Chapter 18
Detecting Speech Interruptions for Automatic Conflict Detection

Marie-José Caraty and Claude Montacié

18.1 Introduction

Research in organization and management has investigated phenomena, such as the causes, effects, and handling of interpersonal or intergroup conflicts (Korabik et al. 1993; Macintosh and Stevens 2008; Thomas et al. 2008). Data on these social and psychological phenomena are collected from people who are involved in the conflict, witnesses of the conflict, or, by extension, looking at a recording of the conflict escalation between the protagonists. A large quantity of audio and/or video metadata can be extracted from these recordings, such as the conversation, face, and gesture interactions. In this chapter, the conversational interactions during political debates have been studied to develop an automatic conflict detector from voice analysis. A reliable detector of conflict would be useful for many applications, such as security in public places, the quality of customer services, and the deployment of intelligent agents. The development of such a system requires modeling of the conversational interactions as well as the search for specific interactions in relation to a given measure of conflict handling (Rahim 1983; Daly et al. 2010).

M.-J. Caraty (✉)
Paris Descartes University, 45 rue des Saints Pères, 75006 Paris, France

STIH, Paris-Sorbonne University, 28 rue Serpente, 75006 Paris, France
e-mail: Marie-Jose.Caraty@ParisDescartes.fr

C. Montacié
STIH, Paris-Sorbonne University, 28 rue Serpente, 75006 Paris, France

F. D'Errico et al. (eds.), *Conflict and Multimodal Communication*,
Computational Social Sciences, DOI 10.1007/978-3-319-14081-0_18

18.1.1 Model of Conversational Interaction

Conversation is a social interaction between two or more people, where taking turns to talk is naturally observed. In the pioneering work of Sacks et al. (1974), an organizational model of turn-taking for conversation that is context-free, capable of context sensitivity, and having a cross-cultural validity was investigated. The constraints of their model were set in reference to the high cross-cultural flexibility of conversation accommodation, with a wide range of interaction in which there is a variety of persons and numbers of persons who are taking part. The authors proposed a model that relies on two components that are related to the turn-constructional units (TCUs, the basic units of talk) and the turn allocation at the end of each TCU for the next unit (the next speaker's TCU). TCUs end with points of possible completion (e.g., gap, query) called transition-relevant places (TRPs), in which the turn transition could be relevant but is not necessary. Observed in any conversation, 14 facts were listed. An excerpt of this list is the following: (a) mostly one party talks at a time; (b) the vast majority of turn-taking transitions is composed of transitions that have no/slight gap and no/slight overlap; (c) the turn size varies; (d) overlapping speech is common, but brief; (e) two basic turn-allocation techniques are used: the "current selects next" technique when a current speaker can select a new speaker (e.g., addressing a question) and the "self-select" technique when a speaker can self-select in starting to talk; (f) repair mechanisms exist for addressing turn-taking violation; e.g., when overlapping speech occurs, one (or more) of the speakers will stop prematurely. A set of rules was edited for addressing turn transitions from TRP in such a way as to minimize the gap or overlap in the transitions. The turn transfer is defined according to the construction of the TCU, regardless of whether the "current speaker selects next" technique is used as well as the eventual application of "self-selection." The rules are based on the purpose of no-gap-no-overlap transitions, for which ability is required in anticipating the precise moment at which a TCU is going to come to a completion point (i.e., a TRP). In related work (De Ruiter et al. 2006), the lexical and syntactic content of TCU was shown to be necessary for this anticipation, while the intonation contour was neither necessary nor sufficient for this projection. According to the turn-taking rule-set applied to a multiparty conversation, overlap is expected in the neighboring transition-relevant places: when a possible completion of the current TCU is wrongly projected by a party or when parties are competing in a self-selection mode for a next turn. In a work that is related to turn-taking organization and that is beyond the ordinary conversation and is mostly unconstrained in terms of a role, a wide range of publications have studied the turn-taking practices and characteristics within various contexts of multiparty interactions. Distinctive features of turn-taking were found in institutional interactions in which a turn-taking organization is more constrained and specialized according to the roles that are assigned to the group members (e.g., interviewer vs. interviewee, chair vs. participant). Studies on turn-taking management were investigated in institutional settings such as in a classroom (Mac Houl 1978; Mehan 1985; Lerner 1995), in courts (Atkinson and Drew 1979), in political interviews (Beattie 1982), in press conferences (Schegloff 1987), in mediation (Garcia 1991), in professional meetings (Boden 1994), in talk shows in which interpersonal conflicts are expressed (Brinson and Winn 1997), in auctions (Heath and Luff 2007), in political debates (Valente and Vinciarelli 2010), and in political meetings that involve large groups of people in which everyone can contribute ideas, opinions, and proposals and in which opposition is also expressed (Mondada 2013). The role of the chair has been analyzed in various studies (Boden 1994; Svennevig 2008; Mondada 2012). Prediction of the speaker order in turn-taking was investigated in news, talk shows, and meetings (Barzilay et al. 2000; Vinciarelli 2009).

18.1.2 Guidelines and Overview

In related work on conflict detection in conversational interactions (Valente and Vinciarelli 2010; Pesarin et al. 2012), turn-taking patterns and overlaps between speakers are shown to be informative with respect to classification into the presence or absence of conflict. The total amount of overlap and the minimum pitch during overlap were found to be the features that correlated the most with conflict (Kim et al. 2012c). A widely adopted classification of interruptions/overlaps is collaborative or competitive in reference to the "cooperative-competitive" dimension of the conflict-handling style. While communication strategies are naturally collaborative, this preponderance is not the case for conflict dialogues, in which competitive strategies are the norm. The detection of competitive interruption is a difficult problem in relation to the search of the TRPs. Spectral content and intonation contour are not sufficient to locate these places. Furthermore, the perception of the conflict can be different in the case of the constrained organization of turn-taking, such as institutional interactions (interview, debate, meeting). Competitive strategies such as those of the moderator or the chairman appear to be natural in this context and are not perceived as conflicting. Our experiments relate to the classification of audio clips into two classes of conflict level (low and high) during the Interspeech 2013 Conflict Challenge. The clips, which were extracted from political debates, have been annotated into conflict levels, using crowdsourcing to model the perception of the people. For our design of the conflict detector, we categorized the overlapping speech into low- and high-level conflict overlap. We made the assumption that these categories can be detected from acoustic cues. We focus our study on a multi-resolution framework for the detection of the overlaps and a multi-expert architecture to include knowledge about overlap in the automatic conflict detector.

This chapter is organized as follows: Section 18.2 presents the speech material that we used for the experiments on conflict detection; it describes and analyzes the statistical characteristics of the corpus while focusing on interruptions and the moderator's role. Section 18.3 describes the Conflict Challenge and the various audio feature sets that were used for our investigations. In Sect. 18.4, the multi-resolution framework of the overlap detectors is outlined, the relation of the types of overlap with the conflict level is introduced and assessed on the Development set, and the results are discussed according to the official measure of the challenge in terms of the UAR. Section 18.5 describes the multi-expert architecture of the conflict detector. Various audio features that are related to the overlap detectors are presented. The results on the conflict detector task on the Test set are discussed. Section 18.6 presents the study's conclusions.

18.2 Speech Material

The SSPNet corpus (Kim et al. 2012a) is an international reference for social signal databases. In the context of political debates, this corpus allows investigations on conflict to occur during interactions between group members. SSPNet was used for our study in analyzing various turn-taking characteristics and testing models for conflict level detection.

18.2.1 SSPNet Corpus

The "SSPNet Conflict Corpus" is a collection of 45 political debates in the French language that were televised in Switzerland. It represents approximately 12 h of speech signals; 1,430 audio clips of 30 s duration were extracted from the corpus. A total of 157 individuals were speaking in the collection of debates (23 females and 134 males). In the various multiparty discussions of the debates, the roles of the group members were distinguished: a member of the group held the role of moderator, and the other members were participants who were taking part in the debate. Four moderators (1 female, 3 males) and 153 participants (22 females, 131 males) were counted in the database. The SSPNET corpus was distributed for the Interspeech 2013 ComParE Challenge. Data were split into the Train, Development, and Test sets: 793 clips were in the Train set, 240 clips were in the Development set, and 397 were in the Test set. Metadata are available for the Train and Development sets.

The clips were annotated in terms of the conflict score in the range -10 to $+10$ by crowdsourcing, to model the perceptions of the data consumers at a nonverbal level; metadata were taken to be low-level conflict (LLC) when the score was lower than 0; otherwise, it was taken to be high-level conflict (HLC). Figure 18.1 shows the distribution of the clips of the Train set as a function of the conflict score range (CSR). The clips are split into the two classes of level conflict (LLC and HLC); the dashed line shows the boundary between the LLC and HLC clips. LLC clips are predominantly represented in the database (63 % for LLC vs. 37 % for HLC).

