# Chapter 6
# Automatic Labeling Affective Scenes in Spoken Conversations


**Firoj Alam, Morena Danieli and Giuseppe Riccardi**

**Abstract** Research in affective computing has mainly focused on analyzing human emotional states as perceivable within limited contexts such as speech utterances. In our study, we focus on the dynamic transitions of the emotional states that are appearing throughout the conversations and investigate computational models to automatically label emotional states using the proposed *affective scene framework*. An affective scene includes a complete sequence of emotional states in a conversation from its start to its end. Affective scene instances include different patterns of behavior such as *who* manifests an emotional state, *when* it is manifested, and *which* kinds of changes occur due to the influence of one's emotion onto another interlocutor. In this paper, we present the design and training of an automatic affective scene segmentation and classification system for spoken conversations. We comparatively evaluate the contributions of different feature types in the acoustic, lexical and psycholinguistic space and their correlations and combination.

F. Alam (✉) · M. Danieli · G. Riccardi
Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
e-mail: firoj.alam@unitn.it; firoj.alam@alumni.unitn.it

M. Danieli
e-mail: morena.danieli@unitn.it

G. Riccardi
e-mail: giuseppe.riccardi@unitn.it



R. Klempous et al. (eds.), *Cognitive Infocommunications, Theory and Applications*, Topics in Intelligent Engineering and Informatics 13,
https://doi.org/10.1007/978-3-319-95996-2_6

## 6.1 Introduction

In social interactions, the emotional states of the interlocutors[1] continuously change over time. The flow of emotional states is influenced by the speaker's internal state, by the interlocutor's response, and by the surrounding environmental stimuli in the situational context [21]. For the development of behavioral analytics systems, there is a need to automatically recognize the flow of speakers' emotions in socially regulated conversations.

Behavioral analytics systems that can analyze the emotional flow may facilitate the automatic analysis of naturally-occurring conversations from different areas of human relationships, including therapist versus patient, teacher versus student, and agent versus customer interactions.

State of the art literature provides evidence that the understanding of emotions is important in human-machine dialogue systems [40] and call center interactions [18].

During the past century, the study of human emotions was a critical field of investigation in humanities. While most of the studies in this area focused on the classification of emotions into basic and complex categories, some psychological studies focused on the emotional process itself. For example, the appraisal theory [43] illustrated the sequential organization of the emotional process. According to this theory, an emotional state can change because of the underlying appraisals and reaction processes. Moreover, the sequential organization model assumes that the appraisal process is always active, and it continuously evaluates an event or situation to update its organism's information. Based on the appraisal phenomenon, Gross proposed the *modal model* of emotion [25]. This model illustrates the emotional process in term of the situation → attention → appraisal → response sequence. Gross's theory states that the emotion-arousal process is induced by a situation, i.e., a physical or virtual space that can be objectively defined. The situation compels an attention of the subject and it elicits the subject's appraisal process and the related emotional response. The response may generate actions that in turn modify the initial situation. Based on this model, Danieli et al. [17] proposed a concept, named *affective scene*. The purpose was to model the complex interplay between the expression of emotions and its impact in different conversational interactions.

In the field of automatic analysis, the research on affective computing mainly focused on analyzing low-level signals such as utterances, turns in spoken conversation or images. However, the present studies lack the modeling of the entire emotional space at the higher level, such as the flow of an emotion sequence throughout the conversation. There could be several reasons such as lack of operationalized concepts and the difficulty in representing them in conversational interactions. The study of Lee et al. [33] suggests that modeling the conditional dependency between two interlocutors' emotional states in sequence improves the automatic classification performance. In [21], Filipowicz et al. studied the how emotional transitions influence the social interactions. They observed that it could lead to different interpersonal

[1] In this chapter, the word 'interlocutor' encompasses the person speaking and expressing (*speaker*) the emotion and the person listening and perceiving (*listener*) that emotion.

behaviors. Regarding the research for designing automatic emotion recognition systems, several behavioral cues have been explored using various modalities such as audio and video, and multi-modal [27, 30, 31, 44].

The importance of summarizing spoken conversations, in terms of emotional manifestations, long-term traits, and conversational discourse, is presented in [48]. Authors demonstrated that such a summary can be useful in different application domains.

The research related to the use of paralinguistics for behavioral analysis from spoken conversations include personality [2–4], overlap discourse [12, 14], overlap in context [15], turn-taking dynamics [11], user satisfaction [16], and interlocutors' coordination in emotional segment [5].

In our study, we focus on designing the *affective scene framework* to automatically label affective scenes. The *motivation* of designing an affective scene framework is also practical because from the automatically labeled affective scenes one can interpret the different patterns of the speakers' affective behavior in situated contexts. For the experiment and evaluation described in this chapter, we utilized call center's dyadic spoken conversations. We investigate the call center agent and customer's affective behaviors such as *customer manifested anger or frustration, however, the agent was not empathic* (see Sect. 6.4). This kind of understanding can help to pinpoint the customer's problem, and it can also help in reducing phenomena like the churn rate.[2] Such an understanding can help to provide a better customer service. We propose the computational models that can automatically classify emotional segments occurring in conversations. For the design and evaluation of the model, we utilize *verbal* and *vocal non-verbal* cues in terms of acoustic, lexical, and psycholinguistic features. The work described in this chapter may contribute to the cognitive infocommunications research field by providing a model of affective scenes that may be adopted to investigate intra-cognitive communications [6] issues.

This chapter is organized as follows. In Sect. 6.2, we briefly review the terminologies used in the affective computing research, which are relevant for this work. In Sect. 6.3, we give a brief overview of the corpus that we used to investigate the affective scene (Sect. 6.4) and design the framework. We present the experimental details in Sect. 6.5, and discuss the results in Sect. 6.6. Finally, we provide conclusions in Sect. 6.7.

## 6.2 Terminology

In this section, we review terminologies from affective behavior research that are relevant for the context of this work.

[2]The *churn rate* is "the percentage of customers who stop buying the products or services of a particular company." In the telecommunication industry, some studies found that the approximate annual churn rate is 30% [24, 49].

**Behavior**: It is defined as "... quite broadly to include anything an individual does when interacting with the physical environment, including crying, speaking, listening, running, jumping, shifting attention, and even thinking" [22].

**Behavioral Signals**: Signs that are direct manifestations of individual's internal states being affected by the situation, the task and the context. Signals can be overt and/or covert. Examples of overt signals are changes in the speaking rate or lips getting stiff. Examples of covert cues are changes in the heart-rate, galvanic skin response or skin temperature.

