Chapter 20
Conflict Cues in Call Center Interactions

Maria Koutsombogera, Dimitrios Galanis, Maria Teresa Riviello, Nikos Tseres, Sotiris Karabetsos, Anna Esposito, and Harris Papageorgiou

20.1 Introduction

This work explores the multimodal nature of conflicts occurring in call center dyadic interactions from a multidisciplinary perspective, including paralinguistic and conversation analysis cues. It is based on a Greek phone company's corpus which consists of conversations revealing how customers interact with call center operators to express concerns in terms of efficiency, provided services, argumentation and negotiation issues, and expressed emotions during the interactional exchanges. In this setting, we consider conflictual the interaction between speakers that pursue individual and at times incompatible goals (Allwood 2007). Conflict holds between the beliefs and goals of two individuals involved in the conversation, which, namely, represent the consumer and service provider roles. Instantiations of emotional behavior related to conflict and to the speakers' roles are, e.g., the expression of frustration or anger on the customers' side, stress detection, disappointment mitigation, and failure in providing the requested services or solutions on the operators' side.

Understanding conflict cues and being able to model them enables the development of technologies that can deal effectively and efficiently with the complexity of

M. Koutsombogera (✉) • D. Galanis • S. Karabetsos • H. Papageorgiou
Institute for Language and Speech Processing,'Athena' Research Centre, Artemidos 6 & Epidavrou, 15125 Athens, Greece
e-mail: mkouts@ilsp.gr

M.T. Riviello • A. Esposito
Department of Psychology, Second University of Naples, Caserta, Italy

International Institute for Advanced Scientific Studies (IIASS), Vietri sul Mare, Italy

N. Tseres
Department of Linguistics, University of Athens, University Campus, 15784 Athens, Greece

F. D'Errico et al. (eds.), *Conflict and Multimodal Communication*, Computational Social Sciences, DOI 10.1007/978-3-319-14081-0_20

expressions and understandings of human behavior and respond to a growing need of applications related to human behavior analysis (Narayanan and Gregoriou 2013). For example, the telecommunications industry suffers from approximately 30 % of churn rate, while it is of great importance to keep a high percentage of customer retention (Jahromi et al. 2010). In this context, it is crucial to detect emotional traits providing information about the speakers' intentions and emotional states. These traits are multimodal, in the sense that they can be inferred from the paralinguistic properties of the utterances, from the structural units of the interaction and their flow, as well as from the linguistic content. At the same time, the perception of the speakers' emotional states and the definition of the appropriate values describing them are more than a trivial issue. This work focuses on most of the aforementioned aspects, keeping aside for the time being the investigation of the linguistic content.

For the needs of our task, we extracted emotionally colored units from our call center corpus. These units were in turn labeled by a human annotator with the values of positive or negative. A large number of speech and other context-related features were extracted for each unit to train mathematical models that can be used to predict the label of an unseen emotional unit. A subset of the corpus was further annotated in terms of turn management types, and the resulting annotations were then associated to the emotional labels. In the next section, we describe the collected corpus and the procedures applied to annotate it as well as a small-scale experiment aimed to assess the perception of emotions from conversations extracted from this domain. In Sect. 20.3 we describe the automatic feature extraction process and the machine learning models proposed for the automatic classification tasks together with the obtained results. Section 20.4 is dedicated to the turn management annotation process, the study of overlapping speech in the corpus, and the association of turn management values to the emotional ones. Finally Sect. 20.5 concludes the presented work and provides future directions.

20.1.1 Related Work on Automatic Emotion and Conflict Detection

A rich set of combined speech features has been investigated and it is often employed, usually related to the temporal, prosodic, as well as the spectral content of the speech signal, to capture any underlying emotional pattern reflected upon these features (Schuller et al. 2010, 2011; Morrison et al. 2007). Previous works have studied emotion recognition and more specifically anger in speech (Neiberg and Elenius 2008; Lee and Narayanan 2005; Burkhardt et al. 2009; Polzehl et al. 2011; Erden and Arslan 2011). On a task of discriminating five emotions (fear, anger, sadness, neutral, and relief) in real-world audio data from a French human-human call center corpus, an average detection rate of 45 % was reported with only 107 acoustic features (Vidrascu and Devillers 2007). Support vector machines (SVMs) and Gaussian mixture models (GMMs) have shown a reasonable accuracy

in detecting anger in voice-controlled telephone services (Neiberg and Elenius 2008; Burkhardt et al. 2009) (average recall of 83 % and 69 %, respectively). Duration measures seem not to play an important role in emotion detection (Burkhardt et al. 2009). Incorporating linguistics and training of "emotion salient" words seems promising (Lee and Narayanan 2005), with fusion of acoustic and linguistic cues slightly improving overall scores as in Polzehl et al. (2011). Approaches to automatic conflict detection rely on extracting and analyzing nonverbal behavioral informative cues (Kim et al. 2012) and exploit turn organization features (Pesarin et al. 2012).