Segmentation metadata are available for each clip, indicating the diarization ("who spoke when"). From these metadata, we can compute the following statistics:

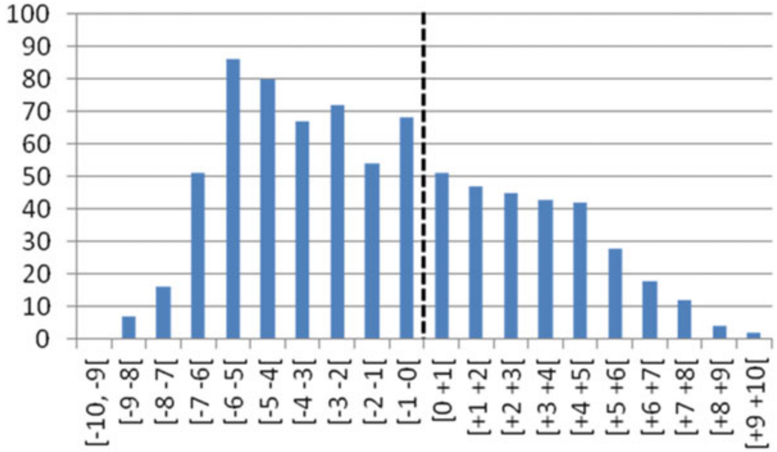


Fig. 18.1 Clip occurrence on the Train set as a function of the CSR

(a) the overlap segment duration, (b) the clip overlap duration as the summation of each overlap segment duration of the clip, (c) the mean overlap duration of a clip as the ratio of the clip overlap duration to the number of overlaps occurring in the clip, and (d) the percentage of overlap duration of the clip as the ratio of the clip overlap duration to the clip duration.

18.2.2 SSPNet Train Set Statistics

We analyzed the statistics of the SSPNet database Train set in focusing on the main characteristics of overlap segments; some statistics of the moderator were also investigated. The Train set includes 793 clips and has a total duration of 23,774 s (two clips' duration is inferior to 30 s), with 82 speakers (one moderator and 81 participants).

We analyzed the 4,143 segments of 23,774 s duration that were obtained by the clip diarization given in the SSPNet database. These segments were split according to the number of speakers that occurred in the segment: (1) 34 segments of a total duration of 89.9 s, which correspond to gaps in which nobody is speaking, (2) 2,638 segments of a total duration of 20,083.5 s, in which a lonely subject is speaking, and (3) 1,471 segments of a total duration of 3,600.6 s, in which two subjects are speaking. No segment was identified that had three or more speakers.

Figure 18.2 shows the histogram for each CSR of the average of the number of interruptions (i.e., the segments of overlapping speech) of the CSR clips. The horizontal dashed line represents the average of the number of interruptions of the Train set clips. Except for the CSR ([−1, 0[), all of the CSRs of LLC have a mean number of interruptions that are below the average value (1.85 = 1,471/793). The HLC clips have more interruptions than the LLC clips.

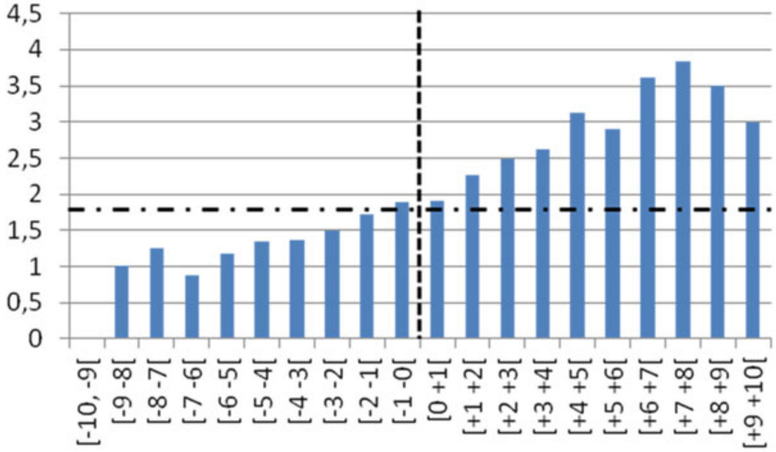


Fig. 18.2 Average of the number of interruptions as a function of the CSR

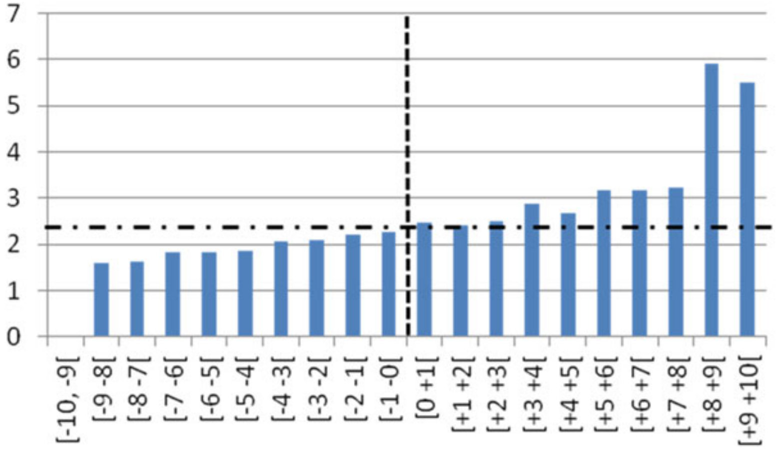


Fig. 18.3 Overlap mean duration (in s) as a function of the CSR

Figure 18.3 shows the histogram of the overlap mean duration for each CSR. The horizontal dashed line represents the average of the overlap duration in the Train set (2.45 s = 3,600.6/1,471). HLC clips have a mean duration of overlap that is higher than the LLC clips.

Figure 18.4 shows the histogram for each CSR of the percentage of overlap duration. The horizontal dashed line represents the mean percentage of the overlap duration of the Train set clips (15.1 % = 3,600.6/23,774). The conflict level is shown to be highly correlated to the percentage of overlap duration as in related work (Kim et al. 2012b).

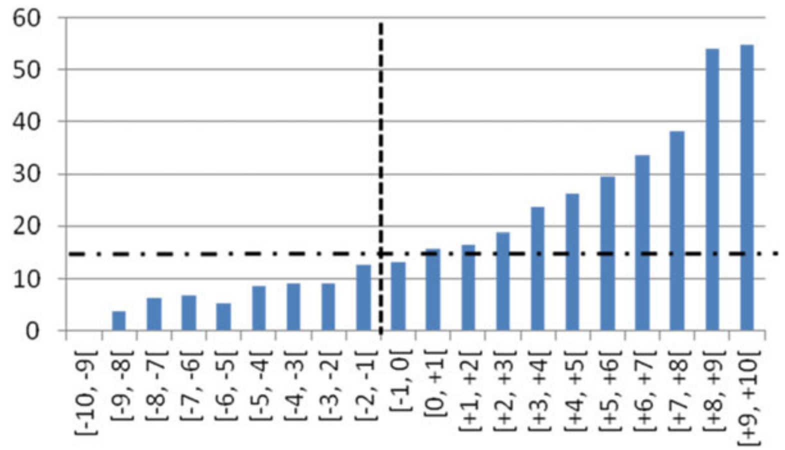


Fig. 18.4 Percentage of overlap duration as a function of the CSR

Table 18.1 Statistics on the speech duration of the moderator

Moderator—spk-050	Train set	Non-Ov	Ov	LLC-Ov	HLC-Ov
Total speech duration (in s)	27,284.7	20,083.5	7,201.2	2,619.0	4,582.2
Speech duration of the moderator (in s)	5,149.7	3,183.7	1,966	1,029.8	936.2
Percentage of the moderator speech duration	18.9	15.9	27.3	39.3	20.4

In the multiparty discussions of the debates, a member of the group held the role of moderator among the participant members who were taking part in the debate. We analyzed various statistics that were related to the moderator from the Train set.

In Table 18.1, the statistics on the speech duration of the moderator are given. The total speech duration (27,284.7 s) that we accounted for is different from the total segment duration (23,774 s) of the Train set; it was estimated as the duration of a segment in which a lonely subject is speaking plus twice the segment duration in which two subjects are speaking. The speech duration of the moderator was computed for the various classes of speech: Nonoverlap (Non-Ov) and overlap (Ov) were split into the two conflict level classes, low-level conflict overlap (LLC-Ov) and high-level conflict overlap (HLC-Ov). The speech duration of the moderator is given (in s). The percentage of the moderator speech duration was computed as the ratio between the speech duration of the moderator and the total speech duration. From the previous statistics, we note in Table 18.1 that the moderator speaks more during LLC-Ovs than during HLC-Ovs (39.3 % vs. 20.4 %) and Non-Ovs (39.3 % vs. 15.9 %).

