Emotion Unfolding and Affective Scenes: A Case Study in Spoken Conversations

Morena Danieli
Dept. of Information Engineering and Computer Science
University of Trento, Italy
morena.danieli@unitn.it

Giuseppe Riccardi
Dept. of Information Engineering and Computer Science
University of Trento, Italy
giuseppe.riccardi@unitn.it

Firoj Alam
Dept. of Information Engineering and Computer Science
University of Trento, Italy
firoj.alam@unitn.it

ABSTRACT

The manifestation of human emotions evolves over time and space. Most of the work on affective computing research is limited to the association of context-free signal segments, such as utterances and images, to basic emotions. In this paper, we discuss the hypothesis that interpreting emotions requires a conceptual description of their dynamics within the context of their manifestations. We describe the unfolding of emotions through the proposed *affective scene* framework. Affective scenes are defined in terms of *who* first expresses the variation in their emotional state in a conversation, *how* this affects the other speaker's emotional appraisal and response, and *which* modifications occur from the initial through the final state of the scene. This conceptual framework is applied and evaluated on real human-human conversations drawn from call centers. We show that the automatic classification of affective scenes achieves more than satisfactory results and it benefits from acoustic, lexical and psycholinguistic features of the speech and linguistics signals.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; I.5 [**Information Systems**]: Pattern Recognition

Keywords

Affective Scene; Emotion; Spoken Conversation; Computational Paralinguistics; Machine Learning

1. INTRODUCTION

In natural human-human conversations emotions and feelings are experienced, expressed, mutually interpreted, and continuously processed by the speakers. Understanding the flow of speakers' emotional states in socially regulated contexts constitutes an important goal for the development of behavioral analytics with relevant societal impact. The possible areas of application of behavior analytics include a wide range of human relationships such as therapist-patient, teacher-student, agent-customer, and employer-employee interactions. Reliable recognition and classification of individual emotional states is crucial for developing natural and effective applications in all those areas. This emerging field of research may benefit from the adoption of holistic methodologies that can combine multiple behavioral cues in order to predict behavior. While much of the research on affective computing focuses on emotion recognition, the modeling of the emotion dynamics associated with social interactions is little explored.

In the past few decades, the automatic emotion recognition research community has focused on the analysis of several different cues of emotional behavior including gesture, voice, face [14, 21], and multi-modal [13] expressions. Understanding human emotional behavior is relevant for application domains such as human-machine dialog [19] and call center interactions [7]. The lack of a complete description of the emotional space, as a sequence, might be partly due to difficulties in representing the multiplicity of factors involved and the lack of operationalized concepts that may help in describing such complexity.

In this paper, we introduce the concept of *affective scenes* for modeling the complex interplay between the vocal expression of emotions and its impact on the communicative situation that it originates from. The concept is built on the psychological model of emotions developed by Gross [10], and it includes an annotation model for basic and complex social emotions. In Section 4, we will explain the Gross's modal model of emotions, which provides valuable insights for determining possible sets of high level features characterizing the process of affective communication.

The analysis of the dynamics of the emotional flow in conversations shows that the mental states of the speakers continuously vary from neutral to emotionally connoted states. While in principle the variations may be endless and random, we want to test the hypothesis that they can be traced back to some prototypical categories. For modeling such variations it is important to capture the origin of these relational events that originate them. To approach this task we describe the *affective scenes* in terms of *who* 'first' expresses the emotional variation, *how* this affects the other speaker's emotional appraisal and response, and *which* modifications occur from the initial through the final state of the scene. Accordingly, the annotation model we propose is conceived

ERM4CT'15, November 13, 2015, Seattle, WA, USA.
 ISBN 978-1-4503-3988-9/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2829966.2829967 .

for focusing the annotators' perception on the variations of the speakers' emotional state.

The paper is organized as follows. In Section 2, we provide a review of literature relevant to our work. We introduce the concept of *affective scene* in Section 3 and we provide affective scene based analysis in Section 5. In Section 4, we illustrate the details of the corpus we used to validate our framework. We also present the task of automatic classification of affective scenes categories in Section 6, and illustrate our experimental details, and discuss the results in Section 7. Finally, we provide conclusions in Section 8.