20.2 Data Collection

Our data collection consists of 135 audio files corresponding to call center human-human dyadic conversations between an operator and a customer. The conversations come from a customer support service of a Greek telecommunications company and are classified according to their content in six major categories: (1) churns, outgoing calls to contractual customers that have requested to cancel their contract with the company (the operators ask the customers the reason for their choice and attempt to change their mind); (2) customers, incoming calls about any issue related to customer service (technical problems, bills, complaints, etc.); (3) telesales, outgoing calls to customers aiming to sell regular phone contracts; (4) upgrade, outgoing calls informing customers about new offers; (5) mobile, outgoing calls aiming to sell mobile phone contracts; and (6) welcome, outgoing calls to welcome the customers to the company's network. Each category consists of unequal number of files.

The overall duration of the corpus is approximately 9.5 h. Each conversation corresponds to a unique customer, while the operator might be the same in more than one conversation. The distribution of audio files per content, duration, and number of speakers of the corpus is shown in Table 20.1.

Table 20.1 Corpus details

			# of speakers	
Categories	# of files	Duration (min)	Operator	Customer
Customers	43	169	18	43
Churns	13	63	6	13
Upgrade	23	91	7	23
Telesales	24	121	12	24
Welcome	17	64	4	17
Mobile	15	52	9	15
Total	135	560	56	135

20.2.1 Data Annotation

The audio files were next annotated with a twofold aim to (a) identify instances of emotional behavior as expressed in the participants' conversation and subsequently (b) assign an emotional label to them. The data annotation was performed by an expert annotator. The selection of the appropriate utterances relied on the annotator's perception of verbal and paralinguistic cues expressing the speakers' sentiments and feelings. Specifically, the annotator's task was to detect units that are emotionally colored, i.e., that deviate from a nonemotional and/or neutral way of speaking in terms of either linguistic expressions or prosodic and paralinguistic properties of speech (such as loudness, intensity, etc.) and carry information about emotions the speakers are actually experiencing. These units may be of varying lengths, i.e., interjections, words, phrases, or utterances. In this respect, the conversational segments that are judged neutral or not emotionally colored by the annotator were left unmarked and unlabeled.

The identified units were in turn annotated as positive or negative. This set of values seems to be appropriate for the goals of the specific task, i.e., to describe the attitude of the speakers toward each other as well as their evaluation on the provided services or on the reported problems. It is important to assess in this domain, on the one hand, whether the customers are eventually satisfied with the services they get or seem to evaluate them negatively and, on the other hand, whether operators express the intent to resolve problematic issues, soothe possible negative effects, or are unable to fulfill the customers' requests.

In this sense, variation or scaling of similar emotions pertaining to either the positive or the negative spectrum is considered to be grouped under one of the two values. For example, no matter if speakers express helplessness, frustration, or anger, the essential part is that they eventually express a negative stance; thus, emotions of the aforementioned values will be labeled as negative.

A second reason for selecting this binary set of values was to avoid ambiguity that is expected to affect the automatic processing phase as well as the evaluation of the data by human judges. Specifically, the higher the number or the granularity of the labels is, the more complex the recognition task becomes, especially when certain labels are semantically close to each other.

Initially, the selected annotation labels consisted of a set of 25 categorical values tailored to the needs of the call center domain and inspired by inventories of categories representing emotions and related states as suggested in the EmotionML (Schröder 2013; Schröder and Pelachaud 2012). In practice, annotating the data with this fine-grained set of labels proved to be a hard task due to the difficulty to assign an appropriate label to speech units showing relatively insufficient perceptual cues in order to disambiguate between labels of semantically similar values. For example, though it was easy to discern between units expressing opposite emotional states, such as satisfaction and anger, there were lots of ambiguous units which were perceptually considered representative of more than one single emotional label (e.g., anger/irritation/frustration).

Furthermore, this binary set of positive and negative values would facilitate the inter-annotator agreement experiments, where different annotators are expected to make faster and more reliable judgments when using binary labels. In addition, this approach will facilitate the evaluators' task when the emotion is not distinctly expressed (as in the case of phone calls, where noise is a constant factor).

The distinct 25 categorical labels initially selected were thus mapped to two values, namely, positive and negative. Table 20.2 shows the diversity of emotion types that positive and negative classes refer to.

The aforementioned process resulted in the annotation of 1,396 speech units. Their distribution according to content categories, label type, gender, and speaker role is depicted in Figs. 20.1 and 20.2 below. The call types where an emotional behavior is detected more frequently are churns and customers. The negative emotional label prevails over the positive one, suggesting that when speakers exhibit an emotional behavior, it is usually targeted to expressions of complaints and dissatisfaction. Furthermore, the majority of emotional units are uttered by customers.