In Table 18.2, the modes of the interruptions that were related to the moderator were analyzed according to the following occurrences: "the moderator interrupted a participant" or "the moderator was interrupted by a participant." Clips were extracted from the video into 30-s duration segments. The mode of interruption that was related to the moderator was defined from an overlap in which the moderator

Table 18.2 Statistics on interruption mode occurrences of the moderator

Moderator—spk-050	Ov	LLC-Ov	HLC-Ov
# of interruptions	1,353	604	749
# of interruptions by the moderator and occurrence percentage	645 (47.7 %)	357 (59.1 %)	288 (38.4 %)
# of interruptions of the moderator and occurrence percentage	198 (14.6 %)	127 (21.0 %)	71 (9.4 %)

was speaking, by examining the previous segment: if in this segment the moderator was speaking, then the moderator was interrupted by a participant; otherwise, the moderator interrupted a participant. Taking off the first segment of each clip, an interruption occurs at the beginning of each overlap segment; the total number of interruptions in the Train set is 1,353 split into 604 interruptions in LLC-Ovs and 749 interruptions in HLC-Ovs. The number of interruptions by the moderator (respectively, the interruptions of the moderator) was computed for the overlaps and their two categories (LLC and HLC) as well as its percentage of occurrence. We note that the moderator interrupted the participants more often than the moderator was interrupted by the participants (47.7 % vs. 14.6 %). Moreover, the moderator interrupted the participants more in the LLC-Ovs than in the HLC-Ovs (59.1 % vs. 38.4 %).

18.3 Conflict Challenge

The Conflict Challenge was one of the shared tasks that was organized during the Interspeech 2013 Computational Paralinguistics Challenge (Schuller et al. 2013), which took place from January 15 to May 24, 2013. The task consisted of an automatic analysis of the group discussions, to retrieve the conflicts. The goal of this competition was to bridge the gap between research in automatic conflict detection and the low compatibility of the results. The task data were split into the Train, Development, and Test sets. The speaker dependence between these sets was reduced to a minimum that was needed in the real-life settings. As usual, the criterion to guide the detection strategy is the maximization of the UAR on the Development set. This set is also used to tune the parameters of the learning algorithms. Metadata are available only for the Train and Development sets. The participants did not have access to the labels of the Test set. However, each participant could upload the instance predictions up to five times, to receive the confusion matrix and the results from the Test set. The official measure of the competition is the UAR. An official system of conflict detection was also provided with the following characteristics: the WEKA data mining tool kit was used as a framework for the classification task (Hall et al. 2009), and the support vector machine (SVM) classifier with linear kernel and sequential minimal optimization (SMO) was used for learning; the official set of features (6,373 features), which is referred to as the IS-2013 set, was a representation of the utterances, and the complexity parameter of the SVM classifier was optimized by using UAR on the Development set.

18.3.1 Audio Feature Sets

In this section, we describe the audio feature sets that we used for analyzing speech segments. This speech representation (Vogt and André 2005; Schuller et al. 2008) is a new paradigm for speech analysis. It contrasts with the standard paradigm for speech analysis (the sequence of observation vectors): regardless of its duration, a speech utterance is represented by a large set of features, which is termed an audio feature set. The feature set is based on several low-level descriptors (LLDs) that are computed from short overlapping windows of the audio signal. These LLDs comprise the loudness, the harmonics-to-noise ratio, the zero-crossing rate, the spectral and prosodic coefficients, the formant positions and bandwidths, the duration of voiced/unvoiced speech segments, and features derived from the long-term average spectrum such as band energies, roll-off, and centroid as well as voice quality features such as jitter and shimmer. Various global statistical functions (functionals) are computed on these LLDs to obtain feature vectors of equal size for each speech utterance. The sequence of LLDs that are associated with speech utterances can have different lengths, depending on the duration; the use of functionals allows us to obtain one feature vector per speech utterance, with a constant number of elements. It avoids the use of the expensive procedures of time warping between sequences of different lengths such as dynamic programming algorithms. Some functionals aim at estimating the spatial variability (e.g., mean, standard deviation, quartiles 1–3), and others aim at the temporal variability (e.g., peaks, linear regression slope). The four audio feature sets that we used for our experiments include the set of features that was provided by the organizers of the Interspeech 2010 (IS-2010) Paralinguistic Challenge (Schuller et al. 2010), the set of features for the Interspeech 2011 (IS-2011) Speaker State Challenge (Schuller et al. 2011), the set of features for the Interspeech 2012 (IS-2012) Speaker Trait Challenge (Schuller et al. 2012), and the set of features for the Interspeech 2013 (IS-2013) Conflict Sub-Challenge (Schuller et al. 2013). All of the features were extracted using the open source openSMILE feature extraction tools (Eyben et al. 2010). The IS-2010 feature set consists of 1,582 audio features, which were computed from 38 LLDs and 21 functionals. The spectral features include loudness, mel-frequency cepstral coefficients, mel-frequency band energy, and line spectral pair frequencies. The prosodic and voice quality features comprise the pitch frequency and envelope, jitter, and shimmer. Functionals such as the mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, linear regression coefficients, and quartile and percentile coefficients were applied on the LLDs. The IS-2011 feature set consists of 4,368 audio features, which were computed from 59 LLDs and 39 functionals. Additional LLDs, such as the auditory spectrum-derived loudness

Table 18.3 Official feature sets of Interspeech Challenges

Feature set	IS-2010	IS-2011	IS-2012	IS-2013
# of LLDs	38	59	64	59
# of functional	21	39	40	48
# of features	1,582	4,368	6,124	6,373

measure, RASTA-style filtered auditory spectra, and statistical spectral descriptors (such as flux, entropy, variance) have been introduced. Additional functionals, such as quadratic regression and linear predictive coefficients and peak distances, allowed a better estimation of the temporal variability. The IS-2012 feature set consists of 6,124 audio features, which were computed from 64 LLDs and 40 functionals. Few LLDs have been added, including the logarithmic harmonics-to-noise ratio, spectral harmonicity, and psychoacoustic spectral sharpness. Functionals that are related to the local extrema, such as the statistics of inter-maxima distances, have been introduced. Useless functionals have been removed to limit the number of the audio features. The IS-2013 feature set consists of 6,373 audio features, computed from 59 LLDs and 48 functionals. A total of 724 audio features were removed from the IS-2012 feature set, and 972 were added. New functionals that were related to the local extrema, such as the modeling of inter-maxima, have been introduced.

Table 18.3 summarizes the main characteristics of the used feature sets. The first three feature sets were used for the detection of overlap, and the last feature set was the official feature set for the detection of conflict.

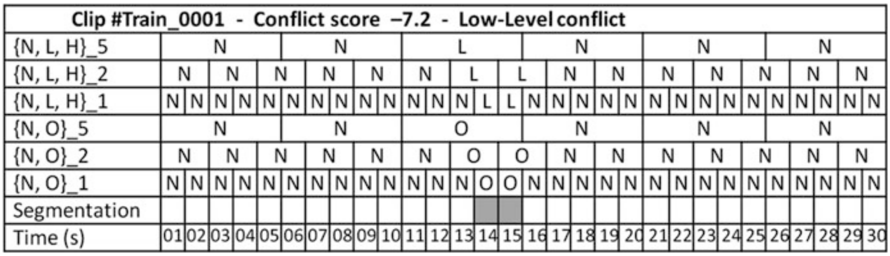
18.4 Interruption Detection

From the previous statistics analyzed in Sect. 18.2, the conflict level was shown to be highly correlated to the mean number of interruptions (cf. Fig. 18.2), the mean duration of overlap (cf. Fig. 18.3), and the percentage of overlap duration (cf. Fig. 18.4). Detecting segments of overlap is a difficult problem without individual microphones (Yamamoto et al. 2005). The main problem is due to the nonstationary characteristics of the speech signal. An alternative approach is the use of a microphone array (Quinlan and Asano 2007). In this case, the estimation of the number of signal sources allows the detection of segments that contain more than one source of speech. Another approach, which is applied for improving the speaker diarization system, is the speech segmentation by a three-class hidden Markov model (Boakye et al. 2008), with the three classes corresponding to nonspeech, speech, and overlapping speech. Mel-frequency cepstral coefficients (MFCC), root mean square (RMS) energy, and linear predictive coding (LPC) residual energy features have been used, and they provided a precision of 66 % and a recall of 26 %. In our approach, we have chosen to develop a multi-resolution framework to estimate the overlap duration percentage. This approach is based on the fusion of various overlap detectors, in which each detector is estimated on the segments of a fixed and chosen duration.