**Affect**: It is an umbrella term that covers a variety of phenomena that we experience such as emotion, stress, empathy, mood, and interpersonal stance [29, 41]. All of these states share a special affective quality that sets them apart from the neutral states. In order to distinguish between each of them, Scherer [42] defined a design-feature approach, which consists of seven distinguished dimensions, including intensity, duration, synchronization, event-focus, appraisal elicitation, rapidity of change, and behavioral-impact.

**Affective Behavior**: The component of the behavior that can be explained in terms of affect analysis.

**Emotion**: There is a variety of definitions of this concept. A few of them are reported below. According to Scherer [42], emotion is a relatively brief and synchronized response, by all or most organismic subsystems, to the evaluation of an external or internal stimulus.

Gross's [26] definition of emotion refers to its *modal model*, which is based on three core features such as (1) what gives rise to emotions (when an individual attend and evaluate a situation), (2) what makes up an emotion (subjective experience, behavior, and peripheral physiology), and (3) malleability of emotion.

According to Frijda "emotions are intense feelings that are directed to someone or something" [23, 41].

**Emotional State**: In the psychological and psychiatric literature the concept "emotional state" is often used as being coextensive with some given emotion, and it is not explicitly defined from itself. However, studies that analyze human emotions in a more situated perspective, have proposed that emotional states are conditions of the psychological and physiological processes that generate an emotional response, and that contextualize, regulate, or otherwise alter such a response [7].

**Empathy:** According to Hoffman [28], "Empathy can be defined as an emotional state triggered by another's emotional state or situation, in which one feels what the other feels or would normally be expected to feel in his situation". For this work, we follow this definition of empathy.

By McCall and Singer [34], empathy is defined based on four key components: "*First*, empathy refers to an affective state. *Second*, that state is elicited by the inference or imagination of another person's state. *Third*, that state is isomorphic with the other person's state. *Fourth*, the empathizer knows that the other person is the source of the state. In other words, empathy is the experience of vicariously feeling what another person is feeling without confounding the feeling with one's own direct experience".

Perry and Shamay-Tsoory [37] "... denotes empathy as our ability to identify with or to feel what the other is feeling."

## 6.3 The Corpus and Its Annotations

The corpus of Italian conversations analyzed in this study consists of 1894 customer-agent phone dialogues amounting ∼210 h. It was recorded on two separate channels with 16 bits per sample, and a sampling rate of 8 kHz. The length of the conversations is 406 ± 193 (mean ± standard deviation) seconds. The corpus was annotated with *empathy (Emp)*, basic and complex emotions. The basic emotions include *anger (Ang)*, and complex emotions include *frustration (Fru), satisfaction (Sat), dissatisfaction (Dis)*. We also introduced *neutral (Neu)* state tag as a relative concept to support annotators while identifying the emotions in the conversational context. The annotation protocols were defined based on the Gross's *modal model* of emotion. More details of the annotation process can be found in [17]. The annotators' task was to identify the occurrences of emotional segments in the continuous speech signal where they can perceive a transition in the emotional state of the speaker. They identified the onsets of the variations and assigned an emotional label. Hence, the emotional segment may consist of one or more turns.[3] Moreover, the annotators were also instructed to focus both on their perception of speech variations, such as acoustic and prosodic quality of pairs of speech segments, and on the linguistic content of the utterances.

For the evaluation of the annotation model, we randomly selected 64 (∼448 min) spoken conversations. The annotators were hired for the annotation task. The demographic information of the annotators consists of Italian native speakers, similar age, and ethnicity but different gender. The goal was to assess whether the annotators can perceive the speech variations at the same onset position, as well as their agreement of assigning the emotional labels on speech segments. The comparison showed that 0.31 of the annotated speech segments were exactly tagged by the two annotators at the same onset positions, while 0.53 was the percentage of cases where the two annotators perceived the emergence of an emotional attitude of the speaker occurring at different, yet contiguous, time frames of the same dialog turns.

To measure the quality of the annotations, we calculated inter-annotator agreement using the kappa measure [8]. It is widely used to assess the degree of agreement among the annotators. The kappa coefficient ranges between 0 (i.e., agreement is due to chance) and 1 (i.e., perfect agreement). Values above 0.6 suggests an acceptable agreement. We found reliable kappa results as shown in Table 6.1. The annotation task was complex, which was also taking more time for the annotation. From our observation, we found that on average the annotation time of a conversation was ∼18 min. The analysis of annotators' disagreement showed that the inevitable portion of the subjectivity of this task had a greater impact in cases where the differences of the emotional labels were more nuanced, like in case of frustration and dissatisfaction.

[3]Turn refers to the spoken content of a speaker at a time. For example, speaker A says something, which is speaker A's turn, then, speaker B says something, which is speaker B's turn.

**Table 6.1** Kappa results of the annotation

| Emotional state | Agent/Customer | Kappa |
|---|---|---|
| Empathy | Agent | 0.74 |
| Anger | Customer | 0.75 |
| Frustration | Customer | 0.67 |
| Satisfaction | Customer | 0.69 |
| Dissatisfaction | Customer | 0.71 |

## 6.4 Affective Scene

In [17], Danieli et al. defined the *affective scene* on the basis of the Gross's *modal model* of emotion [25, 26]. The concept of ***affective scene*** is defined as "*an emotional episode where an individual is interacting, in an environment, with a second individual and 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 of the 'unfolding of emotions' throughout the closure event when individuals disengage themselves from the communicative context. To describe the affective scenes, we have considered three factors:

- *who* showed the variation of their emotional state *when*,
- *how* the induced emotion affected the other interlocutor's emotional appraisal and response, and
- *which* modifications occurred by such a response with respect to the state that triggered the scene.

In Fig. 6.1, we present an example of an emotional sequence, which shows an affective scene. In the example, we see there is a sequence of three events.

1. *who:* customer 'first' manifested frustration.
2. *how:* agent appraised customer's emotion and responded with empathic behaviors.
3. *which:* customer manifested satisfaction at the end of the conversations.

The lack of empathic response from the agents may cause different patterns of emotional response from the customer. It can lead to the manifestation of customer's anger, dissatisfaction, or frustration.

From the manual annotations of our corpus, we selected a subset containing 566 conversations where we investigated different patterns of affective behaviors. Some of the selected patterns are presented below.

- agents were neutral when customers manifested either anger, frustration or dissatisfaction.
- agents were neutral when customers manifested either anger, frustration or dissatisfaction followed by satisfaction at the end of the conversations.