Table 20.2 Coarse categorical emotion labels used and mapping to detailed values

Coarse	Fine grained
Positive	Pleasure, satisfaction, excitement, interest, politeness, certainty, relief, trust, surprise, reassurement
Negative	Anger, annoyance, irritation, disappointment, frustration, anxiety, worry, helplessness, confusion, doubt, uncertainty, irony, indifference, surprise, suspicion

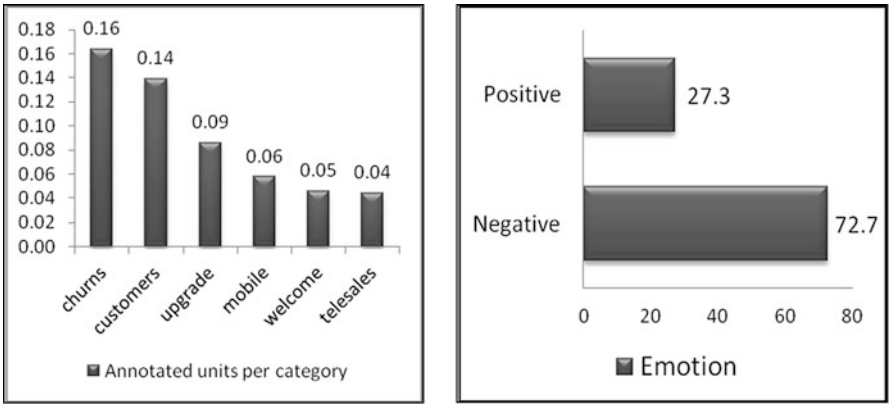

Fig. 20.1 Annotated units per file type—ratio of the total duration of the annotated units over the total duration of the audio files according to the call types (*left*) and the binary emotional labels (*right*)

Fig. 20.2 Distribution of emotionally annotated units according to gender (*left*) and speaker role (*right*)

Table 20.3 Agreement results between annotators within the same language (GReek and ITalian) and between the naive annotators and the expert one

Agreement results			
GR1 vs. GR2	*Expert vs. GR1*	*Expert vs. GR2*	*Average*
80.4 %	89.3 %	79.8 %	83.2 %
IT1 vs. IT2	*Expert vs. IT1*	*Expert vs. IT2*	*Average*
40 %	44.6 %	65.30 %	50 %

20.2.2 Perceptual Experiments and Inter-annotator Agreement

To measure inter-annotator agreement, the extracted and annotated units were assigned to two Greek nonexpert annotators (GR1, GR2) to label them as positive, negative, or neutral. The "neutral" value was given as an option so as not to bias the annotators toward the positive or negative direction. The annotators had no prior knowledge of (a) the context of the discussion and (b) the content of the files the units were extracted from. The two annotators agreed on the 74 % of speech units, showing a high agreement with the expert annotator. The out-of-context assessment was then compared to the expert annotator's assessment, resulting in an agreement average of 83.2 %, cf. first row of Table 20.3.

In parallel, driven by works reporting on psychological experiments carrying out a comparative analysis of subjective perceptions of emotional states (Riviello et al. 2011; Esposito and Riviello 2011), a cross-cultural pilot experiment was carried out to investigate the role of paralinguistic information and language in the perception of emotional information. Specifically, our goal was to test through a small-scale experiment the human ability to infer emotional information through only perceptual cues and how effective it is compared to the knowledge of the linguistic content.

The experiment involved the assessment of the speech units by two Italian annotators (IT1, IT2) who do not have any knowledge neither of the Greek language nor of the context of the discussions. Inter-annotator agreement was measured among the two Italian annotators as well as between each of them and the expert annotator (cf. Table 20.3).

Table 20.4 Inter-annotator agreement between the Greek (left) and the Italian (right) subjects

GR inter-annotator agreement				IT inter-annotator agreement			
	Positive	Negative	Neutral		Positive	Negative	Neutral
Positive	4.2 %	17.6 %	2.7 %	Positive	3.8 %	1.2 %	0
Negative	0.4 %	69.5 %	1.2 %	Negative	9.4 %	36.1 %	0
Neutral	0.1 %	4 %	0.3 %	Neutral	24.2 %	25.2 %	0.1 %

The low agreement score (40 %) between the Italian annotators is mostly because one of them used the "neutral" label frequently, while the other did not, as can be shown in the detailed inter-annotator agreement scores in Table 20.4. The "neutral" label was mainly attributed to units whose paralinguistic properties could not drive the annotator to infer whether those units have a positive or negative value. This also explains why the neutral instances annotated by one of the Italian subjects are equally attributed to a 50 % of positive and 50 % of negative labels (cf. Table 20.4) by the other Italian subject.

This pilot experiment suggests that paralinguistic cues are essential for the perception of emotions in speech as well as that lexical or linguistic information drastically improve the annotation's accuracy. Thus, these preliminary results show that the decoding of positive and/or negative emotion in speech units largely depends on the native language knowledge and the communication context. Native speakers seem to be favored in comparison to the nonnative ones because of their ability to infer linguistic and the semantic contents in addition with the exploitation of prosodic and paralinguistic information. This assumption, however, needs to be verified by further experimentation including more elaborate conditions as well as an adequate number of nonnative subjects.

20.3 Automatic Emotion Classification Experiments

In order to automatically classify the data at hand, the speech units were shuffled and grouped into a training (TR) and a testing (TE) set, respectively, in such a way that the resulting sets refer to disjoint speakers. Also, to avoid bias toward one or another category during the training and the testing phases, the corpus splitting resulted in parts that contain a similar proportion of positive/negative, operator/customer, and male/female speech units (cf. Table 20.5).