18.4.1 Clip Segmentation and Relabeling

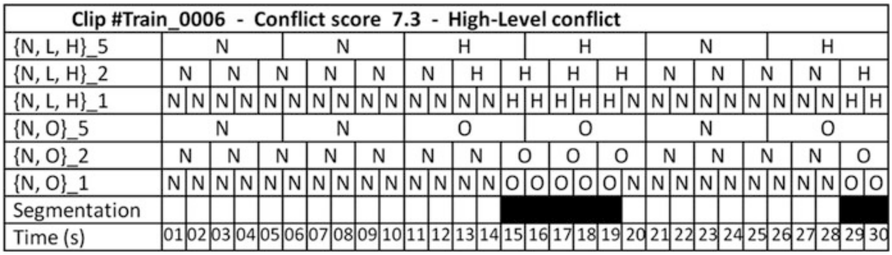
The clips were segmented into consecutive audio segments. Three segment durations were chosen for the multi-resolution: 1, 2, and 5 s. For a given duration of segment, two segment-based detectors were designed: (1) the first detector is a two-class classifier that is referred to as an {N, O}-detector; it classifies a segment into Non-Ov (N) or Ov (O), and (2) the second detector is a three-class classifier that is referred to as an {N, L, H}-detector, which classifies a segment into Non-Ov (N), LLC-Ov (L), or HLC-Ov (H). Then, for multi-resolution detection, six SVM-based overlap detectors have been developed: (1) three two-class SVM classifiers, which we called {N, O}_1, {N, O}_2, and {N, O}_5, for the three durations, and (2) three three-class SVM classifiers, which we called {N, L, H}_1, {N, L, H}_2, and {N, L, H}_5, for the three durations. These labels (N, O, H, and L) were computed from the SSPNet corpus metadata using speaker segmentation and conflict metadata. The Train and Development sets were relabeled using the multi-resolution framework of overlap localization. For each clip, diarization and conflict information are now represented by 102 labels: 60 labels for {N, O}_1 and {N, L, H}_1, 30 labels for {N, O}_2 and {N, L, H}_2, and 12 labels for {N, O}_5 and {N, L, H}_5. These new labels will be used for the training and testing of the various overlap detectors.

In Figs. 18.5 and 18.6, the row called *Time* gives the time in seconds in the range from 1 to 30 (i.e., the clip duration), and the row *Segmentation* is the representation of the diarization metadata of the clip: N-segments are colored in white, L-segments



Clip #Train_0001 - Conflict score –7.2 - Low-Level conflict																														
{N, L, H}_5	N					N					L					N					N					N				
{N, L, H}_2	N		N		N		N		N		N		L		L		N		N		N		N		N		N		N	
{N, L, H}_1	N	N	N	N	N	N	N	N	N	N	N	N	N	L	L	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
{N, O}_5	N					N					O					N					N					N				
{N, O}_2	N		N		N		N		N		N		O		O		N		N		N		N		N		N		N	
{N, O}_1	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Segmentation														▨	▨															
Time (s)	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Fig. 18.5 Train set relabeling for the Train_0001 clip of low-level conflict



Clip #Train_0006 - Conflict score 7.3 - High-Level conflict																														
{N, L, H}_5	N					N					H					H					N					H				
{N, L, H}_2	N		N		N		N		N		N		H		H		H		H		N		N		N		N		H	
{N, L, H}_1	N	N	N	N	N	N	N	N	N	N	N	N	N	N	H	H	H	H	H	N	N	N	N	N	N	N	N	N	H	H
{N, O}_5	N					N					O					O					N					O				
{N, O}_2	N		N		N		N		N		N		N		O		O		O		N		N		N		N		O	
{N, O}_1	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	O	O	N	N	N	N	N	N	N	N	N	O	O
Segmentation															■	■	■	■	■										■	■
Time (s)	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Fig. 18.6 Train set relabeling for the Train_0006 clip of high-level conflict

in gray, and H-segments in black. The other rows contain the relabeling according to the various detectors. For the three rows {N, O}_*x* (*x* ∈ {1, 2, 5}), a segment is labeled O when it contains a part of overlap and, otherwise, N. For the three rows {N, L, H}_*x* (*x* ∈ {1, 2, 5}), overlap segments are labeled according to the conflict level of the clip: L for LLC-Ov and H for HLC-Ov.

Figure 18.5 gives an instance of metadata relabeling for the LLC clip #Train_0001. For this clip, an LLC-Ov occurs over 13.01 and 14.4 s. The relabeling is O for the segments 14 and 15 of {N, O}_1, the segments 7 and 8 of {N, O}_2, and the segment 3 of {N, O}_5. The relabeling is L for the segments 14 and 15 of {N, L, H}_1, the segments 7 and 8 of {N, L, H}_2, and the segment 3 of {N, L, H}_5.

Figure 18.6 gives an instance of metadata relabeling for the HLC clip #Train_0006. For this clip, HLC-Ovs occur over 14.9 and 18.9 s and over 28.3 and 30 s. The relabeling is O for the segments 15, 16, 17, 18, 19, 29, and 30 of {N, O}_1, for the segments 8, 9, 10, and 15 of {N, O}_2, and for the segments 3, 4, and 6 of {N, O}_5. The relabeling is H for the segments 15, 16, 17, 18, 19, 29, and 30 of {N, L, H}_1, for the segments 7, 8, 9, 10, and 15 of {N, L, H}_2, and for the segments 3, 4, and 6 of {N, L, H}_5.

18.4.2 Two-Class {N, O} Classifiers

Using relabeling, three two-class SVMs ({N, O}_1, {N, O}_2, {N, O}_5) were estimated on the Train set. Each SVM classifies a segment of a given duration (1, 2, and 5 s) into overlap (O) or Non-Ov (N). To account for the imbalanced class distribution, the upper-represented category (N) was down-sampled by a given factor. A factor of 4 was applied for the {N, O}_1 detector, a factor of 3 for the {N, O}_2 detector, and a factor of 2 for the {N, O}_5 detector. We investigated the effects of different feature sets on the accuracy of the overlap speech detection. Table 18.4 gives the accuracy rates (N-Acc. and O-Acc. in %) of the two-class

Table 18.4 Accuracy rates of the detectors {N, O} on the Development set according to the feature sets. In bold, the best feature set

Detectors {N, O}	Feature set	N-Acc. (%)	O-Acc. (%)	UAR (%)
{N, O}_1	**IS-2010**	**86.7**	**73.9**	**80.3**
{N, O}_1	IS-2011	87.7	72.3	80.0
{N, O}_1	IS-2012	87.8	71.6	79.7
{N, O}_2	**IS-2010**	**85.1**	**75.1**	**80.1**
{N, O}_2	IS-2011	87.3	71.6	79.5
{N, O}_2	IS-2012	87.4	71.7	79.5
{N, O}_5	**IS-2010**	**82.7**	**78.7**	**80.7**
{N, O}_5	IS-2011	84.9	75.3	80.1
{N, O}_5	IS-2012	84.0	75.7	79.8

classifiers on the two classes (N and O) on the Development set. Three audio feature sets were compared: IS-2010, IS-2011, and IS-2012 (cf. Sect. 18.3.1). For all two-class classifiers, the overlaps are more difficult to detect than the Non-Ovs, and IS-2010 was the best feature set, with a UAR (in %) of slightly over 80 % for the three detectors. For the architecture of the conflict detector, we chose to use only the best two-class classifiers {N, O}_1, {N, O}_2, and {N, O}_5 with the IS-2010 audio feature set. Our assumption is that only the best overlap classifiers are relevant for the detection of conflict.

18.4.3 Three-Class {N, L, H} Classifiers

Previous studies presented different typologies of overlaps: overlap and backchannel with overlap (Gravano and Hirschberg 2011) and competitive and collaborative overlaps (Oertel et al. 2012). A backchannel indicates that the speaker producing them follows and understands the other speaker. They are generally words, onomatopoeias, or other sounds produced in the background (Clancy et al. 1996). Collaborative or competitive interruptions are manifested by speech overlap, but only overlap from a competitive interruption can it be related to a conflict (Kurtié et al. 2012). In competitive overlaps, the incoming speaker attempts to forcefully take over the turn. In collaborative overlaps, the incoming speaker assists the current speaker in his or her speech. We chose to build classes of LLC-Ovs and HLC-Ovs by making the hypothesis that they would be separable acoustically and useful for conflict detection. This choice is supported by the observation that some of the LLC-Ovs of the Train set were backchannel with overlaps and/or collaborative overlaps.

Using relabeling, three three-class SVM classifiers ({N, L, H}_1, {N, L, H}_2, and {N, L, H}_5) were estimated on the Train set. Each SVM classifies a segment of a given duration (1, 2, and 5 s) into an H, L, or N. To account for the imbalanced class distribution, the upper-represented category (N) was down-sampled by a given factor. A factor of 8 was applied for the {N, L, H}_1 detector, a factor of 6 for the {N, L, H}_2 detector, and a factor of 3 for the {N, L, H}_5 detector. We investigated the effects of different feature sets on the accuracy rate of the overlap speech detection. Table 18.5 gives the accuracy rates of the three-class classifiers on the Development set. Three audio feature sets were compared: IS-2010, IS-2011, and IS-2012. IS-2010 was the best feature set for {N, L, H}_1, having a UAR of 61.1 %. IS-2011 was the best feature set for {N, L, H}_2, with a UAR of 61.3 %. IS-2010 was the best feature set for {N, L, H}_5, with a UAR of 63.5 %. The LLC-Ovs are more difficult to detect than the HLC-Ovs. Furthermore, the detection rate of the LLC-Ovs appears to decrease with the duration of the analyzed segment: 44.7 % for {N, L, H}_5 (5 s), 35.9 % for {N, L, H}_2 (2 s), and 31.7 % for {N, L, H}_1 (1 s). A possible explanation would be that the detector {N, L, H}_5 allows a better estimation of the overlap durations than the other detectors and, consequently, a better discrimination of the LLC- and HLC-Ovs. Indeed, the duration of the LLC-Ovs is lower on average than the HLC-Ovs (1.98 s vs. 2.75 s). For the architecture of the conflict detector,

Table 18.5 Accuracy rates of the detectors {N, L, H} on the Development set according to the feature sets. In bold, the best detector

Detectors	Feature set	N-Acc. (%)	L-Acc. (%)	H-Acc. (%)	UAR (%)
{N, L, H}_1	IS-2010	78.0	31.7	73.5	61.1
{N, L, H}_1	IS-2011	79.9	32.7	70.5	61.0
{N, L, H}_1	IS-2012	79.4	31.4	71.4	60.7
{N, L, H}_2	IS-2010	79.5	32.6	71.2	61.2
{N, L, H}_2	IS-2011	78.1	35.9	70.0	61.3
{N, L, H}_2	IS-2012	80.5	31.5	68.0	60.0
{N, L, H}_5	**IS-2010**	**77.5**	**44.7**	**68.3**	**63.5**
{N, L, H}_5	IS-2011	76.4	40.0	67.7	61.4
{N, L, H}_5	IS-2012	80.8	38.2	67.4	62.1

we chose to use only the best three-class classifier: {N, L, H}_5 with the IS-2010 audio feature set. Our assumption is that only the best overlap classifier is relevant for the detection of conflict.