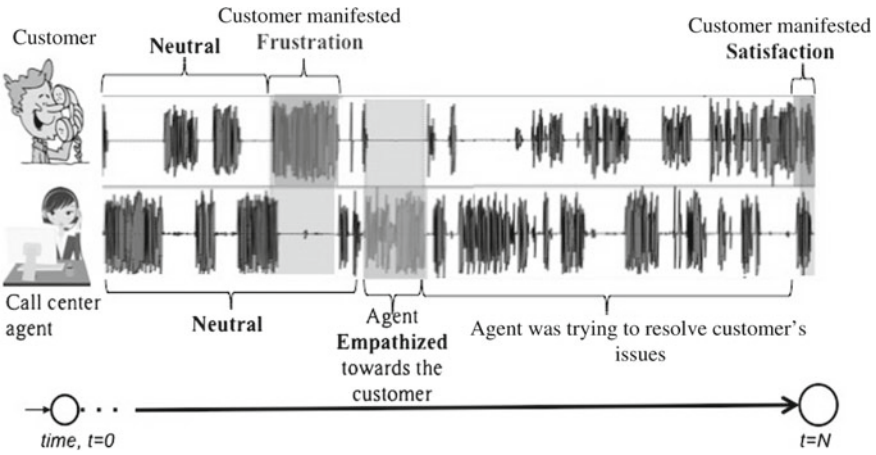


**Fig. 6.1** An annotated example of a call center conversation. The interaction between call center's agent and customer shows that customer first manifested frustration, then agent empathized towards the customer, and finally customer manifested satisfaction at the end of the conversation

- agents were empathic, and customer manifested either anger, frustration or dissatisfaction.
- agents were empathic, and customer manifested either anger, frustration or dissatisfaction followed by satisfaction at the end of the conversations.
- agents were empathic, and customer manifested satisfaction.

In Fig. 6.2, we depict the distribution of the behavioral patterns between speakers that we mentioned above. In the figure, *A:Emp-C:Neg* represents agents (A) were empathic (Emp), and customers manifested negative emotions and *C:Neg,Sat* rep-

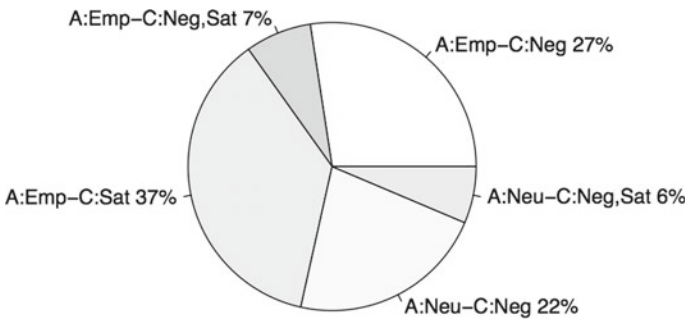


**Fig. 6.2** A set of conversations consisting of different patterns of affective behavior analyzed from emotional state sequence. For instance, the pattern (A:Emp-C:Neg,Sat) indicates that, in this corpus, there were 7% of conversations where agents showed empathy and the customer negative emotions, however, at the end they were satisfied. This pattern can explain the success of the operator in handling the state of the customer

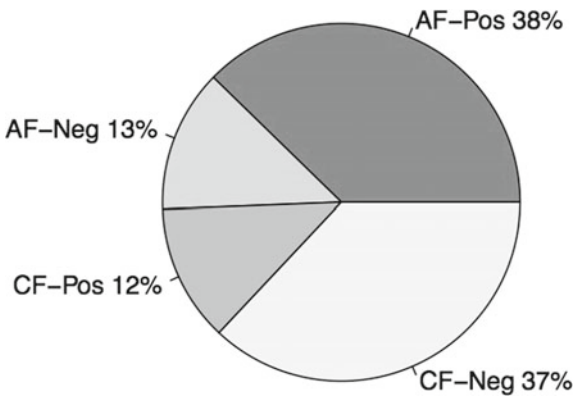


**Fig. 6.3** The proportion of conversations, in which either agent (AF) or customer (CF) expressed emotion at the start of the conversation, and the manifestation of customer's emotional state at the end of the conversation. *Pos* and *Neg* represent customer manifested either positive or negative emotion

resents customer manifested negative emotions followed by satisfaction at the end of the conversations. For the sake of this analysis, we grouped anger, frustration, and dissatisfaction into negative. From this analysis, we see that in 22% conversations customers manifested negative emotions whereas agents were neutral. In 27% conversations, even if the agent were empathic, but customers manifested negative emotions. These two scenarios might be an important factor for call center managers. Manual intervention might be necessary for these conversations. Other scenarios are also interesting to look into, for example, the pattern *A:Emp-C:Neg,Sat* occurred in 7% conversations, where customers' manifestation of satisfaction followed customers' negative emotions. It may be agents were empathic, and customer's emotional arousal reduced over the course of the conversations.

The other patterns that we wanted to investigate are the manifestation of emotions at the start, and at the end of the conversations. It includes *who* is manifesting emotion at the start of the conversations, i.e., agent first (AF) or customer first (CF), and *which* types (positive (Pos) or negative (Neg)) of emotions customer were manifesting at the end of the conversations. In Fig. 6.3, we present such an analysis. The critical part is that in 13% (AF-Neg) and 37% (CF-Neg) conversations customer manifested negative emotions at the end of the conversations. We observed that customer manifestations of negative emotion at the start of the conversations might lead to the higher chance of the manifestations of negative emotions at the end of the conversations. In this analysis, we did not consider what is the emotional manifestations of the agents at the end of the conversations. Such analytical results indicate that more insights can be found from affective scenes, which leads to the fact that it is necessary to automatically detect affective scenes.

## 6.5 Affective Scene Framework

In Fig. 6.4, we present the proposed affective scene framework. Several architectural decisions can be made, however, we found this approach is a good starting point towards the automation of the affective scenes in conversations. The whole architectural pipeline can also be useful to process visual signals too. As shown in Fig. 6.4, at first we separated the conversation channel-wise. This process can be done in two different manners. Either two channels will input independently to the system, or a classification module can be designed to separate the two speakers' speech channels. After that, an automatic segmenter was used to separate speech and non-speech segments (see Sect. 6.5.1). Then emotion segment classifiers (Sect. 6.5.2) were employed, which automatically assign each speech segment with an emotion label. After that, using the emotion sequence labeler module (see Sect. 6.5.3), we combined the segment classifiers' output and segment meta information, i.e., start and end time boundary of the segment, to design emotional sequence for the whole conversation.