2. BACKGROUND

In the past few decades, there has been effort to design and develop methodologies for affective interaction. Methodologies based on behavior observation protocols can identify recurrent categories of emotional exchange by utilizing data that is generated by listening to recorded conversations, and then manually coding the relevant emotional transitions. However, it is well known that such methods are widely acceptable but highly time-consuming. In the affective computing literature, there are evidences of behavior analysis methods in domains where traditional observational protocols can be applied. In [4], the authors proposed an approach to automate a manual human behavior coding technique for couple therapy. They use acoustic speech features in order to classify the occurrences of basic and complex emotional attitudes of the subjects such as sadness, blame, and acceptance.

Focusing on sequential organization is important when studying emotion in real setting, as reported by Goodwin & Goodwin [9] in the field of conversation analysis. Their results show that powerful emotional statements can be built by the use of sequential positions, resources provided by the environment where the action occurs, and *artful orchestration* of a range of embodied actions such as intonation, gesture, and timing. In [16], Lee et al. reported that modeling the conditional dependency between two interlocutors' emotional states in sequence improves the automatic classification performance where they used a corpus in which actors manifested emotional episodes.

Following the sequential organization and the modal model of emotion [11] in our experiment:

1. we model the unfolding of discrete emotional states using a real-life corpus,
2. then, we analyze the ways and occurrences of a given emotion, which is felt and expressed by one speaker, and its impact on another's emotional state,
3. we study how the emotional state feeds back to the initial situation, and define affective scenes in terms of the emotional episodes,
4. then, we label each conversation with affective scenes.

3. AFFECTIVE SCENE FRAMEWORK

3.1 Definition of Affective Scene

The emotional states of individuals engaged in a conversations are characterized by continuous variations. We hypothesize that the linguistic and contextual structures of such variations can be objectively described by exploiting the correlation between the continuous variations in speakers' emotional states and variations in the situational context of the interaction. In the psychological literature there are some competing models aiming to capture and describe such variation, including Scherer's dynamic theory of emotion differentiation and sequential checking [20]. For our experiment we refer to a general, yet clear and flexible, psychological model that may account for the interplay between variation of emotional state and variation of the situational context, i.e. the *modal model* of emotions [11]. According to the modal model, the emotion-arousal process is believed to be induced by a *Situation*, a physical or virtual space that can be objectively defined. The *Situation* compels the *Attention* of the subject and triggers the subject's *Appraisal* process and the related emotional *Response*. The *Response* may generate actions that in turn modify the initial *Situation*. In this study, we focused on the *affective scenes* that ensue in such communicative situations.

We define the ***affective scene*** in the context of a dyadic human communication, but it may be generalized to multiparty communication. The affective scene is ***an emotional episode where one individual is affected by an emotion-arousing process that (a) generates a variation in their emotional state, and (b) triggers a behavioral and linguistic response. The affective scene extends from the event triggering the unfolding of emotions on both individuals, throughout the closure event when individuals disengage themselves from their communicative context.***

While this process is continuous in terms of the human response signals, we describe the unfolding as a sequence of discrete emotional episodes that have an initial state, a sequence of states, and a final state. In order to describe the framework of affective scenes we focused on *who* 'first' shows the variation of their emotional state, *how* the induced emotion affects the other speaker's emotional appraisal and response, and *which* modifications such a response may cause with respect to the state that triggered the scene.

3.2 Corpus Based Analysis

We applied the definition of affective scene to the analysis of dyadic spoken conversations between customers and agents collected in call centers. The details of the corpus are provided in Section 4. The analysis of the conversations provided evidence for several occurrences of *prototypical* emotional sequences. For example, we observed situations where customers call in order to complain about an unfulfilled service request they made a few weeks before. The customers are *frustrated* due to the delay, and the manifestations of their emotional state triggers the start of affective scene instances. This scenario represents the term *who* in our definition: in this scenario, *who* initiated emotion? → customer. The agents may understand the point of view of the customers, and *empathize* with their distress. In addition, the agents may take all the required actions in order to solve the customers' problem. The appropriate response and actions by the agents may impact on the emotional states of the customers. In this case, *how* the agent shows an emotional response that reflects the second term of our definition, such as *how* agent is responding? → by empathizing.