The TR set (1,150 units) was used for training two different machine learning algorithms to discriminate between emotionally positive and negative speech units. The TE set (246 units) was used for assessing the algorithms' performance on unseen positive and negative speech examples.

Table 20.5 Corpus annotation statistics—percentage of annotated units per label, speaker role, and gender

	Negative (%)	Positive (%)	Operator (%)	Customer (%)	Female (%)	Male (%)
Train	72	28	38.5	61.5	78.7	21.3
Test	76	24	30.5	69.5	55.3	44.7
Total	72.7	27.3	37.1	62.9	74.6	25.4

20.3.1 Automatic Feature Extraction

Audio Features. To extract audio features, the extended set of speech features as proposed by the Interspeech 2010 Paralinguistic Challenge was exploited (Schuller et al. 2010). The speech features are computed using openSMILE, the audio feature extraction front-end component of the open-source Emotion and Affect Recognition (openEAR) toolkit (Eyben et al. 2010). A total of 1,582 acoustic features were extracted for each speech unit, including mainly descriptive statistical functionals (DSFs) computed over low-level descriptors (LLDs), i.e., speech features derived on a frame-level analysis.

Given the varying lengths of our speech units, the DSF extraction provides static feature vectors for speech units of different sizes, and therefore, their use is beneficial with respect to the nature of the problem (Schuller et al. 2007, 2011). The resulting feature vector acoustically describing each speech unit includes many speech feature types and statistical functionals, to cover prosodic (e.g., loudness, pitch, pitch envelope, etc.), spectral (e.g., energy, log Mel frequency bands, MFCCs, line spectral pairs, etc.), as well as voice quality (e.g., jitter, shimmer, etc.) quantities. Specifically, the 1,582 first level functionals are obtained from 21 DSFs applied over 34 LLDs plus their deltas and 19 DSFs applied over 4 LLDs and their deltas together with two more features regarding pitch onsets and turn durations. The LLDs are computed on a frame rate of 10 ms with a window size of 40 ms (except for MFCCs and LSPs where the window size was set to 25 ms) and then are smoothed with a moving average low-pass filter. The DSFs, among others, include lower-order moments, extremes, percentiles, quartiles, regression coefficients, peaks, etc. Additional details can be found in Schuller et al. (2007; 2011).

Additional Features. For each speech unit, the three following features were also taken into account: (a) the speaker role, i.e., whether a unit is uttered by a customer or an operator, (b) the gender (male/female), and (c) the speech unit's duration that may vary from 0.5 to 12 s.

20.3.2 Classification Experiments

LIBSVM (Chang and Lin 2011), a popular open-source implementation of support vector machines (SVM) learning method, was used for the classification experiments. Given a set of training instances labeled as positive (+1) or negative (−1), the

SVM selected model learns an optimal hyperplane that separates these classes. The learned model can then be used to predict the label (category) to be attributed to an unseen speech unit. In our experiments, the SVM model selected uses a radial basis function (RBF) kernel which, as shown in previous works, is effective in similar emotion classification tasks (Mower et al. 2011).

The SVM model was initially trained considering only the 1,582 features that openSMILE extracted from each unit. We used WEKA's implementation (Hall et al. 2009) of information gain (IG) and Pearson correlation (PC) to select the N most relevant to the task features ($N = 1, 2, 3, 1{,}582$). The feature selection was performed on the whole training set. At a next level, we repeated the same experiments using the additional abovementioned features (role, gender, and duration). To distinguish among the two different sets of features used, the first method using 1,582 features was indicated with SVM-RBF1, and the second one using 1,585 features was named SVM-RBF2. Both methods use PC as a feature selection method, since it performed slightly better than IG. A majority classifier was used as baseline, i.e., a classifier that assigns the label that dominates in the training set, the negative in our case, to all instances of the test set.

20.3.3 Results

The obtained classification results from the SVM-RBF1 and SVM-RBF2 procedure are illustrated in cf. Fig. 20.3. The x-axis indicates the number of the features

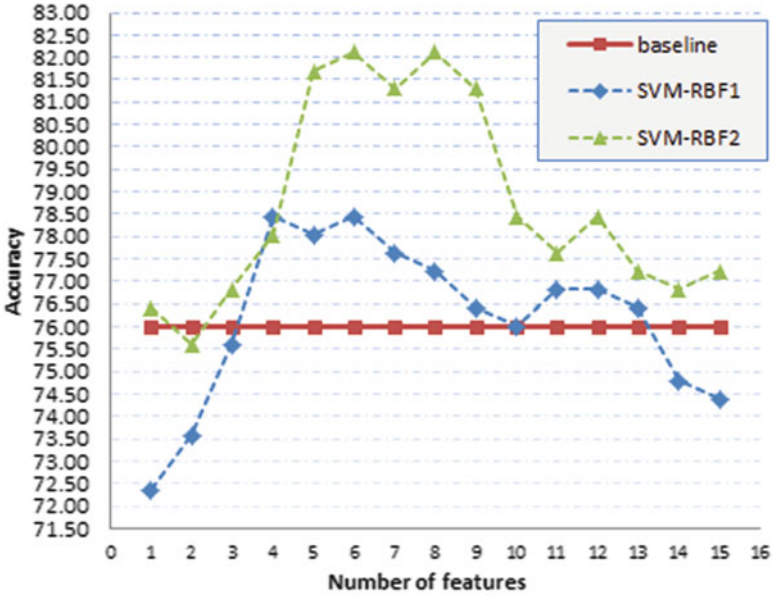

Fig. 20.3 Accuracy results obtained with different feature sets

Table 20.6 Precision and recall scores for SVM-RBF2 ($N = 6$)