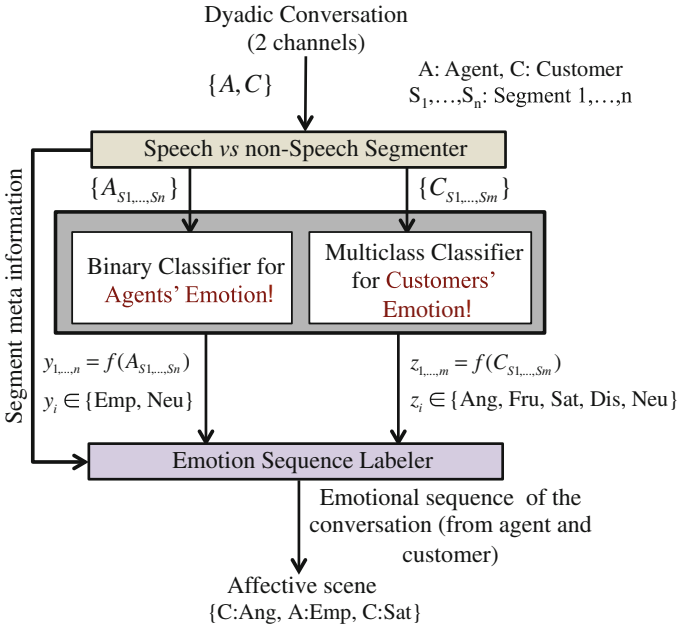


**Fig. 6.4** The architecture of the proposed automatic affective scene classification

### 6.5.1 Classification of Speech and Non-speech Segments

To classify speech and non-speech segments within a conversational channel we used an off-the-shelf speech and non-speech segment classifier. It has been designed using forced aligned transcriptions and trained the model using Hidden Markov Models (HMMs). The transcriptions were obtained from 150 conversations, which consists of approximately 100 h of speech. To train the model, Mel Frequency Cepstral Coefficient (MFCC) features were extracted from the conversations with 25 ms per frame and 100 frames per second with a step size of 10 ms. The model was trained using 32 Gaussian mixtures with a beam size 50, and utilized Kaldi [39] speech recognition framework. The F-measure of the segment classifier is 66.42% [1].

### 6.5.2 Segment Classifiers

We designed separate emotion classifiers to deal with customer and agent channels' speech segments. For the agent channel, we designed a binary classifier in which the class labels include *empathy (Emp)*, and *neutral (Neu)*. Whereas for the customer channel we grouped anger and frustration into negative, then designed a multiclass classifier in which class labels are *negative (Neg)*, *dissatisfaction (Dis)*, *satisfaction (Sat)*, and *neutral (Neu)*. We investigated each system using acoustic, lexical, and psycholinguistic features, also with their decision-level combination. For the experiment, we separated the conversations into the training, development, and test set and their proportion was 0.70, 0.15, and 0.15, respectively. We separated them at the conversation level while we maintained speaker independence for the agent conversations. Due to the unavailability of the customer information in the corpus, we assumed that each conversation is from an independent speaker.

#### 6.5.2.1 Feature Extraction

Acoustic features

We extracted acoustic features from speech signal using openSMILE [19] and the configurations, which we made publicly available.[4] The approach of the feature extraction process is that we first extract low-level features and then project them onto statistical functionals. The success of this approach has been reported in the literature in which studies have been conducted in different paralinguistic tasks [3, 46, 47].

We extracted low-level acoustic features with 100 frames per second. The frame size for voice-quality features was 60 ms with a Gaussian window function and $\sigma = 0.4$. For the other low-level features, the frame size was 25 ms with a Hamming

[4] https://github.com/firojalam/openSMILE-configuration.

window function. The low-level features include zero crossing rate, MFCC (1–12), root mean square frame energy, fundamental frequency, pitch, harmonics-to-noise ratio, spectral features, voice probability, Mel-spectrum band 1–26, centroid, max, min, and flux.

We apply many statistical functionals, here, we report a few of them. More details can be found in [1]. The set includes range, max, min, the geometric and quadratic mean, linear and quadratic regression coefficients, arithmetic and quadratic mean, quartiles and interquartile range and percentile.

After applying statistical functionals onto low-level features, the resulted total number of acoustic features consists of 6861.

Lexical features

We extracted lexical features from automatic transcriptions, and we used an in-house developed ASR system [13] to get the transcriptions. The word error rate of the ASR system is 31.78% on the test set. To understand how the results differ from training data we also evaluated the system on the training set and obtained an WER of 20.87%. To design the classification model, the textual information needs to be converted into vector form and the widely used approach is bag-of-words or bag-of-ngrams. For this work, we converted the transcriptions of each segment into bag-of-words vectors. The applied logarithmic term frequencies (tf) and inverse document frequencies (idf) method. Since contextual information provides better classification results, therefore, we extracted trigram features. The n-gram approach results in a large dictionary, which increase computational cost and also introduce overfitting. To avoid such drawbacks, we removed the stop-words and filtered out lower frequent words and only kept the 10 K most frequent n-grams.

Psycholinguistic features

We used LIWC [36] to extract the psycholinguistic features from the automatic transcriptions. It is a knowledge-based system containing dictionaries for several languages in which word categories are associated with a set of a lexicon. The system computes frequency or relative frequency for each word category by matching the transcriptions. The word categories are used as features for the machine learning tasks. The word categories include linguistic, psychological, paralinguistic, personal concern and punctuations. Since the transcriptions were Italian, therefore, we used Italian dictionary, which contains 85 word categories. Moreover, the LIWC system extracts 6 general and 12 punctuation categories, which results in 103 features. We removed features that are not found in our dataset. For example, punctuations are not available with transcriptions. Hence, we extracted 89 psycholinguistic features.

#### 6.5.2.2 Feature Selection

To reduce the dimensionality, we applied *Relief* feature selection algorithm [32]. The motivation of using this technique is that it showed improved classification

performance and at the same time it reduced computational cost. We observed that in previous paralinguistic tasks [2]. Our feature selection process is as follows. We rank the feature set according to Relief feature selection scores. Then, we generate learning curves by incrementally adding batches (e.g. we added 200 features each time) of ranked features. After that, we select the optimal set of features [2]. To obtain a better performance during the feature selection process, we discretize the feature values into 10 equal-frequency bins [50]. The main reason is that the feature selection algorithm does not work well with the continuous-valued features. The approach of equal-frequency binning is that it divides data into $k$ bins, where each bin contains an equal number of values. For our experiment, we have chosen to use the value of k as 10. To get an unbiased estimate on the classification results, we experimented feature selection and discretization approaches on the development set.

