 – 128	aren't	frustrated
129 – 143	what	neutral
144 – 155	they	neutral
156 – 195	seem	sad
196 – 313	silence	–

Profile				
Angry	Happy	Sad	Neutral	Frustrated
–	–	52.58%	22.41%	25.00%

Table 7. An emotional profile calculation, the true utterance-level label is "sad"

at multiple time scales by taking advantage of the inherent interdependency between the manifestations of emotion over varying time scales. Such units of time may include: phoneme, syllable, word, phrase, utterance, turn, subdialog/topic, dialog, etc. The style of emotion expression is non-constant over these time units. Segments of expression may be highly prototypical or non-prototypical as a dialog progresses. Emotional profiles can be used both to estimate this prototypical ebb and flow and to identify "relevance sections," sections which contain emotional content that biases an evaluator to assess a dialog in a certain manner. Computational techniques that incorporate this dynamic modeling and relevance detection can be used to automatically assess emotion expression at any of the modeled time scales using either the causal or non-causal relationships that exist between the scales.

Emotional profiles can be computed at the utterance level, for example, directly from the phone-level HMM classification results obtained by training on only prototypical emotions (Section 2.3). After classification, the output of the recognizer consists of emotional labels for each phoneme. Emotional labels can include any label x $\in \{$angry, happy, sad, neutral, frustrated$\}$. The profiles are created by calculating the percentage of time during which each emotion is represented within the utterance. See Table 7 for an example of an emotional profile calculation.

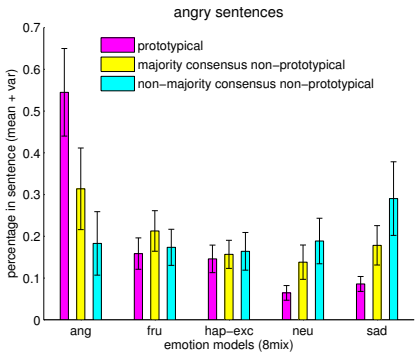

Figure 2. Emotional profiles for all utterances labeled as "angry" by at least one evaluator)

Figure 2 presents the mean and variance of the profiles for the perceptual ground truth of anger. The three bars represent training on prototypical emotions and testing on: prototypical emotions, majority-vote consensus non-prototypical emotions, and non-majority-vote consensus non-prototypical emotions. This depiction of the emotional profiles demonstrates the increasing difficulty in classifying ill-defined emotional utterances. As the emotion expressions become increasingly non-prototypical, the representation of the five emotional classes within the recognition results approaches the uniform distribution. The increasingly uniform nature of an emotional profile suggests that the emotional utterance is non-prototypical and therefore may not impact the dialog-level emotional perception of the evaluators as strongly as prototypical emotional utterances. Emotional profiles are also beneficial because they express the confidence with which each of the emotion labels is assigned to an utterance. These profiles inherently provide more information than a single hard label, allowing the system to utilize the information inherent even in a misclassification. Consequently, when classifying a meta emotion state, such as the dialog-level emotion, it may be beneficial to utilize the emotional profiles for classification, rather than the acoustic feature properties or a majority-vote over the estimated hard labels.

Dynamic interaction modeling at the dialog level captures the influences that exist between interlocutors during an interaction and the temporal characteristics of emotion evolution internal to each individual. During a multiperson interaction, an individual's emotion state changes as a function of the emotion state of his interlocutors [2]. This change may occur slowly, permitting a convergence that promotes a more efficient or positive interaction atmosphere; or the emotion state may diverge drastically from that of the others. Independent of the gradient of emotional change, an individual's emotion flow tends to progress smoothly within any two overlapping windows in emotional portions of the dialog. This temporal smoothness of emotional flow should also be considered during the modeling.