	Negative (%)	Positive (%)
Recall	97.33	33.90
Precision	82.35	80.00

Table 20.7 Best features

Ten best features (PC measure)	
1	Speaker role
2	logMelFreqBand_sma[6]_amean
3	logMelFreqBand_sma[7]_amean
4	lspFreq_sma[1]_quartile1
5	logMelFreqBand_sma[1]_percentile99.0
6	logMelFreqBand_sma[6]_quartile1
7	logMelFreqBand_sma[6]_quartile2
8	logMelFreqBand_sma[7]_quartile1
9	logMelFreqBand_sma[7]_quartile2
10	logMelFreqBand_sma[5]_amean

used, i.e., the N best according to PC, and the y-axis shows the accuracy, i.e., the proportion of the correctly classified instances from the TE. More specifically, the baseline method achieves a 76.01 % accuracy, while SVM-RBF1 and SVM-RBF2 achieve a maximum accuracy of 78.45 % (for $N = 6$) and 82.11 % (for $N = 6$), respectively. In general, SVM-RBF2 has a higher accuracy for almost all values of N. This is most probably due to the fact that the speaker role (customer/operator) feature has the highest PC value.

The effectiveness of the best classifier SVM-RBF2 ($N = 6$) was further analyzed using precision and recall measures for the negative and positive labels. In particular, as shown in Table 20.6, the classifier achieves high precision and recall scores for the negative label. On the other hand, the recall scores for the positive label are much lower than the negative one.

Table 20.7 includes a ranked list of the ten best features according to PC measure.

A further qualitative analysis of the erroneously classified speech units shows that the majority of them are attributed to female (79 %) rather than male (21 %) speakers in accord with the different gender distribution in the corpus. In addition, operators' speech units are slightly worse classified (53 %) than customers' (47 %) ones. This is largely due to the nature of the conversations: for example, operators, in repeated attempts to reassure customers that their problems are being dealt with and a solution is underway or in explaining a misunderstanding with regard to a certain procedure, often speak loudly and in a severe manner. This may produce a clash between their positive intents, as expressed in the verbal content of the utterance and the paralinguistic properties of their speech that leads to a classification of these cases as negative instead of positive. The aforementioned cases account for the 37 % of the errors. On the other hand, there are cases of customers' negative instances, expressed, however, in a calm and quiet manner; hence, these instances are erroneously classified as positive.

20.4 Turn-Taking Structure and Conflict

Exploring the turn-taking structure in the conversation is closely linked to the investigation of conflict, since conflicting exchanges may be traced in the structure of the floor management. While compliance with the turn-taking rules guarantees a successful flow of conversation, irregularities may indicate that there is a tendency for disagreement and intensity, negotiation of opinions, and social situations and thus may be associated with conflict. In an attempt to identify turn-taking cues or patterns related to conflict, we added an extra data annotation level to enrich part of the corpus with this information.

Conversations from the churn category were manually annotated with turn management labels. Specifically, all turn transition points in the audio files were marked on the time axis and were assigned with a label describing the type of turn transition. The ELAN[1] editor was used with a set of labels mainly inspired by the MUMIN schema (Allwood et al. 2007) and with the addition of the backchannel value as shown in Table 20.8 below.

Churn files were selected as a representative content category with regard to conflict, since in this type of conversations, the participants pursue their own individual and conflicting goals. The customers have acknowledged issues which lead them to the decision of quitting their contract, while the operators attempt to change their mind after inspecting their problems and suggesting solutions. A typical attested behavior is customers complaining on noneffective services and/or company's unfair behavior. The clients generally express distrust in the company's services and the feeling of not being adequately protected in their customers' rights. On the other hand, the operators try to soothe the customers' negative feelings performing a series of planned actions devoted to resolve the inconsistencies among customers' wishes and actual services provided by the phone company. We thus believe that the churn category would be of particular interest to further study its turn-taking organization.

Table 20.8 Turn management annotation labels

Turn management labels	
Turn take	A speaker initiates the turn by introducing a new topic
Turn grab	A speaker takes the turn without being offered to do so, possibly by interrupting
Turn accept	A speaker accepts a turn that is offered
Turn offer	A speaker offers the turn to another speaker
Turn yield	A speaker yields the turn being under pressure or interrupted
Turn complete	A speaker completes a turn
Turn hold	A speaker attempts to keep the turn
Backchannel	A speaker produces acknowledgments and backchannels

[1] http://tla.mpi.nl/tools/tla-tools/elan/

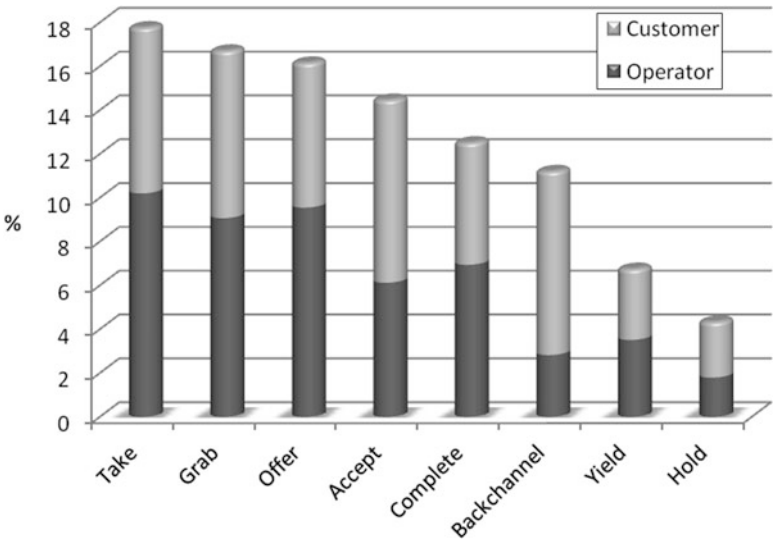

Fig. 20.4 Percentage distribution of turn management labels assigned to customers and operators

In 63 min of 13 churn files, 1,455 turn management segments were annotated with the respective set of labels and their distribution is depicted in Fig. 20.4. The second most frequent label, *turn grab*, indicates that there is a large number of interruptions performed by both participants and seems to function like an effective cue in tracing conflict. The distribution of labels per speaker role, i.e., customer and operator, may vary, as shown in Fig. 20.4. Most of the differences in speaker roles are explained by the label semantics, i.e., it is expected that the operators perform more turn offers by, i.e., asking customers questions, and that customers accept the turn more frequently than the operators. There are nevertheless interesting differences, especially regarding the *turn grab* and *turn yield* labels, the distribution of which implies that both speakers are engaged on a more or less equal basis in interruption instances and thus in conflictual situations.

Another conversational feature related to conflict according to the literature (Sacks et al. 1974; Schegloff 2000; Schuller et al. 2013) is that of overlapping speech, which is considered as a "violation" of the social rule that one party should speak at a time and therefore may be informative of speakers' interrupting attempts to grab the floor. Recent work has outlined the role of overlaps as a reliable cue accounting for the presence of conflict and a sign of competition for having the floor, focusing on their frequency and duration (Grezes et al. 2013).

In this respect, we calculated the overlaps between turn management labels by directly exploiting the annotations, i.e., by extracting all instances where there is an overlap between the conversational actions of the two speakers in a turn transition point, as it is, for example, in the case where a speaker grabs the turn and at the same time the other speaker yields the turn. Overlapping labels account for 31.6 % of the

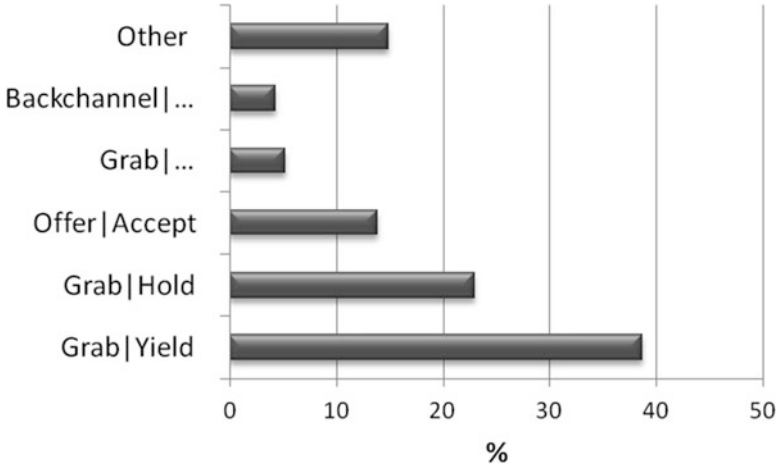

Fig. 20.5 Percentage of overlapping cases between turn management labels

annotations in the turn management layer. Annotations of turn management that do not overlap are, in their majority, cases of turn initiation, greetings, and starting or ending the discussion. In the latter cases, no overlapping speech is observed.

Most of the overlaps (67 %) occur with the *turn grab* label, which is meant to indicate interruptions. Specifically, *turn grab* overlaps mostly with *turn yield*, followed by *turn hold*, as well as with some other labels, as shown in Fig. 20.5 below. It is attested that not all instances of simultaneous talk constitute interruptions; there might also be collaborative or delayed completions, overlaps at transition relevant places (c.f. the offer-accept pair of values), signaling feedback, or backchanneling. However, the high frequency of the *turn grab* labels (paired with other labels) is an indicator of pure interruptions that express the intention to claim the turn and implies conflict between the goals and beliefs of the two speakers.

To better account for conflict cues emerging from the corpus, the turn-taking structure was associated with the expressed emotions exploring the distribution of emotionally annotated units overlapping with a turn management label at any point in the units' lengths, i.e., either at the beginning, middle, or the end of a speech unit. This was done by extracting from the annotated corpus speech segments attributed both to a turn management and an emotion label. During a given interval, it was often attested that an emotional unit overlapped with two turn management units, when simultaneous speech occurs. For example, a negative emotion is expressed by a speaker during simultaneous speech, where two turn management labels are annotated, e.g., *turn grab* by one speaker and *turn yield* by the other. In this case, two overlaps are measured for the emotional unit perspective (one for grab and one for yield, respectively) and one overlap from the turn management perspective. The results show that 83 % of the emotionally labeled speech units overlap with at least one (but also more than one) turn management label and 35 % of turn management labels overlap with emotional labels because of the high number of

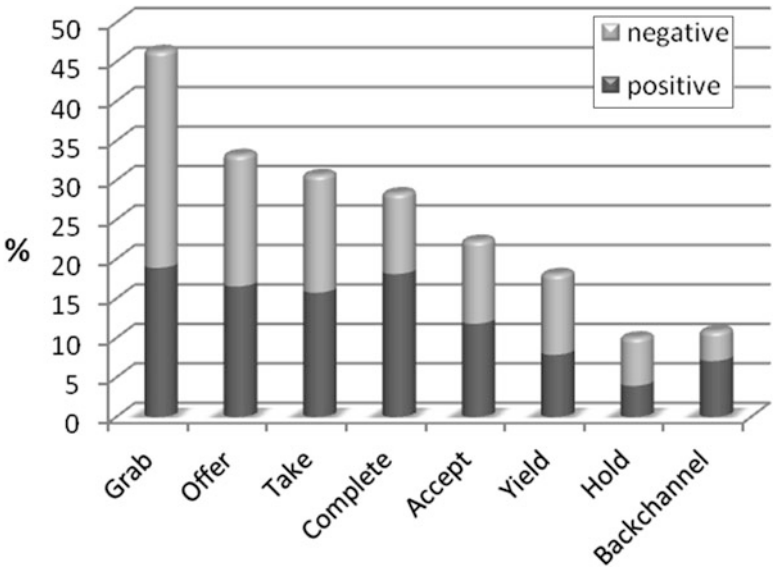

Fig. 20.6 Percentage distribution of turn management labels overlapping with an emotion label

turn managements with respect to the emotional speech units. Turn management labels that are not associated to emotions are mostly turn transitions (e.g., new topic introduction, greetings in the end of the conversation).

As far as the distribution of the labels is concerned (cf. Fig. 20.6), the findings are interesting in that labels that are related to interruption points and that may indicate conflict, such as *grab* or *yield*, overlap more with negative than with positive emotions. A t-test showed the overlaps of negative labels with turn management are statistically significant for *turn grab* with $p = 0.008$, *turn yield* with $p = 0.030$, and *turn hold* labels with $p = 0.031$. In particular, the significance of *turn hold* labels indicates that the speaker who is being interrupted attempts to hold the turn signaling a conversational cue related to conflict.

On the other hand, turn management labels related to the normal conversational flow and regular turn exchanges, i.e., *accept* and *complete*, overlap more with positive emotions. A t-test showed that, in this case, the positive emotional labels' overlaps are statistically significant for turn accept ($p = 0.003$) and turn complete ($p = 0.014$).

The reported results provide some evidence with regard to the relation between turn management labels (representing the rules and their aberrations in terms of how the exchange of turns is performed) and emotional states expressed by the speakers, indicating the whereabouts of potential conflict points within the conversational structure. Before generalizing though, these findings need to be further investigated and compared to other audio files categories, such as the welcome calls, where the distribution and the correlation between emotion and turn management labels remain to be explored.

20.5 Conclusions and Future Work

In this work, we approached the notion of conflict occurring in call center interactions as a complex problem which is decomposed in subtasks related to (a) perceiving and perceptually decoding the emotions occurring in such interactions, (b) automatically classifying them, and (c) exploring the turn-taking structure to find cues and patterns related to conflict.

With regard to the perceptual decoding of vocal emotional expressions, the high agreement scores between Greek annotators indicate the existence of salient perceptual cues allowing to adequately perceive the emotional trace of an utterance, independently of the context. The small-scale perceptual experiment involving native and nonnative raters showed that familiarity with the linguistic content largely improves the assessment of emotion in positive and negative classes. Moreover, an SVM-based algorithm that classifies emotional units extracted from the conversations as positive or negative was presented, the best version of which (SVM-RBF2 $N = 6$) obtained an accuracy score that is 6 % higher than a majority classifier. Finally, in a subset of our corpus (churns), we measured the distribution of turn-taking types, and we explored the association of overlapping speech as well as of overlapping values in turn-taking and emotion to cue the presence of conflict. In this case, it was found that overlapping speech occurs mostly with conflict-related labels in turn-taking (e.g., grab, yield) and that these labels are more correlated with negative emotions.

In future, we plan to improve the classification task by using speaker diarization and speech segmentation techniques to automatically segment recordings and come up with conversation units of variable duration that includes additional features coming from the turn-taking structure and overlapping speech points. Moreover, to allow for generalizations, a cross-lingual study on the human ability to decode perceptually emotional vocal expressions derived from call center dyadic interactions is foreseen, involving more subjects and distinct experimental conditions. Future work will also investigate various ways of incorporating and modeling the temporal sequence and transitions of emotional states, both within the same speaker, and between the two speakers, to show conflict escalation and de-escalation, and discourse structures to improve the automatic classification of conflictual conversations from a business perspective.

Acknowledgments The research leading to these results has been partially funded by POLYTROPON project (KRIPIS-GSRT, MIS: 448306). Also, the participation to Dagstuhl Seminar 13451 "Computational Audio Analysis" held from Nov 3 to 8, 2013, in Wadern, Germany, inspired Anna Esposito to contribute to this work.

References

Allwood J (2007) Cooperation, competition, conflict and communication. Gothenbg Pap Theor Linguist 94:1–14

Allwood J, Cerrato L, Jokinen K, Navarretta C, Paggio P (2007) The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Multimodal corpora for modeling human multimodal behaviour. J Lang Resour Eval 41(3–4):273–287

Burkhardt F, Polzehl T, Stegmann J, Metze F, Huber R (2009) Detecting real life anger. In: ICASSP 2009, Taipei, Taiwan, 19–24 Apr

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3). http://www.csie.ntu.edu.tw/∼cjlin/libsvm

Erden M, Arslan LM (2011) Automatic detection of anger in human-human call center dialogs. In: Interspeech 2011, Florence, Italy, 28–31 Aug

Esposito A, Riviello MT (2011) The cross-modal and cross-cultural processing of affective information. In: Apolloni B et al (eds) Proceedings of the 2011 Conference on Neural Nets WIRN10: Proceedings of the 20th Italian Workshop on Neural Nets. IOS Press Amsterdam, The Netherlands, pp 301–310

Eyben F, Wollmer M, Schuller B (2010) openSMILE—the Munich versatile and fast open-source audio feature extractor. In: ACM multimedia, Florence, Italy, pp 1459–1462

Grezes F, Richards J, Rosenber A (2013) Let me finish: automatic conflict detection using speaker overlap. In: Interspeech 2013, ISCA, Lyon, France

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11:10–18

Jahromi AT, 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

Kim S, Filippone M, Valente F, Vinciarelli A (2012) Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and Gaussian processes. In: ACM international conference on multimedia, Nara, Japan, pp 793–796

Lee CM, Narayanan S (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2):293–303

Morrison D, Wang R, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Comm 49(2):98–112

Mower E, Mataric MJ, Narayanan S (2011) A framework for automatic human emotion classification using emotion profiles. IEEE Trans Audio Speech Lang Process 19(5):1057–1070

Narayanan S, Gregoriou P (2013) Behavioral signal processing: deriving human behavioral informatics from speech and language. Proc IEEE 101(5):1203–1233

Neiberg D, Elenius K (2008) Automatic recognition of anger in spontaneous speech. In: Interspeech 2008, Brisbane, Australia, 22–26 Sept, pp 2755–2758

Pesarin A, Cristani M, Murino V, Vinciarelli A (2012) Conversation analysis at work: detection of conflict in competitive discussions through semi-automatic turn-organization analysis. Cogn Process 13(Suppl 2):533–540

Polzehl T, Schmitt A, Metze F, Wagner M (2011) Anger recognition in speech using acoustic and linguistic cues. Speech Comm 53(9–10):1198–1209, Special Issue on Sensing Emotion and Affect—Facing Realism in Speech Processing

Riviello MT, Chetouani M, Cohen D, Esposito A (2011) Inferring emotional information from vocal and visual cues: a cross-cultural comparison. In: IEEE 2nd international conference on cognitive computation, Budapest, pp 1–4

Sacks H, Schegloff E, Jefferson G (1974) A simplest systematics for the organization of turn-taking in conversation. Language 50:696–735

Schegloff E (2000) Overlapping talk and the organisation of turn-taking for conversation. Lang Soc 29(1):1–63

Schröder M (ed) (2013) Emotion markup language (EmotionML) 1.0, W3C Proposed Recommendation 16 Apr 2013. http://www.w3.org/TR/emotionml/

Schröder M, Pelachaud C (ed) (2012) W3C vocabularies for EmotionML, W3C Working Group Note 10 May 2012. http://www.w3.org/TR/emotion-voc/

Schuller B et al (2007) The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Interspeech 2007, Antwerp, pp 2253–2256

Schuller B et al (2010) The INTERSPEECH 2010 paralinguistic challenge—age, gender, and affect. In: Proceedings of 11th international conference on spoken language processing, interspeech 2010—ICSLP, Makuhari, Japan, 26–30 Sept, pp 2794–2797

Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Comm 53(9/10):1062–1087, Special Issue on Sensing Emotion and Affect—Facing Realism in Speech Processing

Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E, Salamin E, Polychroniou A, Valente F, Kim S (2013) The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Interspeech 2013, ISCA, Lyon, France

Vidrascu L, Devillers L (2007) Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features. In: International workshop on Paralinguistic Speech—between models and data, ParaLing, pp 11–16