#### 6.5.2.3 Classification Experiments and Evaluation

Classification method

We designed the classification models using SVM [38] and applied linear kernel for lexical and acoustic features. For the psycholinguistic features, we applied Gaussian kernel of the SVM. We optimized the parameters $C$ and $G$ by tuning it on the development set. The range of values that we experiment for $C$ and $G$ is $[10^{-5}, 10^{-4}, ..., 10]$. In order to get the results on the test set, we first combined the training and development dataset. After that, we trained the models using the optimized parameters. Since we have three classifier's decisions, therefore, to get a single decision we applied *majority voting*.

Undersampling and oversampling

One of the important problems in designing machine learning model is the imbalance class label distribution. For example, the original distribution of the Emp versus Neu segments was 6% and 94% respectively. The original distribution of customer's channel emotions such as Neg, Dis, Sat, and Neu were 1%, 1%, 2% and 96%, respectively. To deal with such a problem, we undersampled the instances of majority classes at the data level[5] and oversampled the instances of minority classes at the feature level.[6] Our undersampling approach is very intuitive in a sense that we capture the variation of segment length. To do that, we defined a set of bins, which varies segment lengths. After that, we randomly selected $N$ segments from each bin. The size of $N$ is experimental and problem specific. For this study, we used the size of $N$ as 1. However, we found that this is an optimal number, which we obtained by running experiments on the development set. The pseudocode of undersampling approach is presented in Algorithm 1. The boundaries i.e., start and end, of each

[5]By data level, we refer to the data preparation phase, i.e., before feature extraction we select segments of the majority class, which is neutral in this case.

[6]By feature level, we refer to the over-sampling process on feature vector for minority classes.

bin in this study are set based on the experimental observation. It can also be done based on percentiles of segments' length *or* applying a supervised approach such as multi-interval discretization by Fayyad and Irani [20]. With such an approach we are providing the variation of segment length as well as reducing the samples and influence of majority class.

**Algorithm 1** Undersampling algorithm, C- Conversation containing segments, N-number of segment from each bin of segment duration, l-class label to be undersampled. It returns randomly selected segments. Bin can be designed in several way, such as equal frequency or equal interval.

**procedure** DOWN- SAMPLING(C, N, l)
  $newSegmentList \leftarrow List$
  $segmentLengthBin \leftarrow \{0.1-4.0, 4.0-7.0, 7.0-10.0, 10.0-15.0, >15.0\}$
  $segIndex = 0$
  **for all** $seg \in C$ **do**
    **if** $seg.tag = l$ **then**
      $dur \leftarrow duration(seg)$
      $segmentLengthBin[dur] \leftarrow seg$
    **else**
      $newSegmentList[segIndex] \leftarrow seg$
      $segIndex = segIndex + 1$
    **end if**
  **end for**
  **for all** $bin \in segmentLengthBin$ **do**
    $bin \leftarrow shuffle(bin)$ ▷ Shuffles the bin elements for randomization
    **for** $i = 1 : N$ **do**
      $newSegmentList[segIndex] \leftarrow bin[i]$
      $segIndex = segIndex + 1$
    **end for**
  **end for**
  $nC \leftarrow newSegmentList$
  **return** $nC$ ▷ Set of segments randomly selected from C's segments.
**end procedure**

For the oversampling, we used SMOTE algorithm [10] and applied it on the training set. After applying the sampling techniques, the distribution for empathy versus neutral segments becomes 30% and 70%, respectively, and the distribution for Neg, Dis, Sat, and Neu becomes 7.5%, 7.5%, 12% and 73%, respectively. Such distributions are still very skewed, however, more balancing does not help in classification performance, which we experimented with the development set. The reason is with too much under-sampling of the majority class we are reducing the variations of the training instances and classifier does not capture much information. On the other hand too much synthetic generation of oversampling, we are generating similar instances and classifier does not gain much information about the minority classes.

Evaluation metric

To measure the performance of the system we used an well-known metric Unweighted Average (UA) recall. It is a widely accepted metric for the paralinguistic tasks [45]. We have extended this measure to take into account the segmentation errors. Similar type of evaluation is typically done by NIST in diarization tasks [9, 35]. UA is the average recall of class labels, which we computed from a weighted confusion matrix, as shown in Eq. 6.1, where the weight for each segment is the corresponding segment length.

$$C(f) = \left\{c_{i,j}(f) = \sum_{s \in S_T} [((y = i) \wedge (f(s) = j)) \times length(s)]\right\} \tag{6.1}$$

where $C(f)$ is the $n \times n$ confusion matrix of the classifier $f$, $s$ is the segment, $length(s)$ is the duration of $s$, $y$ is the reference label of $s$, $f(s)$ is the automatic label for $s$. The indices $i$ and $j$ are the reference and automatic/classified class labels of the confusion matrix.

### 6.5.3 *Affective Scene Labeling*

Using the *emotion sequence labeler* module as shown in Fig. 6.4, we combined the emotion sequence from both agent and customers' channels by utilizing the output of the emotion segment classifier and the meta information of segments from the speech versus non-speech segmenter. An example of such a sequence is as follows, *C: Fru* $\rightarrow$ *A: Emp* $\rightarrow$ *C: Sat* where customer 'first' manifested emotion, then agent empathized towards the customer, after that customer's emotion regulated towards satisfaction, which is a similar representation we presented in Fig. 6.1.

## 6.6 Results and Discussion

In Table 6.2, we present the performances of the emotion segment classification systems, which we investigated using different types of features such as acoustic, lexical (automatic transcriptions), and psycholinguistic, and also their decision level combination. To compute the baseline, we have randomly selected the class labels based on the prior class distribution. In addition, we also present average results (Avg) to get an understanding of the performance of the whole system.

We obtained the best results with *majority voting* for both segment classification systems. To understand whether the obtained best results are statistically significant or not we run the significant test. For which we used McNemar's test. From the results, we observed that the results of majority voting are statistically significant with $p < 0.05$ compared to any other classifier's results. Moreover, compared to the baseline, a relative improvement of the system with agent's emotions is 42.2%, and

**Table 6.2** Segment classification results in terms of weighted average recall using acoustic, lexical, and psycholinguistic features together with decision level fusion. Ac—acoustic features; Lex—lexical features; LIWC—psycholinguistic features; Maj—majority voting. Values inside parenthesis represent feature dimension

| Exp | Agent | Customer | Avg |
|---|---|---|---|
| | Binary {Emp, Neu} | Multiclass {Neg, Dis, Sat, Neu} | |
| Baseline | 49.3 | 24.4 | 36.9 |
| Ac | (2400) 68.1 | (4600) 47.4 | 57.7 |
| Lex | (8000) 65.6 | (5200) 56.5 | 61.0 |
| LIWC | (89) 67.3 | (89) 51.9 | 59.6 |
| Ac+Lex | (2600) 68.3 | (5800) 49.2 | 58.8 |
| Maj(Ac+Lex+LIWC) | 70.1 | 56.9 | 63.5 |

the system with the customer's emotions is 61.51%. The performance of the system with customer's emotions is lower compared to the other system, which is due to the complexity of the task in terms of class distributions, and multiclass classification task.