18.4.4 Audio Characteristics of Overlaps

Previous studies (Smolenski and Ramachandran 2011; Shokouhi et al. 2013) have shown that the audio characteristics of overlapping speech are different from speech in which a lonely speaker occurs. We looked for the discriminating cues (1) between Ov and Non-Ov and (2) more specifically between HLC-Ov and LLC-Ov. For these investigations, we chose to study the segments that had a 5-s duration in the Train set for the best accuracy results of the 5-s-based {N, O} and {N, L, H} detectors (see, respectively, Tables 18.4 and 18.5 in Sect. 18.4). The 38 low-level descriptors (LLDs) of the IS-2010 feature set have been used as audio characteristics. The relevance of the LLDs was analyzed with respect to the classes Non-Ov/Ov, which are referred to as {N, O}, and the HLC-Ov/LLC-Ov, which are referred to as {H, L}. For each LLD, the relevance is given by the information gain (Rauber and Steiger-Garcao 1993), which is computed on the segments of 5-s duration with the following formula: $H(\text{class}) - H(\text{class/LLD})$, where H is the Shannon entropy. Four steps were defined to compute the entropy: (1) filtering of the IS-2010 features according to a given LLD, (2) clustering of the segments of the Train set using the filtered features, (3) computation of the contingency table from the class and the cluster associated with each segment, and (4) estimation of the entropy from the table of contingency. Table 18.6 gives the information gain computed on the Train set of the five best-ranked LLDs (over 38 LLDs) in discriminating LLC-Ovs and HLC-Ovs. The most relevant LLDs are the logarithmic powers of mel-frequency bands and, more precisely, the high-frequency bands and the normalized loudness. These results show that various acoustic differences exist between the two types of overlaps.

Table 18.6 Information gain of the five best-ranked LLDs of the IS-2010 audio feature set in discriminating LLC-Ovs and HLC-Ovs

Low-level descriptors (LLD)	Inf. gain	Rank (/38)
Log power [3,934–5,649 Hz]	0.130	1
Log power [2,682–3,934 Hz]	0.119	2
Log power [1,768–2,682 Hz]	0.107	3
Normalized loudness	0.102	4
Log power [5,649–8,000 Hz]	0.102	5

Table 18.7 Information gain of the five best-ranked LLDs of the IS-2010 audio feature set in discriminating Ovs and Non-Ovs

Low-level descriptors (LLD)	Inf. gain	Rank (/38)
Fundamental frequency (F0)	0.141	1
Log power [614–1,101 Hz]	0.129	2
Log power [0–259 Hz]	0.127	3
Jitter (DDP)	0.124	4
First mel-frequency cepstral coef.	0.121	5

Table 18.7 gives the information gain that is computed on the Train set of the five best-ranked LLDs (over 38 LLDs) in discriminating Ovs and Non-Ovs. According to the information gain rank, the most relevant LLDs are the fundamental frequency, the logarithmic powers, especially in low-frequency bands, the jitter, and the first mel-frequency cepstral coefficient. The usual representation techniques and algorithms are designed and interpreted for speech signals in which a lonely subject is speaking. In the case of overlapping speech in which two or more subjects are speaking, the usual algorithms are not adapted (e.g., the pitch algorithm); the computation of one fundamental frequency has no sense, and its computation was shown to be the most discriminant cue for detecting Ov/Non-Ov. For a speech representation such as the logarithmic power in the mel-frequency bands, the low-frequency bands in which the first two formants of the speaker occur were also shown to be discriminant. Last, the jitter DDP (difference of differences of periods) related to the pitch and the first mel-frequency cepstral coefficient related to the energy of the segment were also shown to be relevant for the discrimination Ov/Non-Ov.

18.5 Conflict Detector

Overlap detectors have been developed and assessed, to incorporate their knowledge in an improved conflict detector (conflict/nonconflict). Incorporating prior knowledge (Krupka and Tishby 2007; Li et al. 2008) in classification systems allowed an increase in the performance in many applications of pattern recognition (e.g., biomedical image, pathological voice). Various methods have been developed for neural network systems (Chen et al. 2000) and SVM classifiers (Decoste and Scholkopf 2002; Lauer and Bloch 2008). As defined by Schölkopf and Smola (2001), the methods developed for including prior knowledge in an

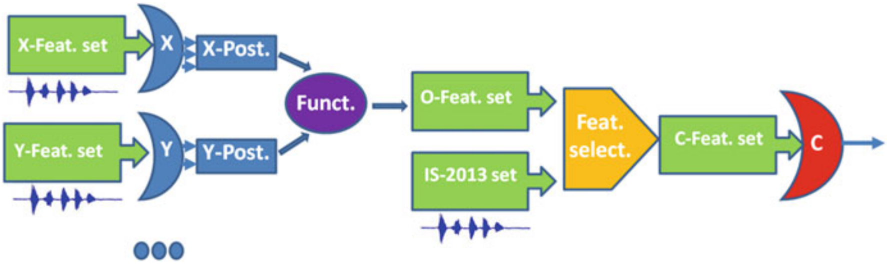


Fig. 18.7 Multi-expert architecture scheme of the conflict detector

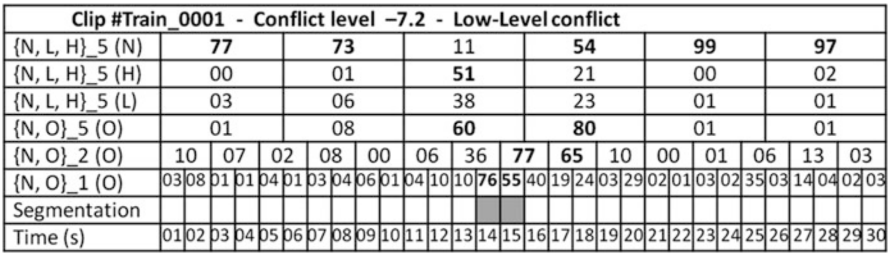
SVM classifier can be divided into three categories: (1) the kernel methods with selection of the most appropriate kernel or the creation of a new kernel, (2) the optimization methods with the addition of constraints, and (3) the sample methods with data generation or modification of data representations. We have chosen the last category by developing an SVM-based detector, using as input a composite feature set. This feature set is a concatenation of selected audio features and posterior-based features that are computed from the posterior probabilities of the overlap detectors. The architecture characteristics of this classification system are close to those used in a mixture of experts (Jordan and Jacobs 1994). These approaches have theoretical advantages, such as a reduction in the hypothesis space and learning consistency. As described in Fig. 18.7, the multi-expert architecture scheme of the conflict detector has consisted of a set of overlap detectors (e.g., X, Y) and a conflict/nonconflict detector (C). A specialized audio feature set (e.g., X-Feat. set) was associated with each overlap detector (e.g., X), to represent the utterances. A conflict audio feature set (Cf-Feat. set) was associated with the conflict detector. This feature set consisted of the selection of the relevant features (Feat. Select.) that were extracted from the overlap feature set (Ov-Feat. set) and the IS-2013 feature set (cf. Sect. 18.3.1). A set of functionals (Funct.) was applied to the posterior probabilities of the overlap detectors (e.g., X-Post and Y-Post) to obtain the Ov-Feat. set.

We chose the overlap detectors giving the best UAR on the Development set (cf. Table 18.4 and 18.5): three two-class (Non-Ov/Ov) SVM-based detectors ({N, O}_1, {N, O}_2, {N, O}_5) and one three-class (Non-Ov/LLC-Ov/HLC-Ov) SVM-based detector ({N, L, H}_5).