The emotional states can vary again so that the call may end with customer's satisfaction. On the other hand, the lack of empathic response from the agents, and/or the fact

Table 1: Types of affective scenes for different communicative situations. Initial State: Initial emotional state of either agent or customer. A: Agent, C: Customer, Emp: Empathy, Ang: Anger, Fru: Frustration, Sat: Satisfaction. As an example, C: Fru means customer manifests frustration. A complete emotion sequence with → indicates the flow of emotions in conversation.

Initial state	Scenarios	Examples
Customer initial discontent	Agent understands, and customer's discontent is attenuated	C: Fru → A: Emp → C: Sat
	Agent understands, but customer emotional state either get worse or does not evolve positively	C: Fru → A: Emp → C: Fru
		C: Fru → A: Emp → C: Ang
	Agent does not understand, and customer emotional state either get worse or does not evolve positively	C: Fru → A: Neu → C: Fru
		C: Fru → A: Neu → C: Ang
Agent preempting of possible customer discontent	Customer emotional state does not vary	A: Emp → C: Fru or Ang
	Customer emotional state evolves into a positive attitude	A: Emp → C: Sat

that the problem cannot be solved immediately, may cause different patterns of emotional variations, including customer anger, dissatisfaction, or further frustration. The type of emotional modification we see here on the customer side reflects the third term of our definition; in this scenario, *which* type of modification? → satisfaction, anger or dissatisfaction.

In order to clarify the scenarios, we illustrate two prototypical communicative situations, as shown in Table 1. The first situation is characterized by a customer's initial discontent, and the second by agent's initial positive attitude towards the customer's state of mind. As it can be seen in the Table 1, the unfolding of the affective scenes from those initial situations may greatly vary from one scenario of communicative situations to another one. In the first example, row 1 in Table 1, we see that the customer 'first' manifests emotion with frustration, then the agent understands and empathizes, and 'finally' the customer changes their emotional state from frustration to satisfaction.

4. CORPUS DESCRIPTION

4.1 Corpus

The corpus includes 1651 customer-agent conversations (186 hours in total) selected randomly over the course of six months, recorded on two separate channels (16 bits per sample, 8kHz sampling rate). The average length of the conversations is 406 seconds, the annotation time of a conversation was, on average, about 18 minutes. The corpus was annotated with respect to a set of emotions including basic emotions such as *anger*, and complex social emotions such as *satisfaction, dissatisfaction, frustration* and *empathy*. The *neutral* tag was introduced as a relative concept to support annotators in their perceptual process while identifying the situation of the context. See Section 4.2 for more information about the annotation.

4.2 Model, Guidelines, and Protocol of Annotation

Following the principles of the *modal model* of emotion [11], we designed and empirically validated the annotation model. The *modal model* of emotions developed by Gross [10, 12] emphasizes the attentional and appraisal acts underlying the emotion-arousing process. In Figure 1, we outline the heuristics of the model. The core *Attention-Appraisal* individual's processes (included the box) are affected by the *Situation* that is objectively defined in terms of physical or virtual space and objects. The situation compels the attention to the individual. It triggers an appraisal process and gives rise to coordinated and malleable responses. This model is dynamic and it shows that the situation may be modified (directed arc from the *Response* to the *Situation*) by the actual value of the *Response* generated by the *Attention-Appraisal* process. Figure 1 also shows an arrow from Response to Situation because the response often changes the original situation and can modify it.

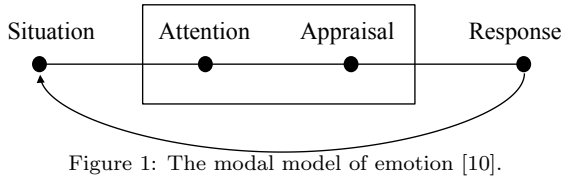

Figure 1: The modal model of emotion [10].

The annotation guidelines were inspired by this model, and provided operational definitions for tagging our set of emotions. In the annotation guidelines, we operationally defined *satisfaction* as *'a state of mind deriving from the fulfilment of individual will or needs'*, and *dissatisfaction* as *'implying some disappointment of individual expectancies'*. It is likely that the issues dealt within the task oriented dialogs of our application domain are solved when the client is satisfied. On the contrary, dissatisfaction signals a general closure and hopeless attitude in the speaker who feels it, and in general towards the chance to be understood and helped by the other speaker.