### *6.6.1 Empathy*

From the Table 6.2, we see that for individual feature-based models, the model using acoustic features provides the best performance compared to models using lexical and psycholinguistic features, respectively. The results of the acoustic features are significantly better than random baseline with a $p$ value less than $2.2E-16$. When there are no transcriptions (i.e., either manual or automatic) available then the use of acoustic features is the ideal situation as it provides a useful and low-computation classification model. The performance of the LIWC's features is better than lexical features. The advantage of this feature set is that its size is very small, i.e., a number of features is 89, which is computationally less expensive. Compared to the baseline results, the performances of all classification systems are higher and statistically significant with a $p$ value less than $2.2E-16$.

In terms of feature and decision level combination, we obtained the best results using *majority voting*. From the statistical significance test, we observe that the results of the majority voting are statistically significant with a $p$ value equal to 0.0004 compared to any other classification models' results. We also conducted experiments how linear combination of acoustic and lexical features performs based on the previous findings in other paralinguistic tasks [3]. We observed that it has not improved performance as shown in Table 6.2.

In Figs. 6.5 and 6.6, we present class-wise correlation analysis on acoustic and lexical features, which we performed correlation analysis on top-ranked features. To

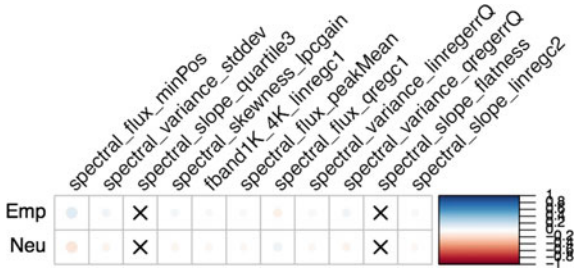


**Fig. 6.5** Correlation analysis of acoustic features. Cells without asteric "×" are statistically significant. The color in each cell represents the correlation coefficients (r) and its magnitude is represented by the depth of the color. The "×" symbol represents the corresponding r is not significant. Description of the features is presented in Table 6.3

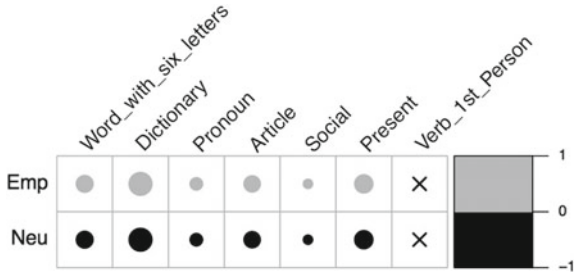


**Fig. 6.6** Correlation analysis of LIWC features. Cells without asteric "×', are statistically significant. The color in each cell represents the correlation coefficients (r) and its magnitude is represented by the depth of the color. The "×" symbol represents the corresponding r is not significant

rank the features we used a feature selection algorithm discussed in Sect. 6.5.2.2. From the feature selection and correlation analysis, our findings suggest that spectral features contribute most to the classification decision. The top-ranked low-level acoustic feature includes spectral, energy, and mfcc in ranked order and the statistical functionals include minimum segment length, standard deviation, quadratic mean and quadratic regression coefficient.

From the Fig. 6.5, we observe that spectral-flux feature is positively correlated with empathy and is negatively correlated with neutral. In the figure, the "×" symbol represents they are not statistically significant. From the figure, we see that even if spectral-slope features appeared in top 10 features, however, they are not highly correlated with any class label. From our analysis of psycholinguistic features, we observed that character length, dictionary word, pronoun, article, social connotation, words convey present tense are positively correlated with empathy and negatively correlated with neutral.

Our findings from lexical feature analysis suggest some words and phrases are positively correlated with empathy and negatively correlated with neutral such as

**Table 6.3** Selected acoustic features and their description obtained from empathy segment

| Feature | Description |
|---|---|
| spectral_flux_minPos | Absolute position of the minimum value of the spectral flux feature |
| spectral_variance_stddev | The standard deviation of the spectral variance |
| spectral_slope_quartile3 | The third quartile (75% percentile) of the spectral slope |
| spectral_skewness_lpcgain | The linear predictive coding gain of spectral skewness |
| fband1K_4K_linregc1 | The slope of a linear approximation of the fband 1000–4000 contour |
| spectral_flux_peakMean | The arithmetic mean of peak spectral flux |
| spectral_flux_qregc1 | The quadratic regression coefficient 1 of the spectral flux |
| spectral_variance_linregerrQ | The quadratic error of the spectral variance, computed as the difference of the linear approximation and the actual contour |
| spectral_variance_qregerrQ | The quadratic error of the spectral variance, computed between contour and quadratic regression line |
| spectral_slope_flatness | The contour flatness of spectral slope, which is a ratio of geometric mean and absolute value of arithmetic mean |
| spectral_slope_linregc2 | The offset of a linear approximation of the contour of spectral_slope |

"posso aiutarla/can I help you", "vediamo un po/let's see". An example of a positive correlation with neutral is "cosa posso esserle/what can I do".

In another study, we have done an experiment with manual segments, which shows that we can reach above 90% UA with a better segmentation approach.

### *6.6.2 Basic and Complex Emotions*

The classification task for basic and complex emotions is much more complex than empathy classification task due to the multi-class classification task. From the results, as shown in Table 6.2, we observed that the performance of LIWC features is lower than lexical features. However, it is higher than acoustic features, and the number of features is very low for this set. The decision level combination has not improved the performance for this case, whereas it shows higher improvement for the classification of empathy. For the linear combination of acoustic and lexical features performance is also lower compared to the lexical features. The feature dimension for the lexical feature set is comparatively higher than the acoustic and LIWC feature sets.