Dynamic interaction modeling has been shown to improve emotion classification accuracy. In [20] a Dynamic Bayesian Network (DBN) structure is proposed to capture the time dependency of emotion evolution and mutual influence between speakers' emotion states in a dyadic spoken dialog (Figure 3). In Figure 3, the EMO_A and EMO_B nodes represent the emotional class label for speakers A, and B in the dialog. The F_A and F_B nodes represent the respective observed acoustic features modeled by Mixture of Gaussian Distributions. The proposed network modeled two aspects of emotion evolution in a dialog, the time dependency and mutual influence. The time dependency of the emotion evolution models the conditional dependence

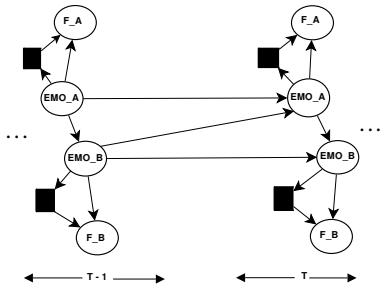

Figure 3. A DBN modeling an emotional interaction between two individuals.

of a person's emotion state on his/her previous emotion state modeled as a first order Markov process. The mutual influence between the two speakers in the dyadic interaction models the effect of an individual's emotion state on his or her interlocutor's emotion history. The paper showed that modeling at the dialog-level could successfully incorporate both time dependency and mutual influence on individual's emotional state to improve emotion recognition performance. The proposed DBN model obtained an absolute 3.67% increase in accuracy (relative 7.12% improvement) over the baseline model, a turn-by-turn static GMM classification without dialog-level modeling.

Emotional interpolation via multi-layered hierarchical modeling is a viable approach for integrating utterance-level emotion profiles into dynamic interaction models. The goal of this modeling is to obtain a global emotion label for a dialog. Emotion profiling is first obtained using low level cues, such as acoustic information. The obtained profile would then serve as the observation features to the dynamic interaction level to estimate the emotion content for the dialog as a whole. This hierarchical structure closely models the structure of the dialog flow between interlocutors' emotions. It could provide more modeling power and a more reliable indication of the emotion content of the dialog than techniques that do not consider the dynamics of inter- and intra-personal interaction.

4. Techniques for hard labeling ambiguous utterances

In certain applications a hard label must be assigned; it is not sufficient to provide soft label or to interpolate the emotion contents of an utterance. In these situations, different techniques must be applied to arrive at a hard label. There exist many techniques for classifying the emotional label of individual utterances. These techniques include Hidden Markov Model classification [5, 21], Support Vector Machine classification [21], Gaussian Mixture Models [20,23], fuzzy logic [17], and many more. There has also been work analyzing the emotional content of an utterance discriminatively by first utilizing neutral speech models [5]. A more thorough overview can be found in [11].

The emotional content of individual utterances can be assessed using fuzzy logic techniques. Fuzzy logic provides a method for dealing with ambiguous emotional utterances when hard labels must be assigned. Fuzzy sets generalize classical sets by allowing partial membership, rather than membership in a single class. Thus, the internal operations of a fuzzy system can make use of the ambiguity expressed in emotional profiles, representing partial membership in one or more emotion categories (sets). Fuzzy sets have been used to represent acoustic features [17], dimensional emotional attributes [17, 19], or emotion category labels [19]. Hard-labels are recovered during defuzzification.

Dialog-level models should also be utilized when assigning hard labels to emotionally ambiguous utterances [20]. In natural communication, humans rarely drastically alter the emotional content of their utterances within small time frames. As individuals interact with each other over a certain time period, conversation partners start to form expectations of their interlocutors emotional communication patterns. Computational methods should leverage this emotional expectation. During classification, an emotional confidence score can be calculated for each emotion label. Utterances with low confidence score can be assigned a new emotional label based on a weighted combination of the temporally closest emotion label with high confidence and the emotional profile of the current utterance. This technique results in a smoothed assessment of emotional content.

When modeling at the utterance level, two user-modeling techniques can be applied, individual-specific models, and averaged models. Individual-specific models seek to model the emotional perceptions of specific users, while averaged models seek to model general trends in emotional communication. While both methods have their value, our recent exploratory analysis has suggested that in certain conditions it may be beneficial to utilize averaged, rather than individual-specific models [24].

5. Conclusion

This paper presents an overview of current methods for emotion classification and possible future directions for dialog interaction-level emotion classification in the presence of prototypical and non-prototypical emotion expressions. Prototypical emotions provide opportunities to create models that capture the properties of well-recognized emotions. These models are of use as they enable an interface to recognize salient emotional behavior in a user population. However, one cannot expect that every emotional utterance will be prototypical. Consequently, it is important to develop techniques to interpret emotional expressions independent of their prototypical or non-prototypical nature. Dialog-level emotion classification can be improved by considering not only the dynamics of the acoustic features but also the

dynamics of the underlying classification. This combination will allow the system to classify the emotion expressed within the context of a dialog based on emotionally clear data, rather than trying to integrate ambiguous emotional content. It is the view of the authors that this technique, in combination with more conventional hard-labeling assignment techniques will result in enhanced automated emotional comprehension by machines.

References

[1] R. Banse and K. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70:614–636, 1996.

[2] J. Burgoon, L. Stern, and L. Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 1995.

[3] C. Busso, M. Bulut, 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*, pages 335–359, Nov. 5 2008.

[4] C. Busso, S. Lee, and S. Narayanan. Using neutral speech models for emotional speech analysis. In *InterSpeech ICSLP*, pages 2225–2228, Antwerp, Belgium, Aug. 2007.

[5] C. Busso, S. Lee, and S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing*, 17(4):582–596, May 2009.

[6] C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2331–2347, Nov. 2007.

[7] C. Busso and S. Narayanan. Recording audio-visual emotional databases from actors: a closer look. In *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, pages 17–22, Marrakech, Morocco, May 2008.

[8] C. Busso and S. Narayanan. Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database. In *Interspeech 2008 - Eurospeech*, pages 1670–1673, Brisbane, Australia, Sept. 2008.

[9] R. Cauldwell. Where did the anger go? the role of context in interpreting emotion in speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 127–131, Newcastle, Northern Ireland, UK, Sept. 2000.

[10] R. Cowie and R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, 2003.

[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, pages 32–80, Jan. 2001.

[12] K. Dautenhahn, C. Numaoka, and AAAI. *Socially intelligent agents*. Springer, 2002.

[13] L. De Silva, T. Miyasato, and R. Nakatsu. Facial emotion recognition using multi-modal information. In *International Conference on Information, Communications and Signal Processing (ICICS)*, volume I, pages 397–401, Singapore, 1997.

[14] L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, May 2005.

[15] E. Douglas-Cowie, L. Devillers, J. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pages 813–816, Lisbon, Portugal, Sept. 2005.

[16] F. Enos and J. Hirschberg. A framework for eliciting emotional speech: Capitalizing on the actors process. In *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, pages 6–10, Genoa,Italy, May 2006.

[17] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49:787–800, 2007.

[18] A. Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, March 2006.

[19] A. Kazemzadeh, S. Lee, and S. Narayanan. An interval type-2 fuzzy logic system to translate between emotion-related vocabularies. In *In Proceedings of Interspeech*, Brisbane, Australia, Sept. 2008.

[20] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *10th Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, UK, Sept. 2009.

[21] Y. Lin and G. Wei. Speech emotion recognition based on HMM and SVM. *Proc. of Int. Conf. on Machine Learning and Cybernetics*, 8:4898–4901, Aug. 2005.

[22] A. Mehrabian. *Nonverbal communication*, chapter 6, pages 104–132. Aldine, 2007.

[23] A. Metallinou, S. Lee, and S. Narayanan. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *IEEE International Symposium on Multimedia*, Berkeley, California, Dec. 2008.

[24] E. Mower, M. Mataric, and S. Narayanan. Evaluating evaluators: A case study in understanding the benefits and pitfalls of multi-evaluator modeling. In *10th Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, UK, Sept. 2009.

[25] E. Mower, M. Matarić, and S. Narayanan. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *IEEE Transactions on Multimedia*, 11(4), 2009.

[26] M. Pantic, N. Sebe, J. Cohn, and T. Huang. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676. ACM New York, NY, USA, 2005.

[27] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, pages 4–15, Jan. 1986.

[28] J. Russell and L. Barrett. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5):805–819, 1999.

[29] K. Scherer. On the nature and function of emotion: A component process approach. *Approaches to emotion*, pages 293–317, 1984.

[30] K. Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165, 1986.

[31] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227 – 256, 2003.

[32] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson. Patterns, Prototypes, Performance: Classifying Emotional User States. *Interspeech*, 2008.

[33] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann. "Of all things the measure is man": Automatic classification of emotions and inter-labeler consistency. In *ICASSP, 2005.*, volume 1, pages 317–320, 2005.

[34] M. Swerts and E. Krahmer. The importance of different facial areas for signalling visual prominence. In *International Conference on Spoken Language (ICSLP)*, pages 1280–1283, Pittsburgh, PA, USA, Sept. 2006.

[35] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, England, 1997.