The operational definition of *frustration* sees it as *'a complex emotional state that arises from the perceived resistance to the fulfillment of individual will or needs'*. For example, in our application domain it is likely that the customer is experiencing frustration in scenarios such as when s/he has to call back many times to solve the same query, when s/he needs to query issues that could have been prevented before in the conversation, and when the agent cannot solve the issues.

Anger is usually described as *the emotion related to one's psychological interpretation of having been offended, defamed, or denied and a tendency to react through retaliation.* In our annotation model, we tag the emergence of anger when the perception of the speech signal shows typical anger voice traits, such as tension, hyper-articulation (cold anger), and raised voice.

We operationally defined *empathy* as *'a situation where an agent anticipates or views solutions and clarifications, based on the understanding of a customer's problem or issue, that can relieve or prevent the customer's unpleasant feelings'*. The annotators need to focus on dialog turns where the agent anticipates solutions and clarifications (attention), based on the understanding of the customer's problem (ap-

praisal), and the agent's acts may relieve or prevent customer's unpleasant feelings (response).

To measure the reliability of the guidelines we calculated inter-annotator agreement by using the kappa statistics [5] over a set of 60 conversations annotated by two expert psycholinguists. We obtained a reliable kappa value of 0.74.

For the corpus annotation task based on the validated guidelines, we recommended the annotators to focus on their perception of speech variations, in particular on the acoustic and prosodic quality of pairs of speech segments, while minimizing any effort to pay attention to the semantic content of the utterances. They were asked to judge the relevance of the perceived variations with respect to the expression of emotions. In particular, the annotation task required:

1. to select pairs of speech segments where the annotators could perceive variations in the melody of speech or in the speaker's tone of voice,
2. to evaluate the communicative situation in terms of appraisal of the transition from a neutral emotional state to an emotionally connoted one, and
3. to identify the onset of the variation and assign a label.

By following the guidelines and the recommendations, we trained two expert annotators with psycholinguistic background. For this annotation task only the emotional categories were annotated.

5. AFFECTIVE SCENES IN SPOKEN CONVERSATIONS

We analyzed 460 annotated conversations containing empathy on the agent side and other basic and complex emotions on the customer side. In Table 1, we report examples of two communicative situations based on the 'initial' manifestations of emotion. In the examples we also report the customers' 'final' emotional state during the conversation.

Based on the initial and final emotional displays we defined and categorized the conversations with different categories of *affective scenes* given in Figure 2, which depicts the emotional sequence examples in Table 1. From the figure we see that after the start of the conversation either agent or customer manifest emotions. Following that there are many emotional transitions between customer and agent, and there is a 'final' emotional manifestation.

Hence, considering the 'initial' emotional displays of customer and agent and 'final' emotional displays of the customer, we have three categories of affective scenes as listed below.

1. Agent or customer manifest emotions at the start of the conversation, therefore we use the terms - Agent First (AF) *or* Customer First (CF).
2. Agent 'first' manifests emotion after the start of the conversation and customer shows positive/negative emotion at the end of the conversation. We use the terms AF-Pos *or* AF-Neg.
3. Customer 'first' manifests emotion after the start of the conversation and customer shows positive/negative emotion at the end of the conversation. To define this scenerio we use the terms CF-Pos *or* CF-Neg.

The negative emotions in this case are anger, frustration and dissatisfaction whereas the positive emotion is satisfaction.

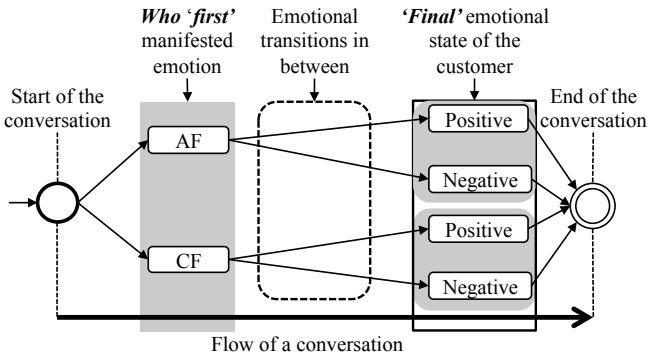

Figure 2: State traces of affective scenes. Starting from the initial state, an affective scene may reach either state **AF** (Agent first manifests emotion) or **CF** (Customer first manifests emotion). Then, following a natural unfolding of emotion states the affective scene may reach either a **Positive** final state (Customer manifests emotion with satisfaction at the end) or a **Negative** final state (Customer manifests emotion with either anger, frustration or dissatisfaction at the end).