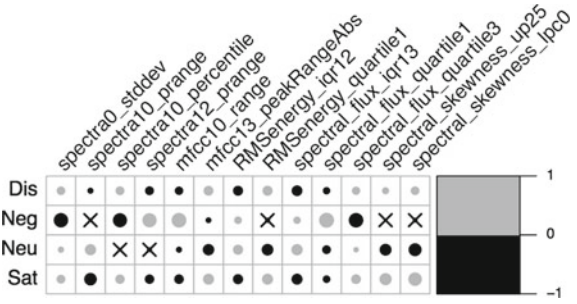


**Fig. 6.7** Correlation analysis of acoustic features Cells without asteric "×" are statistically significant. The color in each cell represents the correlation coefficients (r) and its magnitude is represented by the depth of the color. The "×" symbol represents the corresponding r is not significant. Description of the features is presented in Table 6.4

We observed that the recall of dissatisfaction is comparatively lower than other emotional categories. It also confuses with satisfaction due to the fact the manifestation of both satisfaction and dissatisfaction appear at the end of the conversation. For this reason, there is an overlap of the linguistic content, which also effects the paralinguistic properties of the spoken content. Using acoustic features, we obtained better performance for negative and neutral, whereas using the lexical feature we obtained better performance for negative and satisfaction. In terms of discriminative characteristics, spectral, voice-quality, pitch, energy, and mfcc features are highly important. The statistical functionals include the arithmetic mean of the peak, quadratic regression, the gain of linear predictive coefficients, flatness, quartile, and percentiles.

We have analyzed them in terms of class-wise correlation analysis as presented in Fig. 6.7. The number in each cell in the figure represents the correlation value

**Table 6.4** Selected acoustic features and their description obtained from the segments of basic and complex emotion

| Feature | Description |
|---|---|
| spectra0_stddev | The standard deviation of the rasta style auditory spectra band 0 |
| spectra10_prange | The percentile range of the rasta style auditory spectra band 10 |
| spectra10_percentile | The percentile of the rasta style auditory spectra band 10 |
| spectra12_prange | The percentile range of the rasta style auditory spectra band 12 |
| mfcc10_range | The range of the derivative of the mfcc 10 |
| mfcc13_peakRangeAbs | The absolute peak range of the derivative of the mfcc 13 |
| RMSenergy_iqr12 | The inter-quartile range: quartile2-quartile1 of the root-mean square energy 12 |
| RMSenergy_quartile1 | The first quartile (25% percentile) of the root-mean square energy |
| spectral_flux_iqr13 | The inter-quartile range: quartile3-quartile1of the spectral flux |
| spectral_flux_quartile1 | The first quartile (25% percentile) of the spectral flux |
| spectral_flux_quartile3 | The third quartile (75% percentile) of the spectral flux |
| spectral_skewness_up25 | The percentage of time the signal is above (25% * range + min) of the spectral skewness |
| spectral_skewness_lpc0 | The linear predictive coding of spectral skewness |

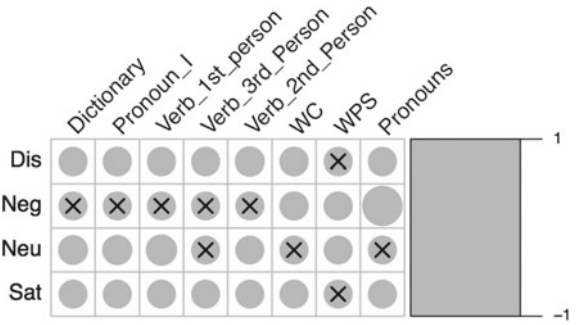


**Fig. 6.8** Correlation analysis of LIWC features. Cells without asteric "×" are statistically significant. The color in each cell represents the correlation coefficients (r) and its magnitude is represented by the depth of the color. The "×" symbol represents the corresponding r is not significant

between class-label and a feature. The color in each cell represents the positive and negative association. The "×" symbol represent the association is not significant with $p = 0.05$. Even though the correlation value is very low, close to zero, however, most of them are statistically significant. Spectral and rasta style auditory spectram features are positively associated with satisfaction. For neutral, spectral features are negatively associated. MFCC and rasta style auditory spectram features are positively correlated with negative emotion. Satisfaction and dissatisfaction are mostly similar; the only dissimilarity exists in the strength of positive and negative association in some features.

The correlation analysis of LIWC features is presented in Fig. 6.8. The highly discriminative LIWC features include personal pronouns, words associated with emotion and verb. Similar to the acoustic features, satisfaction and dissatisfaction are quite similar in their correlation with LIWC features. However, there exists a disassociation in the strength of the correlation. First three features, such as words containing in the dictionary, pronoun (I), and 1st person verb, are negatively correlated with neutral, and these features are positively correlated with the negative emotion.

The correlation analysis of lexical features shows that negative emotion is highly associated with negative words whereas satisfaction represents mostly positive words such as "grazie mille/thank you so much/", "benissimo/very well/" and "perfetto/perfect/". The difference between satisfaction and dissatisfaction is that dissatisfaction represents some negativity, however, there is not much lexical difference. It might be because the annotators mostly focused on the tone of the voice, which distinguished the annotation of satisfaction and dissatisfaction. It is needed to mention that the LIWC and lexical feature analysis has been done based on ASR transcription.

To understand the upper-bound of the classification system, we designed a system by exploiting manual segments with which we obtained UA 70.9% using acoustic features.

Note that, we do not present any figure of the correlation analysis of lexical features as it is very difficult to make a general conclusion from such graphical representation.

## 6.7 Conclusions

In this chapter, we address the design of an affective scene system to automatically label the sequence of emotional manifestations in a dyadic conversation. Our presented framework shows the complete pipeline from speech segmentation to sequence labeling, which might be useful in analyzing affective and behavioral patterns in different applicative scenarios. The results of the automatic classification system using decision combination on call center conversations are significantly better than random baseline. We investigated different feature sets such as acoustic, lexical, and psycholinguistic features. Moreover, we also investigated feature and decision level combinations for designing emotion segment classifiers. The investigation of the feature sets suggests that lexical and psycholinguistic features can be useful for the automatic classification task. In both emotion segment classification systems, we obtained better results using decision combination. In future, as an extension of this research, one can explore the study of affective scenes in multi-party contexts and social domains.