18.5.1 Posterior Probabilities

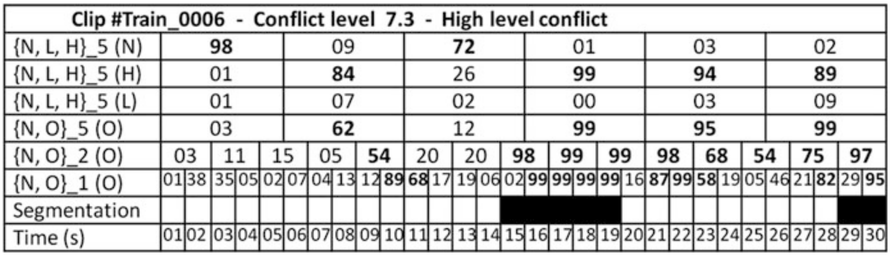
Logistic regression models (Hosmer and Lemeshow 2000) were used to obtain the posterior probabilities from the four overlap detectors ({N, O}_1, {N, O}_2, {N, O}_5, and {N, L, H}_5). These posterior probabilities of the overlap detectors provide information about the uncertainty of belonging to one class: for example, the probability of 60 % of a segment to be an overlap and 40 % to be a nonoverlap. There are various strategies for computing these probabilities, such as Platt's method (Platt 2000), isotonic regression (Zadrozny and Elkan 2002), and Bayesian methods (Sollich 2002). These probabilities are useful to integrate expert classifiers such as overlap classifiers in a global decision process. This approach is a flexible architecture for making decisions without global optimization. The method of computation of the posterior probabilities depends on the chosen set of clips. The goal is to obtain a consistent computation of the posterior probabilities from the different corpora (Train, Development, and Test sets). For the Train and Development sets, the posterior probabilities have been computed by performing cross-predictions on the union of these two sets. This process consists of splitting the data set into s disjoint folds and predicting class posterior probabilities of each instance of a fold from a model trained on the $s-1$ other folds. Sixteen folds have been chosen that have participant independence between two folds. For the Test set, the posterior probabilities have been computed from a model trained on the union of the Train and Development sets. A total of 120 posterior probabilities were computed for each clip: 60 for the {N, O}_1 detector, 30 for {N, O}_2, 12 for {N, O}_5, and 18 for {N, L, H}_5.

Figures 18.8 and 18.9 give an instance of the posterior probabilities from the four overlap detectors ({N, O}_1, {N, O}_2, {N, O}_5, and {N, L, H}_5), respectively, for the LLC clip #Train_0001 and the HLC clip #Train_0006. The row called *Time* gives



Clip #Train_0001 - Conflict level –7.2 - Low-Level conflict																														
{N, L, H}_5 (N)	**77**					**73**					11					**54**					**99**					**97**				
{N, L, H}_5 (H)	00					01					**51**					21					00					02				
{N, L, H}_5 (L)	03					06					38					23					01					01				
{N, O}_5 (O)	01					08					**60**					**80**					01					01				
{N, O}_2 (O)	10		07		02		08		00		06		36		**77**		**65**		10		00		01		06		13		03	
{N, O}_1 (O)	03	08	01	01	04	01	03	04	06	01	04	10	10	**76**	**55**	40	19	24	03	29	02	01	03	02	35	03	14	04	02	03
Segmentation														▨	▨															
Time (s)	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Fig. 18.8 Overlap posterior probabilities as percentages for the Train_0001 clip with low-level conflict



Clip #Train_0006 - Conflict level 7.3 - High level conflict																														
{N, L, H}_5 (N)	**98**					09					**72**					01					03					02				
{N, L, H}_5 (H)	01					**84**					26					**99**					**94**					**89**				
{N, L, H}_5 (L)	01					07					02					00					03					09				
{N, O}_5 (O)	03					**62**					12					**99**					**95**					**99**				
{N, O}_2 (O)	03		11		15		05		**54**		20		20		**98**		**99**		**99**		**98**		**68**		**54**		**75**		**97**	
{N, O}_1 (O)	01	38	35	05	02	07	04	13	12	**89**	**68**	17	19	06	02	**99**	**99**	**99**	**99**	16	**87**	**99**	**58**	19	05	46	21	**82**	29	**95**
Segmentation															■	■	■	■	■										■	■
Time (s)	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Fig. 18.9 Overlap posterior probabilities as percentages for the Train_0006 clip with high-level conflict

the time from 1 to 30 s (clip duration); the row *Segmentation* is the representation of the diarization metadata of the clip: N-segments are colored in white, L-segments in gray, and H-segments in black. The other rows contain the posterior probabilities presented as a percentage. For the three rows {N, O}_*x* (*x* ∈ {1, 2, 5}), a segment of posterior probabilities that was higher than 50 was detected as O; otherwise, it was detected as N. The posterior probabilities that were associated with the {N, L, H}_5 detector are presented in the three other rows {N, L, H}_5 (N), {N, L, H}_5 (L), and {N, L, H}_5 (H) for, respectively, nonoverlap (N), low-level-conflict (L), and high-level-conflict (H). For a given segment, the higher probability (in bold) corresponds to the class that was detected.

In Fig. 18.8, the class O was detected for the segments 14 and 15 of {N, O}_1, the segments 8 and 9 of {N, O}_2, and the segments 3 and 4 of {N, O}_5. Class H was detected for the segment 3 for {N, L, H}_5. There are three wrong detections: the class O instead of N for the segment 9 of {N, O}_2 and the segment 4 of {N, O}_5, and the class H was detected instead of L for segment 3 for {N, L, H}_5.

In Fig. 18.9, the class O was detected for the segments 10, 11, 16, 17, 18, 19, 21, 22, 23, 28, and 30 of {N, O}_1, for the segments 5 and 8 through 15 of {N, O}_2, and for the segments 2, 4, 5, and 6 of {N, O}_5. Class H was detected for the segments 2, 4, 5, and 6 of {N, L, H}_5. There are ten wrong detections: the class O instead of N for the segments 10, 11, and 28 of {N, O}_1, the segments 5, 13, and 14 of {N, O}_2, the segment 2 of {N, O}_5, the class N instead of O for the segments 15 and 29 of {N, O}_1, and the class H instead of N for the segment 2 of {N, L, H}_5. We note that there was no wrong decision for segments 21, 22, and 23 of {N, O}_1, for the segments 11 and 12 of {N, O}_2, and for segment 5 of {N, O}_5 and {N, L, H}_5; after listening, an overlap occurs effectively from 20.3 to 22.2 s but was not labeled in the metadata.

18.5.2 Overlap Feature Sets

One hundred and twenty posterior probabilities were computed for each clip. These values depend on the time and represent the temporal shape of a conflict in terms of the overlap. There are specific temporal shapes for conflict escalation (Kim et al. 2012c), but the 797 clips of the Train set are insufficient to model these temporal shapes. We have chosen to apply statistical functionals to the posterior probabilities; the purpose was to obtain an overlap feature set that is related to the percentage of overlap duration. Three functionals have been chosen: mean, correlation, and covariance. The mean functional was applied to the posterior probabilities of {N, O}_1, {N, O}_2, {N, O}_5 for the class O and to the posteriors of {N, L, H}_5 for the classes N, L, and H. The correlation functional was applied between the posterior probabilities of the class O for all combinations of {N, O}_1, {N, O}_2 and {N, O}_5. Table 18.8 gives a list of the ten features that were computed by the mean and correlation functionals.

Functional covariance is a functional of a functional. It was applied to the mean and correlation functionals. The interest of this functional is to reveal the cofactors.

Table 18.8 List of the features computed by the mean and correlation functionals

Mean and correlation functionals	Feature name
Mean (post ({N, O}_1 (O)))	O1
Mean (post ({N, O}_2 (O)))	O2
Mean (post ({N, O}_5 (O)))	O5
Correlation (post ({N, O}_1 (O)), post ({N, O}_2 (O)))	O12
Correlation (post ({N, O}_1 (O)), post ({N, O}_5 (O)))	O15
Correlation (post ({N, O}_2 (O)), post ({N, O}_5 (O)))	O25
Correlation (post ({N, O}_1 (O)), post ({N, O}_2 (O)), post({N, O}_5 (O)))	O125
Mean (post ({N, L, H}_5 (N)))	N5
Mean (post ({N, L, H}_5 (L)))	L5
Mean (post ({N, L, H}_5 (H)))	H5

Table 18.9 Information gain of the 15 best-ranked LLDs of the audio feature set, including the Ov-2 feature set and the IS-2013 feature set

Features	Information gain	Rank (/6,428)
Cov_O125_O125	0.43862	1
Cov_O12_O125	0.43758	2
Cov_O1_O12	0.43586	3
Cov_O1_O125	0.43177	4
Cov_O15_O125	0.42914	5
Cov_O12_O12	0.42858	6
Cov_O25_O125	0.41965	7
Cov_O12_O15	0.41957	8
Cov_O12_O25	0.41431	9
Cov_O15_O15	0.41429	10
Cov_O15_O25	0.41325	11
Cov_H5_O1	0.40915	12
Cov_H5_O15	0.40849	13
Cov_O1_O25	0.40705	14
Cov_H5_O12	0.40509	15

Two overlap feature sets have been defined. The first feature set, called Ov-1, consisted of 28 features; it was computed by the covariance functional applied to the features that are related to the {N, O} detectors (O1, O12, O15, O2, O25, and O125). The second feature set, called Ov-2, consisted of 55 features; it was computed by the covariance functional applied to the features that are related to the {N, O} and {N, L, H} detectors (O1, O12, O15, O2, O25, O125, N5, L5, and H5). These two feature sets will allow a contrastive test to measure the contribution of the {N, L, H}_5 detector in the detection of conflict. The method of information gain was used to analyze the feature relevance of the Ov-2 set in comparison with those of the IS-2013 set. Table 18.9 gives the information gain computed on the Train set and the rank on 6,428 features (55 features from the Ov-2 set and 6,373 features from the IS-2013 set) of the most relevant features for the conflict detection. The best feature is the Cov_O125_O125 feature (which is equal to O125 multiplied by O125).

Table 18.10 Characteristics of the conflict feature sets

Feature set	Selected feat. set	# of selected feat.	# of selected feat. from Ov features	# of selected feat. from IS-2013
Ov-1 and IS-2013 (6,428 features)	Cf-1	315	15	300
Ov-2 and IS-2013 (6,401 features)	Cf-2	335	45	290

Table 18.11 UAR in the conflict detection task on the Development set according to the conflict feature sets

Feature set	IS-2013	Cf-1	Cf-2
# of features	6,373	315	335
UAR (devel. set) (%)	79.1	87.4	88.3

The 12th rank of the Cov_H5_O1 feature shows that the {N, L, H}_5 detector is relevant for the detection of conflict. A total of 36 out of 55 features of the Ov-2 set have better information gain than those of the IS-2013 set. These results show the interest of the overlap feature sets for the detection of conflict.

18.5.3 Conflict Feature Sets

From two initial feature sets (Ov-1 and IS-2013 and Ov-2 and IS-2013), two conflict feature sets (Cf-1 and Cf-2) were selected by a backward selection algorithm when maximizing UAR on the Development set for the conflict detection task. Table 18.10 gives the characteristics of the Cf-1 and Cf-2 sets of the conflict detector using these feature sets. The Cf-1 feature set consists of 315 features (15 features from the Ov-1 set and 300 features from the IS2013 set). The Cf-2 feature set consists of 335 features (45 features from Ov-2 set and 290 features from the IS2013-set).

Table 18.11 gives the accuracy (UAR in %) of the conflict detection on the Development set using the various feature sets (IS-2013, Cf-1, Cf-2). The results show an improvement of 8.3 % using the Cf-1 set and 9.2 % using the Cf-2 set on the Development set compared to the baseline results that use the IS-2013 set (UAR of 79.1 %). These results show also that the majority of the features of the Cf-2 set are relevant and not redundant. It confirms that the two types of detectors ({N, L, H} and {N, O}) are relevant for the detection of conflict.

18.5.4 Conflict Detectors

Two conflict detectors have been developed. Figure 18.10 resumes the architecture characteristics of the first conflict detector, called the simple overlap-based conflict detector (SO-conflict detector). This detector was based on a set of overlap detectors

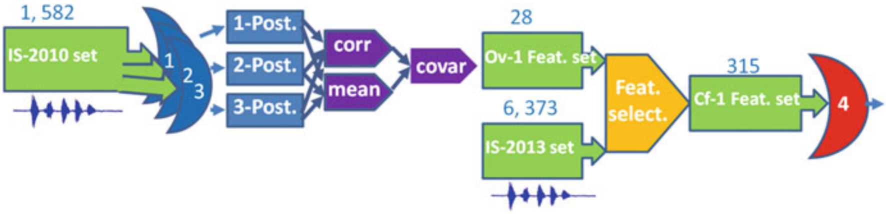


Fig. 18.10 Architecture scheme of the SO-conflict detector

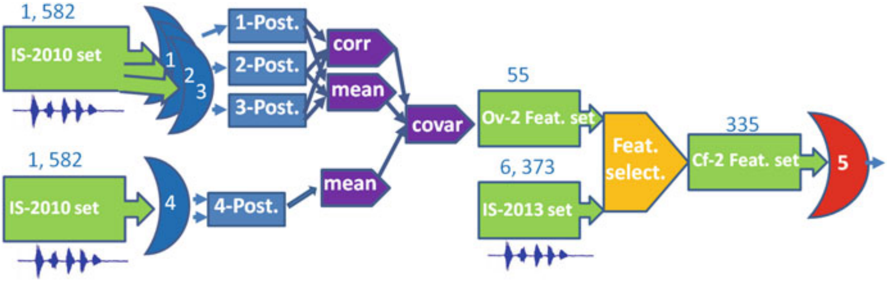


Fig. 18.11 Architecture scheme of the AO-conflict detector

(1, 2, and 3) that correspond to the three multi-resolution-based {N, O} detectors and a conflict detector (4). The IS-2010 feature set (1,582 features) was used for the overlap detectors. The Cf-1 feature set (315 features) was associated with the conflict detector. The Cf-1 feature set was obtained by a backward selection algorithm from the Ov-1 feature set (28 features) and the IS-2103 set (6,373 features).

Figure 18.11 resumes the architecture characteristics of the second conflict detector, called the advanced overlap-based conflict detector (AO-conflict detector). This detector was based on a set of overlap detectors (1, 2, 3, and 4) and a conflict detector (5). The IS-2010 audio feature set (1,582 features) was used for the overlap detectors. The Cf-2 feature set (335 features) was associated with the conflict detector. The Cf-2 feature set was obtained by a backward selection algorithm from the Ov-2 feature set (55 features) and the IS-2103 set (6,373 features).

18.5.5 Conflict Detection on the Test Set

The Test set of the Interspeech 2013 Conflict Challenge (Schuller et al. 2013) consisted of 397 clips with no information or metadata available. Table 18.12 gives the results obtained on the Test set during the Conflict Challenge. Experiments gave a UAR of 83.4 % for the SO-conflict detector and a UAR of 85.3 % for the AO-conflict detector. These results show an improvement of 2.6 % (SO-conflict detector)

Table 18.12 Assessment on the Test set. In bold, the best conflict detector

Conflict detector	# of features	UAR (%) on Test set
SO-conflict detector	315	83.4
AO-conflict detector	**335**	**85.3**
IS-2013 baseline system	6,373	80.8
Grèzes et al. (2013)	1	83.1
Räsänen and Pohjalainen (2013)	349	83.9

and 4.5 % (AO-conflict detector) on the Test set compared to the baseline results with the IS-2013 set (UAR of 80.8 %) for the conflict detection task. These results confirm also that the two types of overlap detectors ({N, L, H} and {N, O}) are relevant for the detection of conflict. The other results are those obtained by the other participants. In Grèzes et al. (2013), a UAR of 83.1 % was obtained on the Test set using a unique feature: the percentage of overlap predicted by an SVM-based regression model. In Räsänen and Pohjalainen (2013), a UAR of 83.9 % was obtained on the Test set using 349 relevant features selected from the IS-2013 feature set. Feature relevance was computed by a random process. We notice that the two better results were obtained by a similar number of features (335 vs. 349).

18.6 Conclusions

This article presents and assesses a detection system of conflict in group discussions from voice analysis. The system was based on a multi-expert architecture and detected two states (conflict/nonconflict). The analysis of the Train set of the SSPNet database has demonstrated that the conflict level was highly correlated with the mean number of interruptions, the mean duration of overlap, and the percentage of overlap duration. The multi-expert architecture enabled knowledge regarding overlaps to be used in the conflict detector.

The concept of LLC-Ovs and HLC-Ovs has been introduced and investigated. Two types of overlap detectors have been developed: the first type aims at detecting whether a speech segment contains overlap, and the second type aims at detecting whether a speech segment contains an LLC-Ov or HLC-Ov. The accuracy of the detectors shows that the LLC-Ovs and HLC-Ovs can be modeled. The high-frequency mel bands and the normalized loudness are shown to be the audio characteristics that are relevant to discriminating these two types of overlap. A multi-resolution framework has been developed for the overlap detectors, to improve the robustness of the detection. Three segment durations have been chosen (1, 2, and 5 s). The experiments have shown that these detectors were not redundant.

A composite set of 335 features, which consist of audio-based features and overlap detector-based features, has been defined for the conflict detection task of the Interspeech 2013 Conflict Challenge. The performance obtained for the Test set gave a UAR of 85.3 %. These results show an improvement of 4.5 % compared to the results of the baseline system of the Conflict Challenge (UAR of 80.8 %).

These experiments have shown the capability of a multi-expert architecture to integrate a piece of conflict knowledge. Other knowledge that is related to the turn-taking patterns, such as the modeling of the moderator role (Vinciarelli 2007), or that is related to the nonverbal interactions, such as the movements of the body, the head, and the arms, could be integrated into the conflict detector.

Acknowledgments Many thanks to Björn Schuller (TUM, Germany), Stefan Steidl (FAU Erlangen-Nuremberg, Germany), and Anton Batliner (TUM, Germany) for the organization of the Interspeech 2013 Conflict Challenge and special thanks to Alessandro Vinciarelli (University of Glasgow, UK) for the SSPNet Conflict Corpus.

References

Atkinson JM, Drew P (1979) Order in court: the organisation of verbal interaction in judicial settings. Humanities Press, Atlantic Highlands, NJ

Barzilay R, Collins M, Hirschberg J, Whittaker S (2000) The rules behind the roles: identifying speaker roles in radio broadcasts. Paper presented at 17th National conference on artificial intelligence, Austin, USA, 30 July–3 Aug, pp 679–684

Beattie GW (1982) Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. Semiotica 39(1–2):93–114

Boakye K, Trueba-Hornero B, Vinyals O, Friedland G (2008) Overlapped speech detection for improved diarization in multi-party meetings. Paper presented at ICASSP Conference, Las Vegas, USA, 31 Mar–4 Apr, pp 4353–4356

Boden D (1994) The business of talk. Organizations in action. Polity Press, London

Brinson SL, Winn JE (1997) Talk shows' representations of interpersonal conflicts. J Broadcast Electron Media 41(1):25–39

Chen Z, Feng TJ, Houkes Z (2000) Incorporating a priori knowledge into initialized weights for neural classifier. Paper presented at international joint conference on neural networks (IJCNN), Como, Italy, 24–27 July, pp 291–296

Clancy PM, Thompson SA, Suzuki R, Tao H (1996) The conversational use of reactive tokens in English, Japanese and Mandarin. J Pragmat 26:355–387

Daly TM, Lee JA, Soutar GN, Rasmi S (2010) Conflict-handling style measurement: a best-worst scaling application. Int J Confl Manag 21(3):281–308

De Ruiter JP, Mitterer H, Enfield NJ (2006) Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. Language 82(3):515–535

Decoste D, Scholkopf B (2002) Training invariant support vector machines. Mach Learn 46(1–3):161–190

Eyben F, Wöllmer M, Schuller B (2010) openSMILE the Munich versatile and fast open-source audio feature extractor. Paper presented at the ACM multimedia conference (MM), Florence, Italy, 25–29 Oct, pp 1459–1462

Garcia A (1991) Dispute resolution without disputing: how the interactional organization of mediation hearings minimizes argumentative talk. Am Sociol Rev 56:818–835

Gravano A, Hirschberg J (2011) Turn-taking cues in task oriented dialogue. Comput Speech Lang 25(3):601–634

Grèzes F, Richards J, Rosenberg A (2013) Let me finish: automatic conflict detection using speaker overlap. Paper presented at the Interspeech conference, Lyon, France, 25–29 Aug, 5 pages

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

Heath C, Luff P (2007) Ordering competition: the interactional accomplishment of the sale of art and antiques at auction. Br J Sociol 58:63–85

Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York

Jordan MI, Jacobs RA (1994) Hierarchical mixtures of experts and the EM algorithm. Neural Comput 6:181–214

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

Kim S, Valente F, Vinciarelli A (2012b) Automatic detection of conflicts in spoken conversations: ratings and analysis of broadcast political debates. Paper presented at ICASSP, Kyoto, Japan, 25–30 Mar, pp 5089–5092

Kim S, Yella SH, Valente FA (2012c) Automatic detection of conflict escalation in spoken conversations. Paper presented at Interspeech Conference, Portland, USA, OR, 9–13 Sept, 4 pages

Korabik K, Baril GL, Watson C (1993) Managers' conflict management style and leadership effectiveness: the moderating effects of gender. Sex Roles 29(5–6):405–418

Krupka E, Tishby N (2007) Incorporating prior knowledge on features into learning. J Mach Learn Res 2:227–234

Kurtié E, Brown GJ, Wells B (2012) Resources for turn competition in overlapping talk. Speech Comm 55:1–23. doi:10.1016/j.specom.2012.10.002

Lauer DF, Bloch G (2008) Incorporating prior knowledge in support vector machines for classification: a review. Neurocomputing 71(7–9):1578–1594

Lerner GH (1995) Turn design and the organization of participation in instructional activities. Discourse Process 19(1):111–131

Li Y, de Ridder D, Duin RPW, Reinders MJT (2008) Integration of prior knowledge of measurement noise in Kernel density classification. Pattern Recogn 41:320–330

Mac Houl A (1978) The organization of turns at formal talk in the classroom. Lang Soc 7:183–213

Macintosh G, Stevens CJ (2008) Personality, motives and conflict strategies in everyday service encounters. Int J Confl Manag 19(2):112–131

Mehan H (1985) The structure of classroom discourse. In: Dijk TA (ed) Handbook of discourse analysis, vol 3. Academic, New York, pp 120–131

Mondada L (2012) The dynamics of embodied participation and language choice in multilingual meetings. Lang Soc 41:1–23

Mondada L (2013) Embodied and spatial resources for turn-taking in institutional multi-party interactions: participatory democracy debates. J Pragmat 46(1):39–68

Oertel C, Wlodarczak M, Tarasov A, Campbell N, Wagner P (2012) Context cues for classification of competitive and collaborative overlaps. Paper presented at Speech Prosody Conference, Shanghai, China, 22–25 May, 4 pages

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(2):533–540

Platt JC (2000) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Bartlett PJ, Schölkopf B, Schuurmans D, Smola AJ (eds) Advances in large margin classifiers. MIT Press, Cambridge, pp 61–74

Quinlan A, Asano F (2007) Detection of overlapping speech in meeting recordings using the modified exponential fitting test. Paper presented at the European signal processing conference, Poznan, Poland, 3–7 Sept, pp 2360–2364

Rahim MA (1983) A measure of styles of handling interpersonal conflict. Acad Manag J 26(2):368–376

Räsänen O, Pohjalainen J (2013) Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. Paper presented at the Interspeech conference, Lyon, France, 25–29 Aug, 5 pages

Rauber TW, Steiger-Garcao AS (1993) Feature selection of categorical attributes based on contingency table analysis. Paper presented at the Portuguese conference on pattern recognition, Porto, Portugal

Sacks H, Schegloff EA, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50(4):696–735

Schegloff EA (1987) Between macro and micro: contexts and other connections. In: Alexander J, Giesen B, Munch R, Smelser N (eds) The micro-macro link. University of California Press, Berkeley, pp 207–234

Schölkopf BAJ, Smola AJ (2001) Learning with Kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge, MA

Schuller B, Wimmer M, Moesenlechner L, Kern C, Arsic D, Rigoll G (2008) Brute-forcing hierarchical functional for paralinguistics: a waste of feature space? Paper presented at the ICASSP conference, pp 4501–4504

Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2010) The Interspeech 2010 paralinguistic challenge. Paper presented at the Interspeech conference, Makuhari, Japan, 26–30 Sept, pp 2794–2797

Schuller B, Batliner A, Steidl S, Schiel F, Krajewski J (2011) The Interspeech 2011 speaker state challenge. Paper presented at the Interspeech conference, Florence, Italy, 28–31 Aug, 4 pages

Schuller B, Steidl S, Batliner A, Noth E, Vinciarelli A, Burkhardt F, van Son R, Weninger F, Eyben F, Bocklet T, Mohammadi G, Weiss B (2012) The Interspeech 2012 speaker trait challenge. Paper presented at the Interspeech conference, Portland, OR, USA, 9–13 Sept, 4 pages

Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S (2013) The Interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion autism. Paper presented at the Interspeech conference, Lyon, France, 25–29 Aug, 5 pages

Shokouhi N, Sathyanarayana A, Sadjadi SO, Hansen JHL (2013) Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems. Paper presented at ICASSP conference, Vancouver, Canada, 26–31 May, pp 2834–2838

Smolenski B, Ramachandran R (2011) Usable speech processing: a filterless approach in the presence of interference. Circuits Syst Mag IEEE 11(2):8–22

Sollich P (2002) Bayesian methods for support vector machines: evidence and predictive class probabilities. Mach Learn 46:21–52

Svennevig J (2008) Exploring leadership conversations. Manag Commun Q 21:529–536

Thomas KW, Thomas GF, Schaubhut N (2008) Conflict styles of men and women at six organization levels. Int J Confl Manag 19(2):148–166

Valente F, Vinciarelli A (2010) Improving speech processing trough social signals: automatic speaker segmentation of political debates using role based turn-taking patterns. Paper presented at the International workshop on social signal processing, Firenze, Italy, 25–29 Oct, pp 29–34

Vinciarelli A (2007) Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. IEEE Trans Multimed 9(6):1215–1226

Vinciarelli A (2009) Capturing order in social interactions. Signal Process Mag IEEE 26(5):133–152

Vogt T, André E (2005) Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. Paper presented at the ICME conference, Amsterdam, The Netherlands, 6–8 July, pp 474–477

Yamamoto K, Asano F, Yamada T, Kitawaki N (2005) Detection of overlapping speech in meetings using support vector regression. Paper presented at the international workshop on acoustic echo and noise control (IWAENC), Eindhoven, The Netherland, 12–15 Sept, pp 2158–2165

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. Paper presented at the international conference on knowledge discovery and data mining, Edmonton, Canada, 23–25 July, pp 694–699