From the analysis of emotional sequences we can see that *Emp* $\rightarrow$ *Sat* appears more frequently than others, 30.7% relative frequency distribution. Some examples of emotional sequence and their distributions are given in Table 2.

Table 2: Examples of emotional sequence (Seq.) with their relative frequency distribution (Dist.) out of 460 conversations

Initial emotional manifestation	Seq.	Dist.
Agent manifested emotion 'first'	Emp $\rightarrow$ Fru	3.5
	Emp $\rightarrow$ Dis	3.5
	Emp $\rightarrow$ Sat	30.7
Customer manifested emotion 'first'	Fru $\rightarrow$ Ang	3.3
	Ang $\rightarrow$ Dis	4.8
	Fru $\rightarrow$ Dis	10.0
	Fru $\rightarrow$ Emp	9.8
	Fru $\rightarrow$ Emp $\rightarrow$ Sat	3.5
	Fru $\rightarrow$ Sat	2.8

6. CLASSIFICATION TASK

To automatically classify the affective scenes categories, described in Section 5, we designed three binary classification tasks:

1. Agent First (AF) *or* Customer First (CF);
2. Agent First, customer manifests Positive emotions at the end (AF-Pos) *or* Agent First, customer manifests Negative emotions at the end (AF-Neg);
3. Customer First, customer manifests Positive emotions at the end (CF-Pos) *or* Customer First, customer menifests Negative emotions at the end (CF-Neg).

Class distribution for each of the classification task is given in Table 3.

Table 3: Distribution of conversations for each classification task.

	Task 1		Task 2		Task 3	
Class	AF	CF	AF-Pos	AF-Neg	CF-Pos	CF-Neg
No. of conv.	213	247	160	53	47	200
%	46.3	53.7	75.1	24.9	19.0	81.0

6.1 Feature Extraction

Recent studies showed that we can characterize emotion and personality of the speaker by utilizing acoustic, lexical and psycholinguistic features [25, 2]. Therefore, for this study we investigated these feature sets to understand their distinctive properties for the classification of affective scenes. Another reason to investigate these feature sets was that we wanted to understand whether they can be useful for the classification without explicitly knowing the emotion sequence. Since each conversation is represented into two channels (agent and customer), therefore, we extracted features from both channels and then concatenated them as shown in Figure 3.

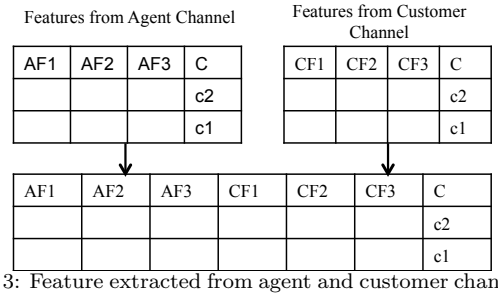

Figure 3: Feature extracted from agent and customer channel and then merged them. Each cell contains feature values.

6.1.1 Acoustic Features

There has been evidence in the literature that paralinguistic properties of speech have distinctive power in classifying emotion while representing them by low-level features and then projecting them onto statistical functionals [23]. Before extracting the features we preprocessed the speech signal to remove silence from the beginning and end. We also removed silence that is longer than one second. For this task, we extracted a vary large set of low-level acoustic features and then projected them onto statistical functionals, using openSMILE [8], based on the feature configuration referred in [24]. We extracted low-level features with approximately 100 frames per second. For reproducibility and transferability the details of the low-level features and statistical functionals are given in Table 4. After projecting low-level features and their delta onto statistical functionals, the feature set contains 6373 features. Since each conversation is comprised of agent and customer channels, therefore, we extracted the same number of acoustic features from the Agent, $A = \{a_1, a_2, ..., a_m\}$ and the Customer, $C = \{c_1, c_2, ..., c_m\}$. Then, merged the features from both channels to form a new feature vector, $X = \{a_1, a_2, ..., a_m, c_1, c_2, ..., c_m\}$.

6.1.2 Lexical Features

Since lexical choices of the speaker provides evidences that represents emotional manifestations. Therefore, for the classification we extracted lexical features from automatic transcriptions, which we obtained using a large vocabulary Automatic Speech Recognition (ASR) system [6]. The Word Error Rate (WER) of the ASR system was 31.78% on the test set and 20.87% on the training set. We designed affective scene instances by concatenating the transcriptions from agent and customer channels. Then, converted them into lexical feature vector in the form of bag-of-words and used tf-idf (term frequency times inverse document frequency) weighting scheme. The idea of tf-idf is that it accounts how often a word appear in a instance but not in other instances. In information retrival domain the term 'document' is normally used to refer to an instance. While converting into the feature vector we removed stop words. In order to take the advantage of contextual benefits we also extracted unigram, bigram and trigram features. This procedure resulted in a very large dictionary. We filtered out lower frequency features by selecting 10K top frequent features to reduce the size of the feature vector.

Table 4: Low-level acoustic features and statistical functionals

Low-level acoustic features
Raw-Signal: Zero crossing rate
Energy: Root-mean-square signal frame energy
Pitch: F0 final, Voicing final unclipped
Voice quality: jitter-local, jitter-DDP, shimmer-local, log harmonics-to-noise ratio (HNR)
Spectral: Energy in bands 250-650Hz, 1-4kHz, roll-off-points (0.25, 0.50, 0.75, 0.90), flux, centroid, entropy, variance, skewness, kurtosis, slope, harmonicity, psychoacoustic spectral sharpness
Auditory-spectrum: band 1-26, auditory spectra and rasta
Cepstral: Mel-frequency cepstral coefficitnts (mfcc 0-12)
Statistical functionals
Relative position of max, min
Quartile (1-3) and inter-quartile (1-2, 2-3, 3-1) ranges
Percentile 1%, 99% and percentile range 1%-99%
Std. deviation, skewness, kurtosis, centroid, range
Mean, max, min and Std. deviation of segment length
Uplevel time 25, 50, 75, 90; rise time, left curvature time
Linear predictive coding lpc-gain, lpc0-5
Arithmatic mean, flatness, quadratic mean
Mean dist. between peaks, peak dist. Std. deviation, absolute and relative range, mean and min of peaks, arithmatic mean of peaks, mean and Std. of rising and falling slope
Linear regression coefficients (1-2) and error
Quadratic regression coefficients (1-3) and error

6.1.3 Psycholinguistic Features

In the past few decades Pennebaker et al. [17] have reported the usefulness of psycholinguistic features in different domains such as understanding gender, age, personality and health. Their efforts resulted the development of Linguistic Inquiry Word Count (LIWC) system[1]. They have reported that these features can capture the meaning in different scenerios such as emotionality, thinking styles, social relationships, and personality. For this study, we also extracted psycholinguistic features from automatic transcriptions using LIWC. It is a knowledge based system containing dictionaries for several languages including Italian. The Italian dictionary contains 85 word categories, and additionally LIWC extract 6 general descriptors and 12 punctuation categories for a total of 103 features. We removed LIWC features that are not observed in our dataset and finally, we obtained 89 psycholinguistic features. The types of descriptors that we extracted using LIWC are reported in Table 5. LIWC calculates the percentage words in a conversation that match each of 85 feature categories in the LIWC dictionary. For the general and punctuation feature categories it does not rely on dictionary. The usefullness of these features has also been evidenced in other paralinguistic tasks such as [3] and [2].

[1]http://liwc.net

Table 5: Psycholinguistic features extracted using LIWC system

LIWC features
General features
Word count, words/sentence, percentage of words exist in dictionary, percentage of words greater than 6 letters, and numerals
Linguistic features
Pronouns, articles, verbs, adverbs, tense, prepositions, conjunctions, negations, quantifiers, and swear words
Psychological features
Social processes: family, friends and humans **Affective processes:** positive, negative, anxiety, anger, and sadness **Cognitive processes:** insight, causation, discrepancy, tentative, certainty, inhibition, inclusive, exclusive, perceptual, see, hear, and feel **Biological processes:** body, health, sexual, and ingestion **Relativity:** motion, space, and time
Personal concern
Work, achievement, leisure, home, money, religion, and death
Paralinguistic features
Assent, nonfluencies, and fillers
Punctuation features
Period, comma, colon, semi-colon, question mark, exclamatory mark, dash, quote, apostrophe, parenthesis, other punctuations, and percentage of all punctuations

6.2 Feature Combination and Selection

In addition to understand the performance of each feature set, such as acoustic and lexical feature sets, we also wanted to understand their combined contribution. Therefore, following the feature extraction, we merged acoustic and lexical features into a single vector to represent each instance in a high-dimensional feature space. Let $S = \{s_1, s_2, ..., s_m\}$ and $L = \{l_1, l_2, ..., l_n\}$ denote the acoustic and lexical feature vectors respectively. The feature-combined vector is $Z = \{s_1, s_2, ..., s_m, l_1, l_2, ..., l_n\}$ with $Z \in R^{m+n}$.

Since each individual feature set is higher dimensional, particularly acoustic and lexical, we applied Relief [15] feature selection technique. It has been shown in the literature [1] that this feature selection technique comparatively performs well, for paralinguistic task, compared to other techniques such as Information gain. Relief estimates the quality of a feature based on how well its values distinguish among instances that are near to each other. For a given instance, it searches for two nearest instances, one from same class and one from different class and estimate weight of an attribute depending on the values of the nearest intstances.

As a part of feature selection process we generated feature learning curves using ranked features from Relief and selected optimal set of features when performance start decreasing.

6.3 Classification and Evaluation

We designed our classification models using Sequential Minimal Optimization (SMO) [18], which is a technique for solving the quadratic optimization problem of Support Vector Machines' (SVM) training. We trained the model using an open-source implementation Weka machine learning toolkit [26]. We chose to use *linear kernel* of SVM in order to alleviate the problem of higher dimensions such as overfitting. In order to measure the performance of the system we used Un-weighted Average, $UA = \frac{1}{2}\left(\frac{tp}{tp+fn} + \frac{tn}{tn+fp}\right)$, where tp, tn, fp, fn are the number of true positives, true negatives, false positives and false negatives, respectively. It has been widely used for the evaluation of the paralinguistic task [22]. Due to the limited size of the corpus we chose to use 10-folds cross-validation. In addition, we optimized the penalty parameter C in the of $[10^{-5}, ..., 10]$.

7. RESULTS AND DISCUSSION

The performance of the system for each feature set is shown in Table 6. We present average UA of 10 folds cross validation, their standard deviation, and number of features for a particular feature set. We also present random baseline results, which we computed by randomly generating class labels based on the prior class distribution.

Table 6: Classification results of affective scenes categories in terms of UA, (average±standard deviation) with feature dimension (Feat.). Ac: Acoustic, Lex-ASR: Lexical features from ASR transcription, LIWC: Psycholinguistic features. {AF,CF}: Agent First, Customer First, AF:{Pos,Neg} Agent First with Positive/Negative emotion of the customer at the end, CF:{Pos,Neg} Customer First with Positive/Negative emotion of the customer at the end

	Task1 {AF,CF}		**Task2** AF: {Pos,Neg}		**Task3** CF:{Pos,Neg}	
Exp.	Avg±Std	Feat.	Avg±Std	Feat.	Avg±Std	Feat.
Random	49.3±7.0	-	49.8±10.1	-	49.0±11.2	-
Ac	58.5±6.7	1000	65.0±13.9	3000	63.9±10.0	4500
Lex-ASR	73.2±6.2	6800	67.5±12.8	5000	70.3±8.1	6800
LIWC	67.8±5.9	89	56.9±11.2	89	49.5±10.6	89

Out of the three classification tasks, we obtained better performance on task 1, {AF,CF}, compared to the other two classification tasks. The higher variation of the classification results of task 2 and 3 is due to the imbalance class distribution and smaller number of instances compared to task 1. From the classification results, we observed that the performance of lexical features outperforms any other single or combined feature set such as acoustic, lexical, acoustic with lexical and LIWC.

The performance of acoustic feature set is better than random baseline and it might be useful when there is no transcription available. In terms of feature dimension, with feature selection we obtained smaller-sized features for this set compared to lexical features for all three tasks. For task 2 and 3 acoustic features performs better than psycholinguistic features. For all three task we found that spectral features are highly relevant for discriminating between classes.

The performance of psycholinguistic features are better than acoustic in task 1, however, in other two tasks its performances are worse. We used linear kernel of SVM for all classification experiments, however, it might not be a better fit for this feature set. Gaussian kernel might be a better option in this case, which we might explore in future.

Linear combination of acoustic+lexical (Z) did not perform well due to the complexity of the large feature space, which we are not presenting here. We will explore it in future by studying it with ensemble methods such as stacking.

Even though the classification performance varies across tasks and feature sets, however, from the results we can infer that automatically categorizing the affective scenes might be future research avenue to investigate. Our conceptual framework can be a good starting point towards defining affective scenes and its automatic classification.

8. CONCLUSIONS

In this paper, we have proposed the conceptual framework of *affective scenes* to describe the dynamics of emotion unfolding in natural conversations. Even though we validated the *affective scenes* framework on call center conversations, nevertheless we believe that the framework is applicable to behavioral analysis of other social scenarios, for example

therapist-patient interactions. Our future research efforts will move towards this extension, as well as to the evaluation of the possible suitability of alternative psychological models, as long as we will be able to experimentally identify the still lacking features of high level emotional flow in dyadic conversations. In this paper, we also investigated automatic classification of *affective scenes* categories by exploiting acoustic, lexical (ASR transcription) and psycholinguistic features. We obtained promising performance using lexical features, and with all other feature sets we are still getting better than random baseline.

9. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union - 7th Framework Programme (FP7/2007-2013) under grant agreement n° 610916 - SENSEI - http://www.sensei-conversation.eu/.

10. REFERENCES

[1] F. Alam and G. Riccardi. Comparative study of speaker personality traits recognition in conversational and broadcast news speech. In *Proc. of Interspeech*, pages 2851–2855. ISCA, 2013.

[2] F. Alam and G. Riccardi. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 955–959, May 2014.

[3] F. Alam and G. Riccardi. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 15–18. ACM, 2014.

[4] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, 55(1):1–21, 2013.

[5] J. Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254, 1996.

[6] S. A. Chowdhury, G. Riccardi, and F. Alam. Unsupervised recognition and clustering of speech overlaps in spoken conversations. In *Proc. of Workshop on Speech, Language and Audio in Multimedia - SLAM2014*, pages 62–66, 2014.

[7] L. Devillers and L. Vidrascu. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Proc. of Interspeech*, pages 801–804, 2006.

[8] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. of the 21st ACM international conference on Multimedia (ACMM)*, pages 835–838. ACM, 2013.

[9] M. H. Goodwin and C. Goodwin. Emotion within situated activity. *Communication: An arena of development*, pages 33–53, 2000.

[10] J. J. Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271, 1998.

[11] J. J. Gross. *Handbook of emotion regulation*. Guilford Press, 2011.

[12] J. J. Gross and R. A. Thompson. Emotion regulation: Conceptual foundations. *Handbook of Emotion Regulation*, 3:24, 2007.

[13] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *Proc. of Multimedia Signal Processing, 2007 (MMSP 2007)*, pages 48–51, 2007.

[14] A. Konar and A. Chakraborty. *Emotion Recognition: A Pattern Analysis Approach*. John Wiley & Sons, 2014.

[15] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Proc. of Machine Learning: European Conference on Machine Learning (ECML)*, pages 171–182. Springer, 1994.

[16] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Proc. of Interspeech*, pages 1983–1986, 2009.

[17] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

[18] J. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT Press, 1998.

[19] G. Riccardi and D. Hakkani-Tür. Grounding emotions in human-machine conversational systems. *Lecture Notes in Computer Science, Springer-Verlag*, pages 144–154, 2005.

[20] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, pages 92–120, 2001.

[21] B. Schuller and A. Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[22] B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *Proc. of Interspeech*, pages 312–315, 2009.

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. Paralinguistics in speech and language state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013.

[24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proc. of Interspeech*, 2013.

[25] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. Acoustic emotion recognition: A benchmark comparison of performances. In *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 552–557, 2009.

[26] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.