## References

1. Alam F (2016) Computational models for analyzing affective behaviors and personality from speech and text. PhD thesis, University of Trento
2. Alam F, Riccardi G (2013) Comparative study of speaker personality traits recognition in conversational and broadcast news speech. In: Proceedings of interspeech, ISCA, pp 2851–2855
3. Alam F, Riccardi G (2014) Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In: Proceedings of international conference on acoustics, speech and signal processing (ICASSP), pp 955–959
4. Alam F, Riccardi G (2014) Predicting personality traits using multimodal information. In: Proceedings of the 2014 ACM multi media on workshop on computational personality recognition, ACM, pp 15–18
5. Alam F, Chowdhury SA, Danieli M, Riccardi G (2016) How interlocutors coordinate with each other within emotional segments? In: COLING: international conference on computational linguistics
6. Baranyi P, Csapó Á (2012) Definition and synergies of cognitive infocommunications. Acta Polytech Hung 9(1):67–83
7. Barrett LF, Lewis M, Haviland-Jones JM (2016) Handbook of emotions. Guilford Publications
8. Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 22(2):249–254
9. Castán D, Ortega A, Miguel A (2014) Lleida E (2014) Audio segmentation-by-classification approach based on factor analysis in broadcast news domain. EURASIP J Audio Speech Music Process 1:1–13
10. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority oversampling technique. J Artif Intell Res 321–357
11. Chowdhury SA (2017) Computational modeling of turn-taking dynamics in spoken conversations. PhD thesis, University of Trento
12. Chowdhury SA, Riccardi G (2017) A deep learning approach to modeling competitiveness in spoken conversation. In: Proceedings of international conference on acoustics, speech and signal processing (ICASSP), IEEE

13. Chowdhury SA, Riccardi G, Alam F (2014) Unsupervised recognition and clustering of speech overlaps in spoken conversations. In: Proceedings of workshop on speech, language and audio in multimedia—SLAM2014. pp 62–66
14. Chowdhury SA, Danieli M, Riccardi G (2015) Annotating and categorizing competition in overlap speech. In: Proceedings of ICASSP. IEEE
15. Chowdhury SA, Danieli M, Riccardi G (2015) The role of speakers and context in classifying competition in overlapping speech. In: Sixteenth annual conference of the international speech communication association
16. Chowdhury SA, Stepanov E, Riccardi G (2016) Predicting user satisfaction from turn-taking in spoken conversations. In: Proceedings of Interspeech
17. Danieli M, Riccardi G, Alam F (2015) Emotion unfolding and affective scenes: a case study in spoken conversations. In: Proceedings of emotion representations and modelling for companion systems (ERM4CT) 2015. ICMI
18. Devillers L, Vidrascu L (2006) Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In: Proceedings of Interspeech. pp 801–804
19. Eyben F, Weninger F, Gross F, Schuller B (2013) Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia (ACMM). ACM, pp 835–838
20. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuousvalued attributes for classification learning. Thirteenth international joint conference on articial intelligence, vol 2. Morgan Kaufmann Publishers, pp 1022–1027
21. Filipowicz A, Barsade S, Melwani S (2011) Understanding emotional transitions: the interpersonal consequences of changing emotions in negotiations. J Pers Soc Psychol 101(3):541
22. Fisher W, Groff R, Roane H (2011) Applied behavior analysis: history, philosophy, principles, and basic methods. In: Handbook of applied behavior analysis, pp 3–13
23. Frijda NH (1993) Moods, emotion episodes, and emotions
24. Galanis D, Karabetsos S, Koutsombogera M, Papageorgiou H, Esposito A, Riviello MT (2013) Classification of emotional speech units in call centre interactions. In: 2013 IEEE 4th international conference on cognitive infocommunications (CogInfoCom). IEEE, pp 403–406
25. Gross JJ (1998) The emerging field of emotion regulation: an integrative review. Rev Gen Psychol 2(3):271
26. Gross JJ, Thompson RA (2007) Emotion regulation: conceptual foundations. In: Handbook of emotion regulation, vol 3, p 24
27. Harrigan J, Rosenthal R (2008) New handbook of methods in nonverbal behavior research. Oxford University Press
28. Hoffman ML (2008) Empathy and prosocial behavior. Handb Emot 3:440–455
29. Juslin PN, Scherer KR (2005) Vocal expression of affect. In: The new handbook of methods in nonverbal behavior research. pp 65–135
30. Kim S, Georgiou PG, Lee S, Narayanan S (2007) Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In: Proceedings of multimedia signal processing, 2007 (MMSP 2007). pp 48–51
31. Konar A, Chakraborty A (2014) Emotion recognition: a pattern analysis approach. Wiley
32. Kononenko I (1994) Estimating attributes: analysis and extensions of relief. In: Proceedings of machine learning: European conference on machine learning (ECML). Springer, pp 171–182
33. Lee CC, Busso C, Lee S, Narayanan SS (2009) Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In: Proceedings of Interspeech. pp 1983–1986
34. McCall C, Singer T (2013) Empathy and the brain. In: Understanding other minds: Perspectives from developmental social neuroscience. pp 195–214
35. NIST (2009) The 2009 RT-09 RIch transcription meeting recognition evaluation plan. NIST
36. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: Liwc 2001. Lawrence Erlbaum Associates, Mahway, p 71
37. Perry A, Shamay-Tsoory S (2013) Understanding emotional and cognitive empathy: a neuropsychological. In: Understanding other minds: Perspectives from developmental social neuroscience. Oup Oxford, p 178

38. Platt J (1998) Fast training of support vector machines using sequential minimal optimization. MIT Press. http://research.microsoft.com/~jplatt/smo.html
39. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, et al (2011) The kaldi speech recognition toolkit. In: Proceedings of automatic speech recognition and understanding workshop (ASRU). pp 1–4
40. Riccardi G, Hakkani-Tür D (2005) Grounding emotions in human-machine conversational systems. In: Lecture notes in computer science. Springer, pp 144–154
41. Robbins S, Judge TA, Millett B, Boyle M (2013) Organisational behaviour. Pearson Higher Education AU
42. Scherer KR (2000) Psychological models of emotion. Neuropsychol Emot 137(3):137–162
43. Scherer KR (2001) Appraisal considered as a process of multilevel sequential checking. Theory Methods Res Apprais Process Emot 92–120
44. Schuller B, Batliner A (2013) Computational paralinguistics: emotion, affect and personality in speech and language processing. Wiley
45. Schuller B, Steidl S, Batliner A (2009a) The interspeech 2009 emotion challenge. In: Proceedings of Interspeech. pp 312–315
46. Schuller B, Vlasenko B, Eyben F, Rigoll G, Wendemuth A (2009b) Acoustic emotion recognition: a benchmark comparison of performances. In: Proceedings of automatic speech recognition and understanding workshop (ASRU). pp 552–557
47. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2013) Paralinguistics in speech and language state-of-the-art and the challenge. Comput Speech Lang 27(1):4–39
48. Stepanov E, Favre B, Alam F, Chowdhury S, Singla K, Trione J, Béchet F, Riccardi G (2015) Automatic summarization of call-center conversations. In: In Proceedings of the IEEE automatic speech recognition and understanding workshop (ASRU 2015)
49. Tamaddoni Jahromi A, Sepehri MM, Teimourpour B, Choobdar S (2010) Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. J Strateg Mark 18(7):587–598
50. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann