PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Computational Modeling of Turn-Taking Dynamics in Spoken Conversations

Shammur Absar Chowdhury

Advisor:

Prof. Giuseppe Riccardi	*University of Trento*

Examining Committee:

Prof. Raffaella Bernardi	*University of Trento*
Prof. Frederic Bechet	*Aix Marseille University*
Prof. Anna Esposito	*Seconda Università di Napoli*

April 2017

Abstract

The study of *human interaction dynamics* has been at the center for multiple research disciplines including computer and social sciences, conversational analysis and psychology, for over decades. Recent interest has been shown with the aim of designing computational models to improve human-machine interaction system as well as support humans in their decision-making process. *Turn-taking* is one of the key aspects of conversational dynamics in dyadic conversations and is an integral part of human-human, and human-machine interaction systems. It is used for discourse organization of a conversation by means of explicit phrasing, intonation, and pausing, and it involves intricate timing. In verbal (e.g., telephone) conversation, the turn transitions are facilitated by inter- and intra- speaker silences and overlaps. In early research of turn-taking in the speech community, the studies include durational aspects of turns, cues for turn yielding intention and lastly designing turn transition modeling for spoken dialog agents. Compared to the studies of turn transitions very few works have been done for classifying overlap discourse, especially the *competitive act* of overlaps and function of *silences*.

Given the limitations of the current state-of-the-art, this dissertation focuses on two aspects of conversational dynamics: 1) design *automated computational models for analyzing turn-taking behavior* in a dyadic conversation, 2) predict the outcome of the conversations, i.e., *observed user satisfaction*, using turn-taking descriptors, and later these two aspects are used to design a conversational profile for each speaker using turn-taking behavior and the outcome of the conversations. The analysis, experiments, and evaluation has been done on a large dataset of Italian call-center spoken conversations where customers and agents are engaged in real problem-solving tasks.

Towards solving our research goal, the challenges include automatically segmenting and aligning speakers' channel from the speech signal, identifying and labeling the turn-types and its functional aspects. The task becomes more challenging due to the presence of overlapping speech. To model turn-taking behavior, the intention behind these overlapping turns needed to be considered. However, among all, the most critical question is how to model observed user satisfaction in a dyadic conversation and what properties of turn-taking behavior can be used to represent and predict the outcome.

Thus, the *computational models for analyzing turn-taking dynamics*, in this dissertation includes automatic segmenting and labeling turn types, categorization of competitive *vs* non-competitive overlaps, functions of silences (e.g., lapse, pauses) and turns in terms of dialog acts.

The novel contributions of the work presented here are to

1. design of a fully automated turn segmentation and labeling (e.g., agent vs customer's turn, lapse within the speaker, and overlap) system.
2. the design of annotation guidelines for segmenting and annotating the speech overlaps with the competitive and non-competitive labels.
3. demonstrate how different channels of information such as acoustic, linguistic, and psycholinguistic feature sets perform in the classification of competitive vs non-competitive overlaps.

4. study the role of speakers and context (i.e., agents' and customers' speech) for conveying the information of competitiveness for each individual feature set and their combinations.
5. investigate the function of long silences towards the information flow in a dyadic conversation.

The extracted turn-taking cues are then used to *automatically predict the outcome of the conversation*, which is modeled from continuous manifestations of emotion. The contributions include

1. modeling the state of the observed user satisfaction in terms of the final emotional manifestation of the customer (i.e., user).
2. analysis and modeling turn-taking properties to display how each turn type influence the user satisfaction.
3. study of how turn-taking behavior changes within each emotional state.

Based on the studies conducted in this work, it is demonstrated that turn-taking behavior, specially *competitiveness of overlaps*, is more than just an organizational tool in daily human interactions. It represents the beneficial information and contains the power to predict the outcome of the conversation in terms of satisfaction *vs* not-satisfaction. Combining the turn-taking behavior and the outcome of the conversation, the final and resultant goal is to design a conversational profile for each speaker. Such profiled information not only facilitate domain experts but also would be useful to the call center agent in real time.

These systems are fully automated and no human intervention is required. The findings are potentially relevant to the research of overlapping speech and automatic analysis of human-human and human-machine interactions. At the same time, the work opens up a new perspective on functions of silence towards information flow.

Keywords
Turn-Taking Dynamics, Behavioral Cues, Overlap Discourse, Competitiveness, Silences, Vocal-Nonverbal Cues, Computational Models

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Prof. Giuseppe Riccardi, who has supported and guided me throughout my Ph.D. with his patience and knowledge.

I would like to thank ICT doctoral school for their supports in many different ways especially to Andrea and Francesca. I am also thankful to the secretaries of our lab - Katalin, Helena, Anna, Carolina, and Piera for supporting and aiding in our regular activities. Also, I would like to thank Veronica for taking care of all the misuse of the servers and being patience all the time.

Furthermore, I would like to thank Morena Danieli, for her understanding and cooperative behavior. Without her guidance and persistent help, this dissertation would not have been possible. I am also grateful to Evgeny A. Stepanov for all the encouragement and unconditional help while listening to all my weird ideas. During my Ph.D., sometimes I blindly believed that he has an answer to all my questions.

A very special gratitude goes out to all my colleagues to make this journey memorable. With a special mention to Orkan, for teaching me about life and share his knowledge; Carmelo and Arindam, for always cheering me up; Michael, for being a good friend; Juan and Fabio for making my research journey interesting and a special thanks to Firoj, for helping me throughout this journey and shaping my research ideas to have more impact and influence in the scientific society.

I am also grateful to all my friends and Trento-family, specially Murtoza, for being part of this journey and helping me in many different ways. I have been having a great social life in Trento, thanks to all members of the BDUnitn family, whom I spent time and shared experiences. My special thanks go to Swapna and Labu for all their help, respect, love and hospitality.

These four and half year journey was not always a smooth ride. During this journey, there were many emotional ups and downs, uncertainty and doubt. But it is Firoj whose strong believe in me and support helped me to tackle these imbalances. No matter the situation, he always made me felt safe, loved and encourage me to stay always positive. I would also like to thank my in-laws, for understanding my absence from family holidays and always encouraging me. I would also like to thank my eternal cheerleader, my handsome little brother, for always wanting the best in me and for me. You make me a better person.

Finally, I would like to thank my parents, for supporting and believing in me. I am grateful to them for making me who I am today, for all their sacrifices, love and encouragement. You both are my role models and thanks for inspiring my dream and giving me the courage to be independent. I always wanted to put Dr. and Prof. in front of my name, similar to my parents, so my heartfelt thanks to my mother and father for helping me to walk towards that dream. The debt that I have to owe you both is unmeasurable. At this point, I owe you my deepest gratitude and love for your dedication and support.

To my parents
Inun Nahar Sultana and *Nurul Absar Chowdhury*

Everything I have accomplish is the result of your courage, compassion, love and sacrifice. You have given me all the tools I could ever need in life and made me who I am today.

To my little and awesome brother
Towhid Absar Chowdhury

For always pushing me to be my best, making my life special and fulfilling it with humor, happiness, love and for helping me to see the world in different angle.

To my best friend and beloved husband
Firoj Alam

For making me feel the most special person in the world; for holding my hands during all the ups and downs of the life; for all encouragement, understanding, love and also for being patience with me during all the deadlines and mostly for providing me the shoulders to cry on.

You all mean the world to me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The interpretation and the understanding of human interaction is one of the greatest scientific challenges. Even though humans learn to interact from an early stage of their life, yet the complex mechanisms that go underneath are still a big mystery for the researchers from different fields. Social interactions do not only convey meaning but also indicate speakers' intentions, expression of emotions, empathy, politeness, and even dominance relationships between the participants. Even after decades of research, the study of *human interaction dynamics* is still the main focus in multiple research disciplines including computer and social sciences, conversational analysis and psychology.

The traditional approach to studying human interaction is observation. This may be done either directly, on the basis of standardized observation protocols, or indirectly, through the collection of audio or audio-visual recordings of social interactions. The recordings are then manually observed and coded with different characteristics such as laughing, interrupting, coordinating, acknowledging, enthusiastic, friendly and polite. The resultant coding is later used in quantitatively measuring and preparing a concise summary out of it. The methodology based on observations depends on human perception and often is done by trained experts thus making the process time-consuming and expensive, as suggested by [2]. The authors report that human experts can only analyze less than 1% of the data in call centers. Apart from being time consuming and labor intensive, the resulted coded data may also have a high degree of variability due to the inter-annotator differences based on perception and human errors. Thus the process of coding human behavior needs automated systems that can 1) detect and analyze low-level cues to process how the conversational behavior is unfolding over time, 2) model the dynamics of the conversation, while 3) understanding intentions of the interlocutors and predicting/summarizing the outcomes.

To provide the machine with such abilities, interdisciplinary research domains such as *behavioral signal processing*, (*BSP*), have emerged. The main focus of the BSP is to design computational methods that model human behavior using behavioral signals [3]. These behavioral signals are manifested using overt and covert cues, which are processed and used by humans explicitly or implicitly, and often time act as a fundamental information carrier in facilitating human analysis and decision making.

Though the behavioral signals differ based on the type of social interactions, however, the information it carries is enormous and informative. Examples of social interaction, includes synchronous/ asynchronous interaction in social media, dyadic phone conversations, face-to-face (e.g., meeting) and face-to-machine (e.g., video-blog). In any interactional scenario, we express our behavior in terms of overt and covert cues. The *covert* cues include physiological signals such as heart-rate, respiratory activity and electrodermal activity. Moreover, the *overt* cues include different verbal expressions such as linguistic phenomena, and non-verbal, e.g., as paralinguistic information, facial expressions, gestures and postures, which are displayed, expressed and observable.

Among all the social interactions, the dyadic conversation is one of the most common real-life inter- personal interaction scenarios. Over the years in various research domains, the dyadic conversation is a core unit of analysis in understanding detailed interaction dynamics. Studies on dyadic conversations can be divided into different domains, based on 1) *mode of communication*, 2) *relationship strength of the interlocutors* and 3) *the style of communication*. The mode of communication, in daily life interaction, includes face-to-face, telephone conversations, emails or even classroom lectures, which dictates how the interlocutors are exchanging information. Relationship strength of the interlocutors also varies, from the personal relationship (husband-wife, siblings, friends, colleagues) to even between two complete strangers. As for the communication styles: a conversation can be casual, e.g., between friends or family; to task oriented for example job interviews or call center inbound calls (as shown in Example 1); or it can involve atypical interactions like therapist-patient. Figure 1.1 shows an example of interaction dynamics in a dyad conversation.

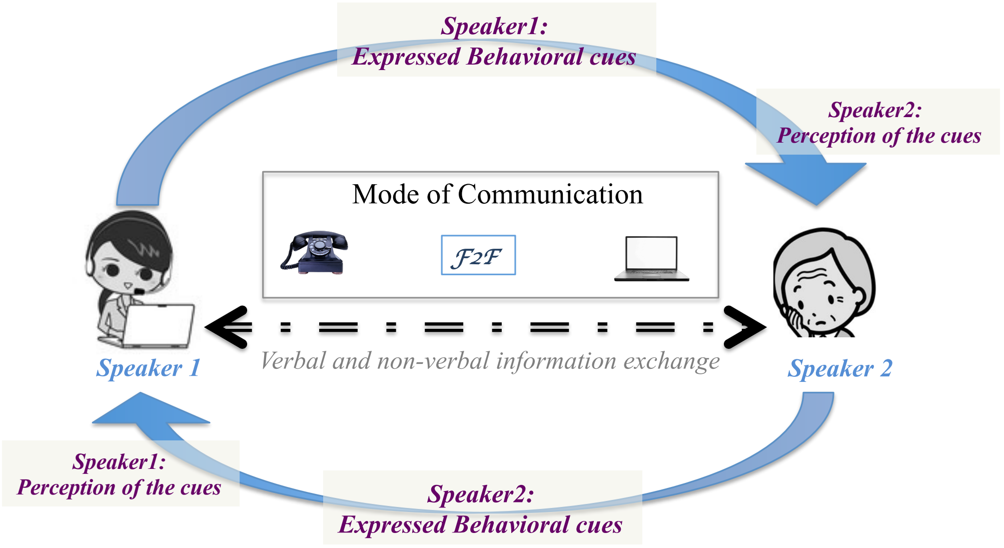


Figure 1.1: Interaction Dynamics in a Dyadic conversation.

Example 1. Example of a Dyadic conversation in call center scenario, where speech overlaps are presented between [and], and (.) represents pauses with length < 1 second[1].

caller: allora quest utenza stamattina è stata sospesa
caller: this morning the service was interrupted
(1.09)
operator: sì
operator: yes
(1.74)
caller: ecco la motivazione cortesemente
caller: please tell me why
(3.7)
operator: allora sì (.) vedo che c è una riduzione della potenza in seguito a una serie di fatture non pagate che
operator: yes (.) (I) see that there is a reduction of the power due to a number of unpaied bill that ...
operator: [fanno un importo ah]
operator: [amount to ...]
caller: [è stata sì è stata disattivata]
caller: [this morning it was suspended]
caller: non è stata ridotta la potenza
caller: it was not a power reduction
(1.1)
operator: ah sì diciamo (.) c è stata a
operator: oh yes let us say (.) it was to
operator: noi risulta lo stato ridotto però probabilmente si è già proceduto al distacco completo
operator: (as far as) we can see the power status was reduced but pherhaps they already interrupted (it) completely
caller: al distacco perfetto ora eh eh su che base mi perdoni
caller: the complete interruption ... perfect! now ehm ehm due to what reason, excuse me?
(1.14)
operator: ah ascolti qui ci sono una serie di fatture malgrado
operator: Listen (please) we have here a number of unpaid bill in spite of
operator: [ci]
operator: [(in spite of) there is]
caller: [mh beni]
caller: [mhm well]
operator: sia il blocco per sisma vedo che c è in <LOCATION> per
operator: the block due to the earthquake I see that there is in <LOCATION>
operator: [il sis]
operator: [the earth-]
caller: [no no no no questo è]
caller: [no no this is]
caller: un blocco che avete un problema voi tra uffici
caller: a block (due to) a problem you have within your (administrative) departments

1.1 Turn-Taking Dynamics

Turn-taking is one of the key aspects of conversational dynamics and is an integral part of human-human and human-machine interaction systems. In everyday interaction, spoken

[1] All the name entities are removed from the example

conversations ordinarily unfold following the norm that each speaker should take a turn of the conversation while coordinating with one another and almost constraining the floor to one party at a time. The practice of taking turns is intuitively familiar and contains many behavioral cues. Each turn a speaker takes is designed to convey something. The design, placing and timing of the turns, in a conversation indicates how the behavior manifestation is motivated by the intention and coordination of the speakers. Thus making turn-taking an important and complex characteristics of conversational dynamics.

The puzzle behind the turn-taking is highlighted in the context of ordinary daily conversation, which lacks a prearranged format for taking turns. The formalization of turn-taking raises many questions which research communities seek to address:

- *How long a speaker will retain the conversational floor?*
- *Is there going to be a speaker change?*
- *If so who is going to be the next speaker and how the local management in the conversation is actually done?* and
- *What was the motive behind the design of each turn and how it is acquired?*

The study of the signals of turn-taking started with Sacks et al. [4] devoting considerable attention to the phenomenon of turn changes, including how the next speaker is selected, in conversations. Theoretically, there are three possible ways of organizing a turn change, i.e., turn-taking signals, which includes a) with no gap and no overlap b) with gap (silence in between), and c) with overlapping speech, as shown in Figure 1.2.

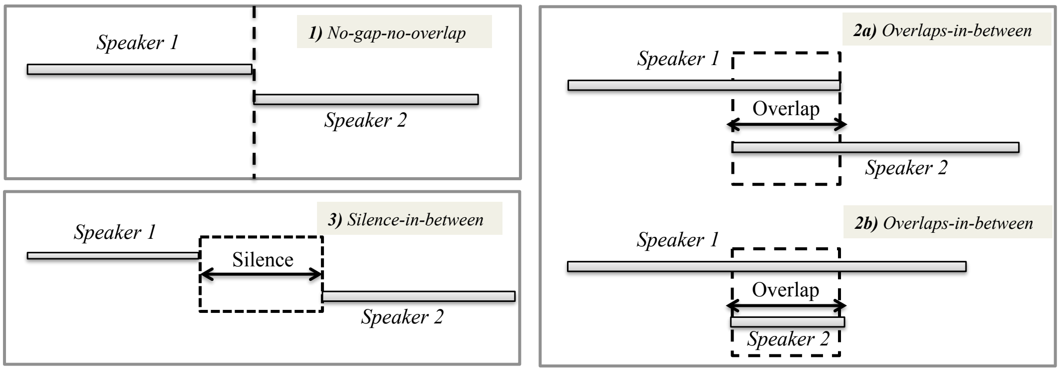


Figure 1.2: Types of turn-taking signals where 1) shows turn changes with "no-gap-no-overlap"; 2) represents turn changes with overlapping speech (2a), and the sudden insertion of another turn in the same conversational flow (2b), and 3) shows turn changes with silence in between.

However, by analyzing substantial conversational data Sacks et al. [4] had observed that the most common case in conversation is *"one-party-at-a-time"*, and that the turn changes mostly occur without any gaps and any overlap (i.e. *"no-gap-no-overlap"* in Figure 1.2). This finding led the authors initially proposed – *the projection theory* – which suggests that the next speaker can anticipates the end of current speaker's turn based on structural and contextual information, and then starts talking at the projected turn-ending. This concept of no-gap-no-overlap has been a inspiration and a ground work of many researches in turn-taking and in conversational analysis community, dedicating their studies over series of rules of conversations.

However, findings in many studies, such as [5], provides enough evidence that the timing of turn-taking is not as precise as it is often claimed. Thus speaker changes are not strictly no-gap–no-overlap and does not follow one-speaker-at-a-time rule. This indicate that in human interaction the concept of minimization of gaps and overlaps are not always aimed and can be used as a tool to express the pragmatic function behind it.

In initial studies of human conversation, *overlapping speech* is considered as a violation of the fundamental rule of turn-taking. However, in a natural spoken conversation, overlapping speech is a universal phenomenon. Studies such as [6] suggests the presence of 44% and 52% of overlaps in face-to-face and in telephone dialogs, respectively, indicating that overlapping is pervasive in human conversations.

Overlapping speech, apart from being a turn-taking signal (speaker change), can signal the speaker's intent behind the overlaps. For example, few studies have proposed that speech overlaps are related to dominance or aggression towards the other speaker [7]. However, the picture is more complex. Not all the overlapping occurrences are related to aggression or conflicts. They can also be cooperative in the conversations, by providing the other speaker with cues about the mutual understanding and supports [8]. Thus distinguishing the overlaps by the intention behind the overlaps and perception by the current speaker is an important behavioral signal for modeling human turn-taking behavior of the conversation. In the computational literature, over the years a widely accepted categorization of overlaps discourse is: **Competitive (Cmp)**, *an attempt to grab the floor*, and **Non-Competitive (Ncm)**, *an attempt to assist the speaker for the continuation of the current turn*.

On the other hand, turn changes with *silence* (*gap*) is the most frequent turn-taking signal. It is also found in many literatures that this type of turn-changing signals is more desirable, as minimizing gaps can risk overlapping speech. On the contrary to gaps, *Lapse between speakers* (long gap) may signal signs of trouble, for example, it can indicate that the upcoming turn will be dis-preferred or disagreeing [9]. Unlike gap and lapse, the concept of *pause* in [4] is not directly associated with turn changes, even if it occurs at *TRP*[2].

Silence has always been characterized in different forms depending on their relation be-

[2] Transition Relevance Place (TRP): are the points change of speakership becomes a salient possibility, whether it is realized or not.

tween speech and language. In addition to pause, gaps and lapse, some other types of silences that can be found in literature are, *stillness*, when it is speaker's turn, the listener listens remaining silence and *eloquent silence*, that includes intentional silence, e.g., showing no interest to reply, ironic silence, grammatical silence, etc. Even though it is most common phenomena in a conversation, but most of the studies associate silence with powerlessness, the death of turn, break in conversational flow and negativity. It was also treated as absence of: speech, meaning and intention [10–12]. At the same time silence has also been reported as sign of power, and politeness depending on the context and the culture [13, 14].

From speech community perspective a considerable work has been done in understanding turn-taking dynamics. The research includes the study of durational aspects of turns, cues for turn yielding intention and lastly for designing turn transition models for spoken dialog agents. Compared to turn transitions very few studies have been done on overlapping speech, especially for classifying *competitive act* of overlapping speech. Studies on overlap include the length, position, and timing of overlaps along with its prosodic and temporal properties. As for silence, in the speech community, most of the studies are in line of the distribution of gaps and pauses along with its durational aspects, with the aim to define the systems' response interval, i.e., waiting time, for spoken dialog systems.

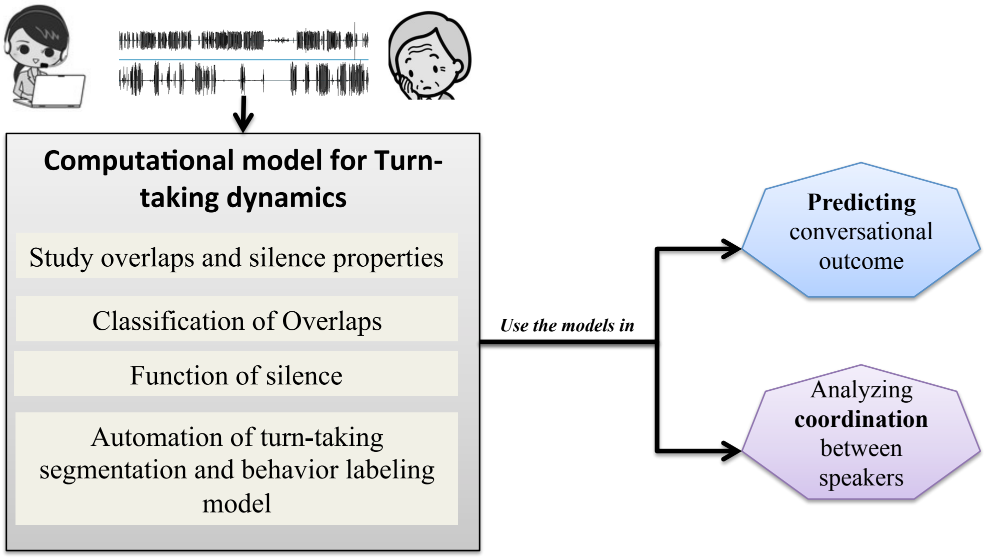

Computational model for Turn-taking dynamics
Study overlaps and silence properties
Classification of Overlaps
Function of silence
Automation of turn-taking segmentation and behavior labeling model
Use the models in
Predicting conversational outcome
Analyzing coordination between speakers

Figure 1.3: Key focus of the dissertation.

Given the limitation in the state-of-the-art of overlap discourse and silence function, this dissertation mainly focused on the following topic, as also depicted in Figure 1.3:

1. **Different aspects and design of Turn Taking Models**

   (a) *Studying the properties of overlaps and silences in turn-taking dynamics.*
   (b) *Designing the computation models for overlap discourse classification and functions of silence.*
   (c) *Automation of segmentation and labeling of turn-taking dynamics* in a spoken dyad conversation.

2. **Studying The Role of Turn Taking Dynamics**:

   (a) **Prediction of the Conversational Outcome:**
   One of the desired outcomes from the conversations is to have a satisfying communication. Over the years, the study of user satisfaction dependent on using spoken or written questionnaires and interviews. In such an evaluation, users are usually asked to fill up questionnaires or rate certain aspects of a conversation that address users' satisfaction, as reported in [15]. Due to the important role it plays in understanding social interactions, user satisfaction has been addressed in different research fields, e.g., Spoken Dialog Systems (SDS), as well in other marketing and designing fields. In SDS, such as problem-solving [16] and tutoring [17], user satisfaction is used as one of the metrics to assess the quality of a dialog system [18, 19]. Thus, the increasing importance of user experience as a quality assessment demands a computational model for observed user satisfaction rather than self-reported measure. To understand the role of turn-taking behavior in predicting observed user satisfaction, *this dissertation has also made a novel contribution to model the state of the observed user satisfaction regarding the final emotional manifestation of the users in an ongoing conversation.*
   (b) **Analyzing Coordination between Interlocutors**
   In a dyadic conversation, one of the important challenges is to understand how different behavioral cues are associated with one another and how we express them in different interaction scenarios. In this thesis, *we investigated, the association of conversational turn-taking behavior with coordination of interlocutors in different emotional manifestations of the speakers*. For this study, in [20], the conversational coordination between the interlocutors is defined as the tendency of speakers to predict and adjust each other accordingly on an ongoing conversation.

The study of designing automatic computational models poses many challenges (see Section 1.2), and at the same time, the outcome of these models has many application sectors such as customer care, education, and healthcare.

1.2 Research Challenges and Addressed Issues

Modeling human interaction is itself a complex task, but its complexity increases even more because the human behavioral patterns and the extracted signals depends on the heterogeneity and variability of mode and style of communication along with the relationship strength and cultural difference. Hence designing a *'universally useful computational system'* a very challenging task. Therefore, the current state-of-art focuses on designing a domain or application specific system(s) with a goal of a specific behavioral aspect in mind.

From designing perspective, a complete pipeline of a computational model includes designing an experimental scenario, data collection focusing on capturing expressed cues in the forms of audio/video/ physiological recordings, and then extracting the patterns to design the model using machine learning algorithms. There are several challenges in each step of this pipeline, shown in Figure 1.4.

At first, setting up an experiment to collect ecologically valid[3] data and obtaining a representative number of samples is a major problem and often impossible to get. Then, the collected data needs to annotate by experts or crowds or unsupervised approach with the predefined turn-taking discourse labels and rules. For the annotation, a concrete guideline with respects to labels' pragmatic functions, speaker intentions among others is needed. A consideration to the agreement of the annotator is also needed in identifying the segment and its associated label due to the inherent nature of subjectivity and variability of each annotator.

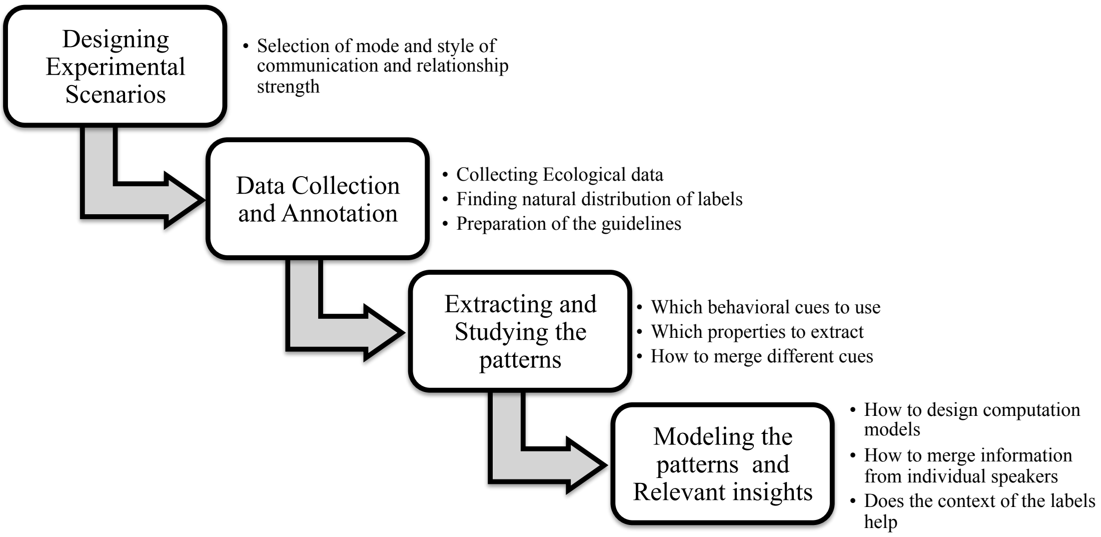

Designing Experimental Scenarios
• Selection of mode and style of communication and relationship strength
Data Collection and Annotation
• Collecting Ecological data
• Finding natural distribution of labels
• Preparation of the guidelines
Extracting and Studying the patterns
• Which behavioral cues to use
• Which properties to extract
• How to merge different cues
Modeling the patterns and Relevant insights
• How to design computation models
• How to merge information from individual speakers
• Does the context of the labels help

Figure 1.4: General processes of investigating human interactions.

[3] Ecological validity often refers to the relation between real-world phenomena and the investigation of these phenomena in experimental contexts [21].

Extracting the right kind of behavioral signals is also very challenging. We express behaviors by overt and covert cues and if only overt cues are considered it also has many channels such as audio, and visual. Considering only one channel makes the computational task a difficult problem. In many interaction scenarios only spoken channel is used, such as telephone conversations. Even though we know which behavioral cues to use, we have to analyze what properties to be considered. Hence, extensive investigations are necessary to deal with such scenarios. In a dyadic conversation, as soon as two people are engaged in a conversation, their internal behavioral states are coupled and become mutually dependent [22]. This coordination is reflected in their turn-taking mechanism, which adds another layer of complexity to the original task. Thus the challenge in modeling turn-taking in human interaction also requires the design of computational models that can capture the interaction dynamics between interlocutors in various information channels. After that, challenges remain to the design of computational models.

One of the important challenge faced in the thesis, which has been hardly addressed in the computational literature, is how to design features for studying silence (*between-speaker and within-speaker*) in the flow of conversation along with how to automatically segment and label turn-taking discourse. There have been studies where thresholds for silence to bridge the turn-constructional units (TCU) has been studied, but all the studies depend on manual turn segmentation.

Therefore, designing computational models involves the following challenges:

- Annotation of ecologically valid data with real behavioral expressions requires an operational definition and guidelines. For example, there has not been any operational definition for annotating and modeling *competitiveness of the overlaps* and *function of long silence* for the call-center scenario.
- In any ecologically valid dataset, natural distribution of the class throws an important challenge for the performance of the computational model.
- Automatically generating the turn-taking sequence poses different challenges such as segmentation (and alignment) of spoken conversations, assigning thresholds and labels to the corresponding segments.
- Since any behavioral construct such as competitiveness is manifested using different verbal and vocal non-verbal cues. Therefore, it is necessary to investigate each linguistic and acoustic information independently and in combination. This requires the investigation of various low-level features, combination strategies at the feature- and decision- level.
- In a dyad conversation, as the dynamics of the turn-taking is mutually dependent on the interlocutors, therefore it is essential to investigate the role of each interlocutor for each

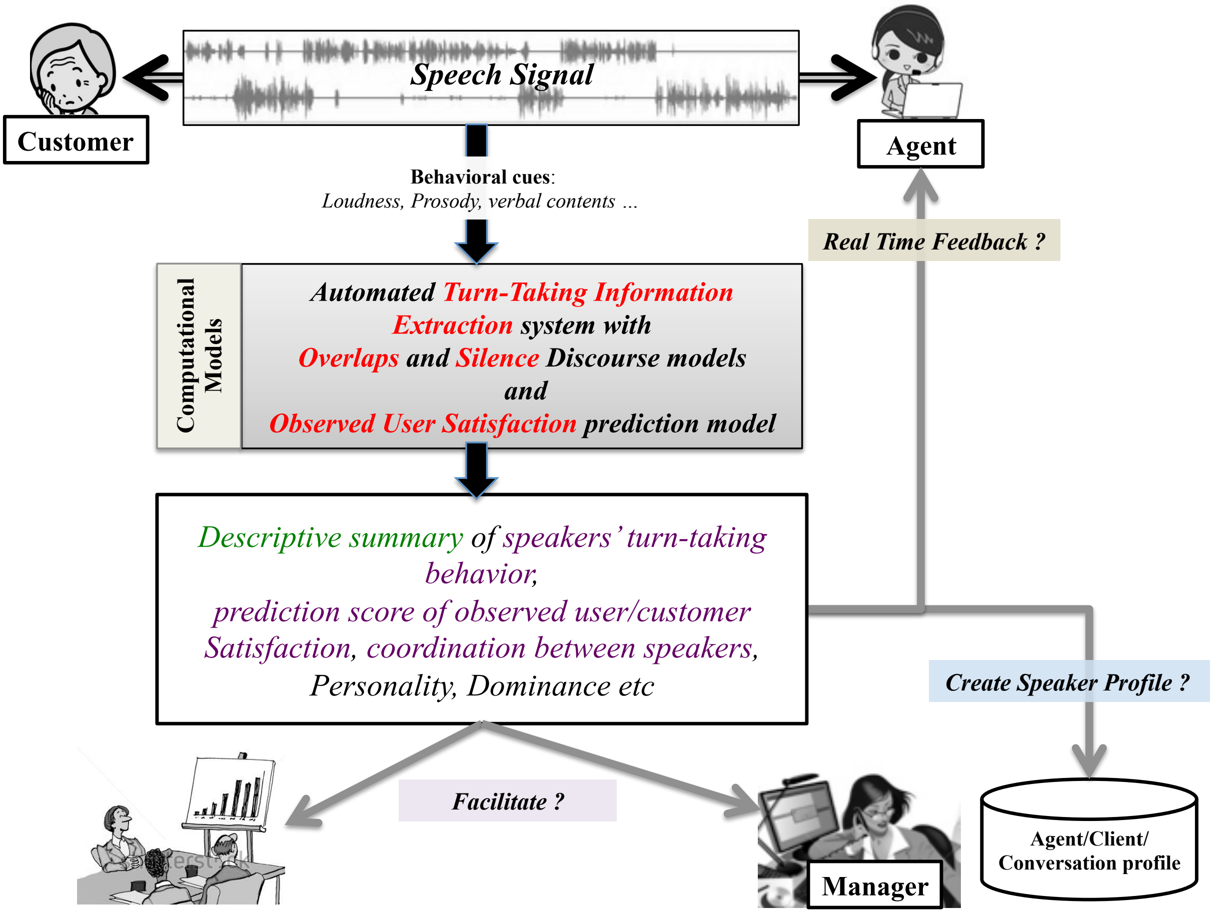

Speech Signal
Customer
Agent
Behavioral cues:
Loudness, Prosody, verbal contents ...
Real Time Feedback ?
Computational Models
Automated Turn-Taking Information Extraction system with Overlaps and Silence Discourse models and Observed User Satisfaction prediction model
Descriptive summary of speakers' turn-taking behavior, prediction score of observed user/customer Satisfaction, coordination between speakers, Personality, Dominance etc
Create Speaker Profile ?
Facilitate ?
Manager
Agent/Client/ Conversation profile

Figure 1.5: An application scenarios for the call centers.

discourse segments and the importance of context for the classification. This also requires study of how to combine the interlocutor's information for the design of computational models.

- To use automated turn-taking information in predicting the outcome of the conversation, the design of turn-taking behavioral features is crucial.
- Designing unsupervised annotation of the outcome of the conversation (i.e., observed user satisfaction) is an important challenge towards the automatic prediction model.
- The way of human interaction differs in different communicative scenarios such as human-human, human-machine. Hence, it is necessary to investigate the capability of the automatic system in different scenarios.
- The design of a complete automated pipeline where no human intervention is required.

So while addressing the above challenges, in modeling turn-taking dynamics the research question that we aimed to answer in this dissertation is

Can we automatically identify and label the turn types
while categorizing the competitiveness in overlaps and functions of long silences?
If so, what behavioral – verbal and vocal non-verbal cues can we use?

To understand what roles turn-taking dynamics plays towards the outcome and coordination of the conversation, the research question we addressed is

How do we model observed user satisfaction (as an outcome)
of a conversation without any human intervention?
Moreover, can we use turn-taking dynamics to predict this outcome
and learn speakers' coordination inside an emotional episode?

While figuring out the answer to this research question we had an application scenario in mind as depicted in Figure 1.5. In this application scenario, agent and customer are interacting in a call center, and the idea is to automatically analyze the turn-taking behavior (competitiveness in overlapping speech, the function of each turn and silences, etc.) of the conversation while predicting the outcome of the conversation. The system then prepare a descriptive summary using the information of the predicted behaviors. The descriptive summary can facilitate domain experts such as call center managers, decision makers and it can also help the agent in real time while accumulating and updating the profile of the speakers.

Apart for the above scenarios, the computational models of turn-taking behaviors have many other application sectors. From tutoring robots [23, 24] to designing the game for children with autism [25] along with designing and improving the naturalness of other spoken dialog agents [26, 27].

1.3 Terminology

This Section highlight the terminology and concepts that are relevant for the dissertation.

Behavior: It is defined as "... quite broadly to include anything an individual does when interacting with the physical environment, including crying, speaking, listening, running, jumping, shifting attention, and even thinking." [28].

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

Turn: A turn is a time during which a single participant speaks.

Turn-taking: Turn-taking is the principal unit of description in conversational structure. It is used for discourse organization of a conversation by means of explicit phrasing, intonation, and pausing, and it involves intricate timing.

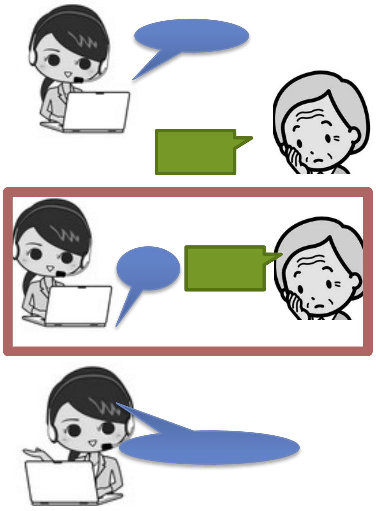


Figure 1.6: A example of overlapping scenario.

Overlapping speech: Overlapping speech is a conversational phenomena where the more than one person is holding the conversation floor (i.e., talking simultaneously) at the same time in the same conversation, as shown Figure 1.6. The alternative terms for overlapping speeches are: double talking and (negative) response times [29], double talk and interruptions [30], simultaneous speech [31], (negative) switch time or switch overlaps [32], and (negative) floor transfer offsets [33]. The term used to call incoming speaker who initiated the overlapping speech is overlapper and the current speaker holding the floor is overlappee through out this dissertation.

Competitive overlap: Competitive overlap represent the pragmatic function of overlaps where the intervening speaker starts prior to the completion of the current speaker while attempting to display interest in the turn even though the current speaker is eager to keep the turn for themselves, and thus both speakers perceive the overlap as problematic. An example of competitive overlaps can be seen in Example 2.

Example 2. Example of Competitive Overlap

```
operator: allora sì (.) vedo che c è una riduzione della potenza in
     seguito a una serie di fatture non pagate che
operator: yes (.) (I) see that there is a reduction of the power due to a number of unpaied bill that
     ...
operator:   [fanno un importo ah ]
operator:  [amount to ...]
caller:   [è stata sì è stata disattivata ]
caller:   [this morning it was suspended]
caller: non è stata ridotta la potenza
caller: it was not a power reduction
```

Non-competitive overlap: Non-competitve overlap represents pragmatic scenario where another speaker starts in the middle of an ongoing turn, and shows no evidence for grabbing the turn for themselves. The intervening speaker use it to signal the support for the current speaker's continuation of speech and both speakers perceive the overlap as non-problematic event. Example of non-competitive overlaps can be seen in Example 3.

Example 3. Example of Non-Competitive Overlap

```
operator: ah ascolti qui ci sono una serie di fatture malgrado
operator: Listen (please) we have here a number of unpaied bill in spite of
operator:   [ci]
operator:   [(in spite of) there is ]
caller:   [mh beni ]
caller:   [mhm well]
operator: sia il blocco per sisma vedo che c è in <LOCATION> per
operator: the block due to the earthquake I see that there is in <LOCATION>
```

Pause: Silences within the same speaker turn. There are very few terminology associated with pause in the literature. They are resumption times [29] and within-speaker silence/pauses. For the dissertation, pause and within-speaker silence are used synonymously.

Lapse-within: Longer (or extended) silences within same speaker turns.

Gap: Short silences at TRP, when current speaker completed his turn without selecting the next speaker and before anyone select themselves as a new speaker, a brief presence of silence. Terminology used to describe this event in conversation are between/inter speaker/turn silences/intervals. Other terminolgy used to describe the similar events include (positive) response times [29], switching pauses [31], (positive) switch time or switch pauses [32], transition pauses [34], alternation silences [30], (positive) floor transfer offsets [33], or just silent or unfilled pauses [35, 36].

Lapse-between: Longer (or extended) silences between turns.

Coordination: The conversational coordination between the interlocutors (speakers) is defined as the tendency of speakers to predict and adjust each other accordingly on an ongoing conversation.

Emotional Episodes/State: The state of an individual's emotions. An emotional state is a product of the psychological and physiological processes that generate an emotional response, and that contextualize, regulate, or otherwise alter such responses [37].

Observed User Satisfaction: This dissertation defines observed satisfaction of a user/speake as the final manifestation of emotional states of the speaker in a conversation. Depending on the states of the final emotion, the observed user satisfaction is categorized into three labels as **Positive** (Pos), **Negative** (Neg), and **Neutral** (Neu).

1.4 Thesis Contributions

The primary focus of this dissertation is to design computational models for turn-taking dynamics using expressed behavioral (overt) cues by investigating a large ecologically valid real call center data. Following which, this dissertation focus on understanding its role by predicting the outcome of the conversation and by analyzing the speakers' coordination in emotional episodes inside the conversation. In the following subsections, a brief description of the main contributions of the thesis is discussed.

1.4.1 Unsupervised Study of Turn-Takings

Understanding turn taking is a critical element in human conversation process analysis, and it varies a lot depending on the mode of communication and style. While most of the previous studies have focused on meeting corpora or other small datasets to study and model turn-takings, this work is concentrated on a *large* dataset of *ecologically valid, inbound Italian call-center spoken conversations where customers and agents are engaged in real problem-solving tasks.*

In this dissertation, we investigate these phone conversations for different characteristics of turns especially for understanding *unsupervised properties of overlapping speech turns using low-level acoustic and lexical features.* Most of the previous studies relied on a small set of prosodic features. Whereas in this study *we investigate a large set of low-level features such as spectral, cepstral features among others, and their derivatives, projected onto statistical functionals, such as mean and range,* which is also a *novelty* in modeling overlaps. The purpose of the study is to understand if there are any visible distinctions between the overlaps in our dataset and if these low-level features could represent properties of these overlaps. While studying the overlapping turns, we also investigated the *duration distribution of silences present in the dataset for understanding the response interval for the automatic system.*

1.4.2 Annotation and Overlapping Speech Classification

Even though overlapping speech is a violation of one-speaker-at-a-time rule, however, it is one of the most common phenomena in spoken conversation. So to understand behavioral manifestations, which unfolds in turn-taking, we needed an operational model to categorize the competitiveness in overlapping speech. However, to model such discourses, we needed a novel overlap annotation guideline.

One of the main and *novel* contributions of the dissertation is to *design of an annotation guideline for segmenting and annotating the speech overlaps with the competitive and non-competitive labels* in a typical call center conversation in a continuous time scale with information from speech signals only. The annotation guideline includes functional rules, which can be transferred to any domain any time.

Most of the effort of the dissertation is directed in modeling overlapping speech classification models. The research first focused on features: the low-level acoustic features, as mentioned in Section 1.4.1, to evaluate the distinguishing capabilities of the features while categorizing competitiveness in overlaps while incorporating both the interlocutors' channel information. Another *novelty* of the dissertation is *the use of lexical and psycholinguistic features* in overlap classification task. This dissertation also includes a *novel* study of *the role of individual speakers with and without context* in providing information regarding the overlap discourse. In terms of model design, this thesis *contributed* to the designs of both *linear and non-linear computational models*. In addition, the thesis also studied different *feature- and decision- combination techniques* and its impact on the performance of the classifications.

1.4.3 Functions of Long Silence

The occurrence of silence inside a spoken conversation is the most natural phenomena. Over the years speech and computer science communities have been studying silence to identify the response interval of the spoken dialog systems.

One of the *novel* contributions of this dissertation is to shed lights on *the function of long silence towards the information flow of a conversation*. The study includes identifying the long silence instances, designing features for the categorization of the functions of between- and within- speaker silences using a hierarchical concept learning[4] technique. In order to better understand and obtain general functional categories from hierarchical tree we selected and merged the clusters (i.e., sub-trees) based on their functional similarity. We have done this selection and merging process with human supervision.

1.4.4 Automatic Turn-Taking Segmentation and Labeling

Another contribution of this dissertation is to design an automated system that can automatically segment and label turns along with the discourse. The system can take only speech signal, in the case of absence of transcription, as input to model the turn-taking discourse. This task has been achieved with the help of the computational models discussed in Sections 1.4.2 and 1.4.3 along with the contributions to create state-of-the-art *human-human Automatic Speech Recognition (ASR) system*, *dialog-act segmentation and classification systems*, and other systems like speech-vs-nonspeech segmenter among others.

1.4.5 Predicting the Conversational Outcome

To understand the role of turn-taking behavior in predicting the outcome of the conversation, the dissertation contributed to *automatically predict observed user satisfaction as a measure of*

[4] As a concept learning technique, we used Cobweb clustering algorithm.

the conversational outcome. Unlike the traditional approach, e.g., self-reported satisfaction, this research model *observed user satisfaction as the final emotional manifestation of the conversation*, which can be either positive, negative or neutral. To design the computational model for prediction, the turn-taking features are engineered using the turn-taking model in Section 1.4.4. Thus the novelties in this part are *defining observed user satisfaction using the final emotional manifestation*, *engineering turn-taking features* and *designing a prediction model for the observed user satisfaction*.

1.4.6 Coordination between Interlocutors inside Emotional Episodes

To understand how different behavioral cues are associated with one another and how we express them in different interaction scenarios, this dissertation contributed to find the association of turn-taking dynamics with different emotional segments. Thus one of the many contributions is, *investigating the coordination of interlocutors behavior in different emotional segments and how conversational turn-taking dynamics are associated with emotional manifestations of the agent and customer*.

1.5 Publications Relevant to the Thesis

The following publications are relevant to this thesis, which are revised in the preparation of the thesis.

1. Shammur Absar Chowdhury and Giuseppe Riccardi, *A Deep Learning Approach to Modeling Competitiveness in Spoken Conversation*, in Proc. of ICASSP. IEEE, 2017, New Orleans, USA.

2. Firoj Alam, Shammur Absar Chowdhury, Morena Danieli, Giuseppe Riccardi, *How Interlocutors Coordinate with each other within Emotional Segments?*, COLING, Osaka, Japan, 2016.

3. Shammur Absar Chowdhury, Evgeny A. Stepanov and Giuseppe Riccardi, *Predicting User Satisfaction from Turn-Taking in Spoken Conversations*, in Proc. of Interspeech-2016, San Francisco, USA.

4. Shammur Absar Chowdhury, Evgeny A. Stepanov and Giuseppe Riccardi, *Transfer of Corpus-Specific Dialogue Act Annotation to ISO Standard: Is it worth it?*, in Proc. of 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož (Slovenia).

5. Giuseppe Riccardi, Evgeny A. Stepanov and Shammur Absar Chowdhury, *Discourse Connective Detection in Spoken Conversations*, in Proc. of ICASSP. IEEE, 2016, Shanghai, China.

6. E. A. Stepanov, B. Favre, F. Alam, S. A. Chowdhury, K. Singla, J. Trione, F. B'echet, G. Riccardi, *Automatic Summarization of Call-center Conversations, IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, Scottsdale, Arizona, USA, 2015.

7. Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi, *The Role of Speakers and Context in Classifying Competition in Overlapping Speech*, in Proc. of Interspeech-2015, Dresden, Germany.

8. Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi, *Annotating and Categorizing Competition in Overlap Speech,* in Proc. of ICASSP. IEEE, 2015, Brisbane, Australia.

9. Shammur Absar Chowdhury, Giuseppe Riccardi, Firoj Alam, *Unsupervised Recognition and Clustering of Speech Overlaps in Spoken Conversations,* Workshop on Speech, Language and Audio in Multimedia (SLAM 2014), Penang, Malaysia.

1.6 Structure of the Thesis

The research on spoken conversation analysis has been done over decades and is still in the center for multiple research disciplines, this dissertation starts with introducing the challenge and the underlining complexity of the spoken interaction.

A brief review of studies done on turn-taking, mainly in overlapping speech and silences, and its related research are then mentioned in Chapter 2.

In Chapter 3, the dissertation introduce the (**SISL**) Human-Human Dyadic Conversation corpus and the sub-corpus used for the experimental and analytical research during the thesis. The chapter also presents the detailed guidelines for the annotation of overlap discourse, evaluation matrices for the annotation quality and corpus analysis. Apart from SISL corpus, the chapter also focus on LUNA human-human Italian corpus and details of the dialog-act annotation of the corpus.

After introducing the corpus and the annotation of the discourse label, the dissertation investigates the properties of overlaps using unsupervised technique in Chapter 4, using low-level acoustic and lexical features. This study helps to investigate whether this low-level features has the power to distinguish the complex discourse of overlaps.

Following the unsupervised study, a detailed description of the design and the technique for modeling overlapping speech categorizer using supervised modeling approaches is presented in Chapter 5. This chapter also includes study of the role of individual speakers with and without context along with acoustic, linguistic and psycholinguistic features in providing information regarding the overlap discourse.

In Chapter 6, the dissertation focus on studying long silences (between- and within- speaker) to understand the function of long silence towards the information flow of the conversation. The chapter includes identifying long silence, clustering them using hierarchical concept learning technique, and selecting and merging them using human supervision for the final analysis.

In order to obtain a descriptive summary of the turn-taking dynamics for the whole conversation we desgined a complete pipeline using different models. In Chapter 7, this dissertation discusses the computational architecture, which shows different components to create aligned speaker channel information and turn segmentation. The system also include a module to create turn labels using heuristic rules, and a discourse classification model to automatically annotate turn-taking dynamics.

To understand the essentials of turn-taking behavior, Chapter 8 focuses on how this behavior can be used to predict the outcome of the conversation. The chapter describes the architecture to automatically predict the observed user satisfaction as a measure of the conversational outcome.

In Chapter 9 this dissertation discusses the contribution towards finding the coordination of the interlocutors behavior in different emotional segments using lexical, psycholinguistic and turn-taking features in terms of regression coefficients, cosine similarity and correlation analysis, respectively.

Then in Chapter 10, the dissertation is concluded with a brief summary of the work and with pointers to motivate future works.

Chapter 2

Relevant Studies

2.1 Fundamentals of Conversations

In a regular conversation, one speaker speaks at a time by taking turns. While small gaps and overlaps between speakers' speech are very frequent and rarely last more than a few milliseconds. This smooth interaction is one of the most essential elements in our daily conversation, which distinguishes it from the other mode of communication such as formal monologue (e.g., broadcast news) and written messages. There has been many research in different fields, however, there is a lack of agreement among researchers in order to define it.

2.1.1 Turn-Taking

For studying this natural phenomenon, Sacks [4] and his colleagues created a field called Conversation Analysis (CA) in 60s and early 70s. The main goal of the field was to study how people interact with each other in different social settings such as informal conversation, medical conversation, and interviews. Among all the aspects explored by CA researchers, they have identified turn-taking as one of the prominent components of a conversation.

The study of turn-taking started very early period of CA research. In [4], Sacks et al. present a characterization of turn-taking in conversations between two or more persons. The authors studied a large set of conversational data and stated a detailed description of fourteen "grossly apparent facts" about human conversation. Some of the facts are:

- "Speaker-change recurs, or at least occurs".
- "One party talks at a time".
- "Occurrences of more than one speaker at a time are common, but brief".
- "Transitions (from one turn to a next) with no gap and no overlap are common. Together with transitions characterized by a slight gap or slight overlap, they make up the vast majority of transitions".

Using the findings of the study, the authors then established a set of rules and constrains that they believed any model of turn-taking in conversation should obey.

The study also proposed a minimal model of turn-taking whose key components are a turn constructional component (*TCU*), which defines a turn. Turns are thus incrementally built out of a succession of turn-constructional units (TCUs as shown in Figure 2.1). A TCU can be made up of sentences, clauses, phrases, and/or individual words.

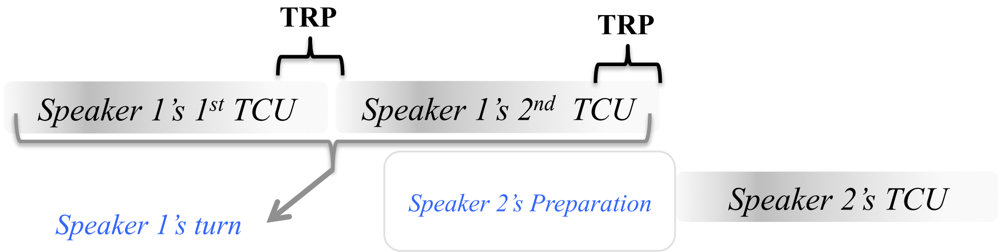


Figure 2.1: Association between TCU completion and transition- relevance, as well as the contingent nature of turn transfer.

2.1.2 Transition Relevance Place

Transition Relevance Places (*TRPs*) is a concept central to the turn-constructional component. TRPs are described as points in an utterance where it would be relevant for another participant to take the floor, as shown in Figure 2.1.

In [4], syntactic constituents' boundaries are considered as an indication to TRP but due to the presence of disfluencies, mentioned in [38], this consideration of syntactic boundaries is found to be problematic. Compared to the previous studies, a more effective definitions of syntactic TRPs are mentioned in [39, 40] . In [39], authors defined that syntactic TRPs are "potential terminal boundaries for a recoverable clause-so-far". Studies in [39, 40], also suggest that reactive tokens such as backchannels (e.g., "okay", "hmm"), assessments (e.g., "really?") and repetitions, in the form of acknowledgment or confirmation, are also examples of complete syntactic units regardless of not being well-formed syntactic constituents.

Apart from syntactic units, another well researched feature for describing TRP is its prosodic characteristics. In [41], the author proposes that the properties of TRPs are prosody, mainly the completion of a tone unit with a non-level nucleus and with a decrease in volume (loudness). A contrasting view with [41] is presented in different studies [39, 40, 42], where the authors considered prosody as markers of intonation completion points for turns containing questions or statements. Prosodic pattern such as lengthening of final word or syllable of a turn has been studied in [35, 43]. Many studies showed that prosodic features can be used to detect which syntactic unit can be considered as a signal of the end of the turn [44–46]. Like other fields, psycholinguists have also investigated prosody as a feature of TRP. In [47], the results shows that humans are able to react spontaneously, sometimes even anticipate beforehand, to the turn boundaries with only using prosodic information.

From the semantic and pragmatic perspective, the turn completion points correspond to the place where the turn constituents a complete meaningful utterance that is coherent with the context of the conversation. Not surprisingly, they have been found a strong correlation with

TRPs [40, 41]. However, all the authors addressing this point acknowledge the difficulty of establishing an operational definition of semantic completion. As there is no simple way to formalize semantic outcome, which till now is considered as a drawback for the field. Moreover, even if they could give a specific definition of semantic completion points, the issue of dependency on syntax completion remains.

Non-verbal aspects of face-to-face conversation are also important with linguistic signals for turn-taking. The daily observation of such signal is making the an eye contact with the listeners to indicate the end of the turn [48]. On the other hand, gestures have little significance in turn-taking [41], although Duncan in [35] did observed that the hand gestures were used during interviews to postpone others turn-yielding signals. The gesture is usually a support to the verbal signal. The studies of non-verbal signal has also been reported in [49].

2.2 Overlapping Speech

2.2.1 Definition of Overlapping speech

In addition to the signals indicating a clear end of a turn, there are other turn-taking signals like overlapping speech, which are described as frequent, but brief phenomena in conversations. According to model in [4], the briefness is explained by the fact that the onset of overlaps is placed at a point, known as TRP, where the current turn's completion is imminent. Therefore a speaker can predict TRP beforehand and select him/herself as next, resulting terminal overlaps. Thus the model explained in [4], explains overlaps as a result of *turn-taking principals*.

In addition to non-competitive overlaps at the TRP, as described above, there is another type of non-competitive overlap that supports the current speakers while confirming the right of the current speaker to continue the turn. This overlap is commonly known as continuer [50], backchannel [51] or response token [52, 53]. Further, two other types of general overlaps that have been indicated in [54, 55] are: 1) collaborative completions, where the second speaker completes the current speaker's turn by overlapping, and 2) choral productions, where speakers producing greetings or a toast by overlapping.

Unlike the definitions of overlaps with non-competitive discourse mentioned in previous paragraph, several studies suggested that the overlap can also be used as a device to compete for the turn in progress. In [56], the authors defines the turn competitive incomings in overlap as those instances in which the overlapper is heard as "wanting the floor" to him/herself at the immediate point in conversation without considering if the current speaker finished his/her turn. Similarly in [57–59], the authors describes overlaps as a form of simultaneous speech, which act as a violation of the current speaker's turn and also as an instrument for exercising power and control in the conversation. The authors further described it as an invasion for the floor which is initiated more than two syllables away from the initial or terminal boundary of a unit type [57]. In [60], the author described the overlapping speech from the overlapper's point of view. The authors defines overlaps as an event when the second speaker start speaking in a point which could not be a TRP and the current speaker cuts off more than one word of overlapper's unit type. Likewise in [61], the author defines overlaps from the perspective of the current speaker. The author stated that an competitive overlaps occurs when a speaker loses the floor before the intend of relinquishing it thus leaving the current utterance incomplete. In [62], the author characterizes competitive overlaps as the instances in which both the speaker perceives the in-overlap speech as problematic and in need of resolution. The author further suggests that the turn competition does not have to be one-sided, i.e., intent from overlapper's side only but both the speakers can compete and aim to drive each other out.

On the other hand, in [63], the author pointed out that to compete for the turn it is not always necessary to have overlapping speech (nor the overlapping speech is sufficient for the recognition of such events) as an utterance can perform a competitive function without overlapping. In [64], the author defined this type of competition as 'silent interruption'. An example of such an event can be a scenario when the overlapper starts the talking in the mid-turn pause from which the current speaker intend to continue the utterance but failed due to the intrusion [65].

Therefore, to summarize, the non-competitive overlaps either *occurs at TRPs* [4] or *begins and ends while the the current speaker still holds the floor* thus the incoming second utterance does not disrupt the current utterance [66, 67] whereas competitive overlaps incoming intrude the internal structure and the syntactic boundaries of a speaker's utterance [66] thus disorganizing the construction of the conversation [67].

2.2.2 Classification Schema for Overlapping Speech

Over years many classification and description of overlapping speech has been proposed. Some of these classification, proposed in the literature, has been mentioned below.

Based on the occurrence of speaker-switch, simultaneous speech, and the completion of first speaker's utterance, [64] devises a categorization scheme for competitive and non-competitive overlapping speech as shown in Figure 2.2. This scheme is later adopted by [68] for studying turn-taking styles. The scheme defines the following categories:

- Smooth-speaker switch: A smooth speaker-switch between the current and next speaker with no presence of simultaneous speech.

- Competitive overlap: Events where simultaneous speech occurs and the utterance of the first speaker remains incomplete.

- Non-competitive overlap: Events where simultaneous speech occurs but does not disrupt

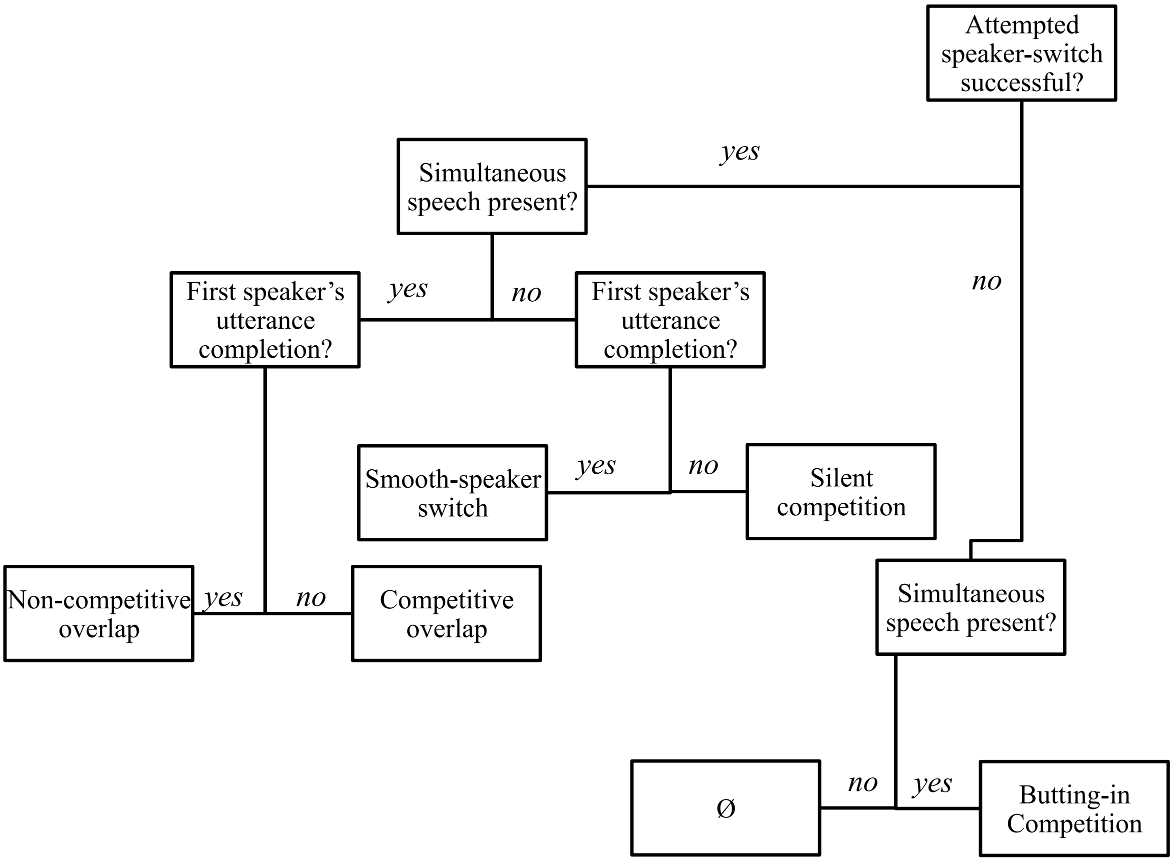

Attempted speaker-switch successful?
yes
Simultaneous speech present?
no
First speaker's utterance completion?
yes
no
First speaker's utterance completion?
Smooth-speaker switch
yes
no
Silent competition
Non-competitive overlap
yes
no
Competitive overlap
Simultaneous speech present?
Ø
no
yes
Butting-in Competition

Figure 2.2: Ferguson's classification of overlaps (non-competitive and compeititve overlapping speech) and smooth speaker-switch.

the flow of the conversation and the utterance of the first speaker is finished even after the overlap.

- Butting-in Competition: an unsuccessful attempt of competitive overlaps, the overlapper stops before gaining control of the floor.

- Silent Competition: An incoming of competitive turn just without overlapping.

Another classification of overlapping speech is proposed by Roger and Schumacher in [69]. They categorize competitive overlaps into successful and unsuccessful competitive overlaps. A schematic representation of their classification scheme is shown in Figure 2.3.

The schema defines the successful competitive overlaps as events in which the current speaker is prevented from completing an utterance by the incoming overlapper's turn while taking the floor; and in unsuccessful events, the overlapper attempts but fails to take the floor.

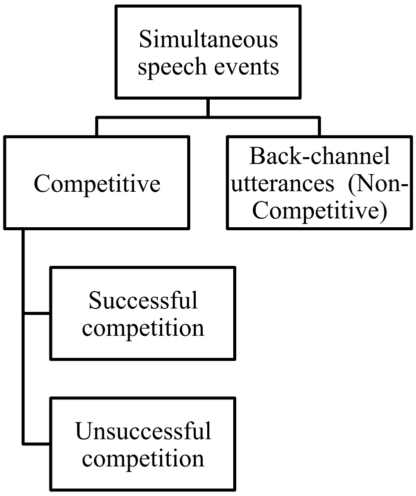

Simultaneous speech events
Competitive
Back-channel utterances (Non-Competitive)
Successful competition
Unsuccessful competition

Figure 2.3: Roger and Schumacher's classification of overlapping speech.

In comparison with the classification presented in [64], successful and unsuccessful competitive will respectively be competitive overlaps and butting-in competitive overlaps while neglecting silent competition events.

2.2.3 Identifying Competitiveness in Overlaps

For identifying competitive and noncompetitive overlaps in a conversation, researchers employed techniques to find properties of overlap that describes the competitiveness of the event. One of the first properties that researchers from conversational analysis and other field studied are the placement of the overlapping speech. In [70], the author investigated the precise placement of overlap onsets and found that they occur systematically at any place in the ongoing turn. The author mentioned three preliminary categorization of overlap onsets, according to their position relative to the TRP.

The onsets are:

1. transitional onset – focuses on completeness of the turn and are located at the TRP.

2. progressional onset – focuses on the flow of the conversation and starts at the silence interval of an ongoing turn, and

3. recognitional onset – focuses on the information recognized and are located at a point where the incoming speaker has gained sufficient understanding of the content of the current turn.

Unlike *transitional* and *progressional* overlap onset, which are the "byproduct of routine turn-taking practices", the *recognitional* onset can results in intrusion of turn and can be viewed as a competitive intend [70]. This type of events is also mentioned in [5]. These findings motivates the author to propose that the positioning of the overlap onset is related to the competitiveness of the overlap. In addition to the placement of the overlap onset, the author also identified if a speaker is aware of the overlaps and does not drop out or resolve the overlaps, this behavior is associated with turn competition [71].

Apart from the overlap onsets, several studies such as [46, 56, 62, 72–74] have claimed that prosodic features, including fundamental frequency height, intensity, speech rate and rhythm are important cues for turn competition in overlap. In [62] the author also showed that the speakers deployed these prosodic features along with cut-offs, sound stretches and repetition or recycling of prior material to indicate competitiveness. The author also suggests that the increase in pitch or volume can be regarded as turn competitive "hitches" that indicates competitive overlaps. These findings are also replicated for other languages in [75, 76] representing that in Italian human machine dialog repetitions and overlaps are not always necessarily competitive but plays an important pragmatic role to indicate the intent.

In [56], the authors proposed that the combination of raised pitch and volume is utilized by overlapper to compete for the turn. In addition, the authors contrast with [70] while suggesting that the timing of the placement of overlap onset within the current speaker's talk, is not a relevant feature classifying competitive and non-competitive overlaps. The authors also insisted that the overlap's lexical design and its pragmatic function also does not provide any distinctive separation between the two overlap classes (competitive vs non-competitive). This claim is later supported by [46, 77]. It is also observed in [77] that competitive overlaps include high pitch and amplitude to grab the attention from the current speaker.

Most of the above mentioned research focused on the prosodic design of overlapping speech and how it is used by interlocutors for competitive and non-competitive intent. But while doing so, these studies solely focused on either on a subset of prosodic features or on a particular position of overlaps.

Some of the subset of prosodic features that has been studied are: pitch and loudness [46, 56]; intensity [73]; fundamental frequency [74], speech rate [78] ; speech rhythm [72].

As for the position of overlaps, most of the studies focused on the overlaps placed clearly prior to possible completion [56, 62, 74]. Unlike the previous studies, in [79], the authors allowed the possibility that an incoming overlapping speech in terminal position may sometimes be competitive. The authors in the study hypothesize that both competitive and non-competitive overlaps can occur in any place in the conversation using different prosodic and positional design. The author also suggests that a combination of fundamental frequency and intensity is one of the most used prosodic feature in competitive overlaps where as recycling of lexical materials plays a major role in describing competitive overlaps. Duration of overlaps is also found to be the most distinguishing feature while classifying competitive and non-competitive overlaps using a decision tree [71, 78].

2.2.4 Research from Speech Communities

Understanding how to differentiate between the competitive and non-compeititve overlapping speech improves the naturalness of many speech technologies such as virtual agents, spoken dialog system (SDS) or even automatic speech recognition systems (ASR). As portrayed in [27], differentiating between turn competitive and non-competitive overlapped incomings is an essential part for a continuous conversation with a virtual agent. One important task for an spoken dialog system is to know/understand when to take the turn and yield the turn to the human partner. As a part of the task, the system should be able to recognize and manage the scenarios where the human partner takes the turn, with competitive intent, when the system is still talking. At the same time, the system should also be able to generate non-competitive overlaps to signal supports or to acknowledge the current speaker [26].

Thus with the aim of improving the quality and naturalness of spoken dialog systems and understand human-interaction, speech community has also been investigating the acoustic and temporal properties of overlap. But compared to the other research area dealing with overlaps and turn-taking in spoken conversations, there have been very few studies on the speakers' competitive and non-competitive turns.

For classifying overlaps different type of features has been explored, such as hand motion and disfluencies [73], body movement features from both speakers and contextual prosodic features from the overlapper [80], gaze, voice quality and contextual features –preceding and during overlaps [81].

Aiming to predict competitiveness, in [82], authors found that incoming of competitive overlaps are not random and context can be used to predict their occurrences. A similar conclusion is observed in [83], suggesting that interruptions are more likely to occur in intonational phrase units (IPUs) rather being random. In [80], the author used body movement features from both speakers along with prosodic feature from the overlapper to investigate the context that surrounds the overlaps. In [81], the author extracted various context features preceding and during overlaps to compare the performance of overlap classification for competitive and cooperative overlaps. While doing so features such as gaze, voice quality were also introduced with the acoustic feature set.

Apart from the few studies mentioned in above paragraph, there are some other important studies which focuses on studying overlaps. Even though the below studies did not differentiate between competitive and non-competitive overlaps but it gave us a ground for starting the investigation on differentiating the pragmatic role of overlaps as competitive vs non-competitive.

To find the importance of modeling overlap and its properties, the authors carried out a quantitative study of overlaps on two meeting corpora from ISCI [84, 85] corpus and two telephone corpora "The Switchboard" [86] and "CallHome English" corpora. This study is conducted using more or less naturalistic spoken interaction which is recorded in a separate audio channel using a close-talking microphone. The study analyzed raw acoustic data to find for recurrent acoustic correlates of overlap. The findings of the study suggests that fundamental frequency along with energy at the onsets of turns in overlap were higher compared to the onsets of turns from silence.

More recently, [26] have analyzed the "Columbia Games Corpus" to identify the prosodic, syntactic and acoustic cues that precede turn changes, turn retentions and backchannels. The findings of the study shows that the inter-pausal units (IPUs), preceding the turn transitions with and without overlap, exhibit comparable turn-yielding cues. The study only considered smooth turn changes and did not address cues that potentially signal competitive and non-competitive overlaps. As stated by [5], that turn-competitive overlaps are usually ignored due to their tendency to break the flow of the conversation.

2.3 Silence

The ambiguous value of silence in daily conversation, be it written or spoken, has arise different theories regarding the importance and function of it for many years. Silence bears distinctive cultural characteristics in communication. In different cultures, silence conveys different meanings and attitudes. Silence, which one person intends as a sign of respect, may be interpreted as rudeness by others.

For over decades researchers from many field has been studying and analyzing silence in human interaction. The research focus includes, but are not limited to, the role of silence in conversation [57, 87, 88]; silence as nonverbal communication [89, 90]; interpersonal silence [91,92], silence as a conflict-management strategy [93], and the use of silence within the context of psychotherapy [94–97].

Early studies on silence mostly focuses on theoretical speculations of the role of silence in human interactions. However, most of the studies on communication primarily focuses on 'talk' relatively ignoring 'silence' [98]. According to [12], the author observed that even in linguistics, the silence is recognized as an empirical datum, which is traditionally defined as the absence of speech sounds.

The range of the interpretations may vary from one culture to another as it is subjective and relative, which indicates silence is both context specific and culture specific [99]. Researchers in communication community recognize silence as a semiotic unit of nonverbal communication [100, 101]. In addition, communication scholars also realize that the meanings associated with silence are not universal in nature, but culturally and contextually defined [102, 103].

In the context of culture, in the eastern cultures, silence is particularly appreciated and associated with several positive impressions in communication, while in the western culture, silence is usually avoided as it is regarded as a kind of social weakness or a sign of withholding and un-cooperative personality [98] or a manifest of the speaker's lack of knowledge [104, 105].

In recognition of the context specificity of the meanings of silence, [106], for example, defines silence as an act of non-verbal communication that "transmits many kinds of meaning depending on cultural interpretation". Silence can also be defined as "the absence of talk" which contains certain communicative purposes [103]. Similarly, the author in [107] insists that silence must bear a communicative function, sometimes peculiar to the interlocutors and sometimes to the context and culture where it appears. Silence has also been reported to have illocutionary force to perform a speech act that seems to exist universally, naturally displaying cultural variance [103], [14], [13], [107].

2.3.1 Role of Silence in Human Interaction

In 1973, the author in [108] investigate the possible function of silence in human interactions when used with other nonverbal cues. The study analyze and states five functions of silence. The functions are 1) linking, 2) affective, 3) revelation, 4) judgmental, and 5) activating. Despite describing the functions of silence in details, the theoritical stance of the author is flawed due to the implicit assumption that the meaning of silence is uniform for different context and culture.

At the same time, the author in [10] defined three major forms of silence. They are 1) psycholinguistic silence, 2) interactive silence, and lastly 3) socio-cultural silence. The author also discuss the function of silence as 1) an indication respect or disrespect for the current authority, 2) a strategy for disapproving a 'violent expression and anger', 3) as a tactic by authorities to create opportunities for sub-ordinates to think independently for themselves, and 4) as a device to rhetorically control behavior.

The study in [109] explains the role of silence in several communication contexts, such the function it plays in human thought processes, its purpose in everyday interpersonal communication, in social and in political life, and also includes its function in counseling and psychotherapeutic contexts.

The study in [103] mentions two primary types of silence. These silences are pauses and hesitations that is used as a tool in verbal turns to take short time for thinking where the second type of silence is the long silence used intentionally, and contain certain meanings and illocutionary force, which is "eloquent silence" [103, 110]. The author in [14] also focused on this eloquent silence and investigated it in Akan society with a socio-pragmatic approach which puts forward that silence embodies social and rhetorical influence, conveys meaning and therefore, has communicative functions. [111] differentiating between intentional and unintentional silence, states that intentional silence conveys meaning in communication.

The study in [112] outline the functions of silence using cognitive, discursive, social, and affective functions. From cognitive perspective the author suggest that the lengths of pauses and hesitations in a conversation is used as a processing time before speaking or listening at the same time, pause can be used as a tool for marking utterance boundaries in discourse, while governing or organizing social relations [113].

Similarly the author in [114] analyzed silence in conversation in interaction level suggesting that silence may signal asymmetry between the speakers along with signaling turn-taking. The author also analyze silence in cognitive level suggesting that silence may co-occur with mental planning of the upcoming utterance. Moreover, as for the function of silent speech segments, the author insists thats they may signal both agreement and disagreement. The author emphasizes in the study that for understanding speaker's intentions behind silence, it is essential to analyze the context.

In literatures over decades silence has range of functions in various contexts of everyday life varied through cultures. The next few sections includes the studies that address the meaning and use of silences.

2.3.2 Conversation Style and Silence

Conversational styles can also be characterized using silence based on pause length, speed and frequency of the conversation [98]. In the study [98] the author reported that New Yorkers perceive slow speakers (Californians) as "withholding and uncooperative".

2.3.3 Politics and Silence

Silence also plays a major part in political speech. As cited in [99] the author in [115] mentioned two types of silence in politics; One of them occurs due to the break down of speech; and the another one occurs due to the failure to utter relevant words which the cited author explained as a political strategy. This type of silence is also been discussed in [14], where the author labels this kind of silence as communicative silence in Akan culture and referred the attitude as "absence of relevant talk" by [101], which overlap with the proposed situation of "irrelevance" in [116].

In [117], the author defines the strategic silence as refusal to communicate verbally, by a public figure which leads to (a) violation of expectations, (b) drawing public attributions and fairly predictable meanings, and (c) seems intentional and directed at an audience. In [118], the authors studied the pragmatic motivations and use of silence in Turkish political talk shows. The author concluded in the same line of the study as [98] and insisting even though silence does not mean lack of knowledge or weakness but in competitive arena speech represents sign of 'Power'.

2.3.4 Politeness and Silence

Based on cultural perspective silence can convey different meaning and attitudes. In some perspective silence represent respect whereas in another silence can be interpreted as an intent of rudeness or insult. The politeness theory suggests that silence can be used as a strategy to avoid face threats and is the most polite speech act, especially in the Eastern culture. For example, in Japanese culture any disagreement, refusal and rejection, are the most common speech acts, performed through silence as a politeness strategy [13].

2.3.5 Power and Silence

Author in [14] suggests that in some African societies, silence appears as a manifestation of power. Silence is used there as tool for the powerful to show the superiority. At the same time, the suppressed/weak remains silence in submission. The author pointed out that silence could be also used as a tool for social control, in some societies it can be used as a way to punish the enraged or those who committed violence.

In [119], the author investigate silence in classroom interaction and found it as one of the important component in the interaction dynamics by studying the observed silence along with nonverbal behaviors of both the students and teachers. The author conclude that even though there is an inequality in status between students and teachers, the participants used silence to negotiate power.

2.3.6 Conflict Management and Silence

In [120], the author investigates how silence can be used in conflict management in an Italian village. The author analyzed a real scenario, he witnessed, between a father and daughter. In the study, the author concluded that the use of silence was strategic in both father and daughter case. The daughter used the silence to avoid any anger verbalization against her father, where as father used it avoid any irreparable damage in family relations. Similarly in [121], the author suggested that in a potential conflict situation, keeping silent helps to manage the situation and substitute for an expressive of negative emotion, where as if the speaker gives way to a verbal expression, it escalates the situation, leading to "everlastingly destructive consequences".

2.4 Turn-Taking Models

Research in psycholinguistics suggests that humans process utterances incrementally [122]. Pointing to the fact that when we hear an utterance, at each point we try to hold a semantic representation of it. So to match the human language processing and to allow natural interactive language-based applications, computational linguists proposed and implemented incremental parsers in [123, 124], where some of them targeted spoken dialog systems domain.

A way to analyze the turn-taking mechanism is by using an artificial agent. The study [125] contains the most detailed work on reproducing human turn-taking behavior in artificial conversational agents. The author's core work is an architecture called "Ymir" [126] that worked with multimodal face-to-face interaction (e.g. hand gestures, gaze, backchannels, including discourse planning), where all the turn-taking management has been done with "Gandalf", an embodied agent acted as a guide for the system. Rules were employed for the agent to manage turn-taking behavior; one such example includes that the system must take the turn after a 50 ms pause following a user utterance, given it is a complete utterance and is turn-yielding.

The TRIPS spoken dialog architecture [127], has been used to develop a number of dialog systems over almost a decade on tasks such as emergency response and evacuation planning [128]. The initial implementation handled turn-taking in the standard rigid way, where as the later version featured incremental interpretation and generation and some other features [129]. Another important feature of TRIPS is that it has separated the discourse from task-related components. Discourse information is captured by a Discourse Context (DC), contains the past log of users and system utterances and the set of current salient entities, but also discourse level obligations, and current turn status. All the components of TRIPS run incrementally, allowing a flexible turn-taking behavior.

In [130] a turn-taking architecture for the Reading Tutor of CMU's Project LISTEN has been described. Works from socio and psycholinguists, discussed in above sections, have uncovered varieties of features that help human to detect the end of turns. But in practice, spoken dialog systems as discussed, have adopted a simple approach. In most of the spoken dialogue system, turn ending is considered when pause, detected using Voice Activity Detector (VAD), lasts longer than a fixed threshold. However, this practice leads to suboptimal behavior in many instances where chances of cases like cut-ins and latency might appear. To overcome these problems several researchers have proposed to use features from dialog. In [131] use of decision trees are shown to classify pauses longer than 750 ms as turn boundary (TB) or turn-internal pause (TIP). Uses of features from semantics, syntax, dialog state, and prosody were able to improve the classification accuracy from a baseline of 76.2% to 83.9%. After many years of research in human-machine domain a large numbers of these systems are developed. Some of the well-known projects are Communicator [132], "How May I Help You?" [133] and many more.

2.5 Summary

This chapter has reviewed some key findings from previous literature regarding events in turn-taking. The study, in the chapter, focused on the fundamental concept of turn-taking such as construction of turns to transition relevance place (TRP) to the complex events such as overlapping speech discourse to functional meaning of silence. The study includes research from many fields, including conversation analysis, psychology among others, and how these research has been incorporated in spoken dialog systems and what is the state-of-the-art in speech communities. From the overview of this chapter, we observed that there are very few studies in finding the function of silence which are methodological. Most of the research on silence highlights the importance of context in understanding its function. Thus pointing that to categorize function of silence, we need to design its feature from its surroundings. Similarly, for overlapping speech research, it is observed that most of the studies from speech focused only on prosodic characteristics of the event, using designed feature to classify competitiveness of the overlap and compare to turn transitions and cues for turn yielding, no research actually focused on classifying discourse of overlaps using high-dimensional acoustic features along with other vocal features such as linguistic cues. So this dissertation aims to address this lacking in the state-of-the-art research to see if computational models can be designed for these turn-taking events.

Chapter 3

Dataset

To design computational models for turn-taking behavioral system and understand what significant role it plays in determining the course of the spoken conversation, one of the important challenge is the annotation of ecologically valid data with real behavioral expressions. For the annotation, operational definitions and guidelines are required. Therefore, the content of this chapter includes information regarding the ecological real data, the design of the annotation scheme, followed by the corpus analysis. This chapter also includes information of other data which is also used in this dissertation.

For the analysis, experiment and the evaluation of the computational models, a dyadic call center spoken conversations scenarios is used. Typical scenario of call center is that both the agent and the customer engage in real conversation to achieve a goal such as information seeking or problem solving. Most of the previous research in turn-taking dynamics, especially overlapping speech discourse, focused on meeting corpora [79] or other small datasets. There are only a very few corpora which have been collected in real-life situations that are large enough to understand and more importantly to model these behaviors.

The Signal and Interaction System Laboratory (**SISL**) Human-Human Dyadic Conversation corpus consists of Italian call-center conversations with real-users and contains manual transcriptions of such dialogs, annotation of overlap discourse, semantic category and affective behavior annotations, including empathy, and basic and complex emotions. For our study we have chosen this data due to its size and naturalness. In addition, we also used Italian LUNA Corpus for a comparative study across corpora on dialog act segmenter and classifier .

Contents published in:
Shammur Absar Chowdhury, Evgeny A. Stepanov and Giuseppe Riccardi, *Transfer of Corpus-Specific Dialogue Act Annotation to ISO Standard: Is it worth it?*, in Proc. of 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož (Slovenia).
Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi, *Annotating and Categorizing Competition in Overlap Speech,* in Proc. of ICASSP. IEEE, 2015, Brisbane, Australia.
Firoj Alam, Shammur Absar Chowdhury, Morena Danieli, Giuseppe Riccardi, *How Interlocutors Coordinate with each other within Emotional Segments?*, COLING, Osaka, Japan, 2016.

3.1 SISL Conversational Discourse Corpus

The SISL Human-Human Conversational Discourse Corpus is a subset of a large Italian call-centers corpus, which has been collected with real-users that were engaged in real conversations with call center agents. The customers are calling the agents to solve some specific problem or for seeking information. The inbound Italian phone conversations are recorded on two separate audio channels with a quality of 16 bits, 8kHz sample rate. The collected corpus have an average duration of 396.6 ± 197.9 for all 10K conversations[1]. Since the motivation behind the collection of the Italian call-centers corpus was to analyze real-life conversation dynamics along with other affective behavior, therefore no prior knowledge has been given to the subjects during the data collection. The data also excluded any personal information regarding the customer except the gender due to its privacy policy.

3.1.1 Transcription

To better understand the interaction between the interlocutors in a conversation, a subset of 955 conversations has been manually transcribed. For the transcription process, an initial speech segment boundaries were given using an automatic turn segmenter [134]. The annotators were allowed to change those boundaries with constrain of minimum speech segment boundary to be 2 seconds with maximum of 10 seconds. The annotators also instructed to mark the Cross-Talk **CTK**. It is defined as a background intelligible speech that are not providing any information to the speakers. The instruction also suggests to ignore background noise such as phone dialing, other background noises, traffic noise, are also ignored. Though the human sounds such as cough, laughter, sneezes, has its importance, however, for our task they are considered as environmental noise and thus labeled as **NOISE** in the transcription.

These manual transcriptions are also used to design an in-house Automatic Speech vs Non-Speech Segmentation and Automatic Speech recognition systems as described in Section 7.2.1.1 and 7.2.1.2. Moreover the linguistic feature design and analysis are also presented in this dissertation are based on this manual transcription along with automated transcription from ASR.

3.1.2 Overlap Discourse Annotation Scheme

A small subset of the conversations are analyzed by an expert psycholinguist who listened each recorded call by applying a systematic direct observation protocol [135] while focusing only on overlapping speech segments.

The observations allowed the psycholinguist to identify different kinds of overlapping speech segments, differing with respects to their pragmatic functions, speaker intentions and linguistic

[1] The original dataset contains 10063 conversations where average ± std is 395.9 ± 198.2 and later some of them has been discarded.

structure. For instance, most of the analyzed conversations showed that overlapping speech segments are co-occurring with greetings at the end of the phone conversation.

The occurrences of speech overlap were characterized by significant variations of their prosodic profiles where some of them showed the intention of the intervening speaker to "grab the floor" of the conversation, i.e., to compete with the other speaker in view of controlling the turn taking structure of the dialog. One such case is the tendency of the agents to interrupt the customer when they believe to have understood the customer's question while the latter insists on providing more information.

Sometimes, however, the intention to "grab the floor" did not show a competitive attitude of the speaker. For example, several overlapping speech segments sound as being collaborative completions by the intervening speaker. Those occurrences could be classified as one out of several forms of back-channeling phenomena.

On the basis of this observational analysis, we designed the annotation guidelines for segmenting and annotating the speech overlaps with the competitive and non-competitive labels. The annotation guidelines include the following:

1. Each overlapping segment may contain more than one overlap instance of the same category. Instances may be separated from each other with a gap less than 40ms.
2. If a speaker thinks aloud during another speaker's turn that is considered an overlap instance.
3. Co-occurrences of "false start" by both the speakers are considered instances of speech overlap if and only if the segments contain complete words and the annotator can infer the speaker's intention on the basis of the perceived intonation of speech.
4. Annotators are asked to reject a conversation or ignore segments if they contain poor quality audio, unintelligible speech, background noise, human sounds like cough, sneezes and laughs.
5. The annotator's judgment includes the appraisal of the speakers' intention on the basis of supra-segmental variations including speech rhythm, accent and intonation along with peculiarities of the semantic content of the utterance.
6. Inferring the annotation label on the basis of the annotator's knowledge of what will occur later in the conversation, i.e., outside the turn being considered, is to be strictly avoided.

Using the above guidelines, the annotators were asked to annotate the segments into one of the following two categories:

Competitive (Cmp): Scenarios where 1) the intervening speaker starts prior to the completion of the current speaker, 2) both the speakers display interest in the turn for themselves, and 3) speakers perceive the overlap as problematic.

Table 3.1: Dialog excerpts from the annotated corpus. Speech overlaps: bold form between [and], Hesitations: (.), Rising intonation: ↗, Falling intonation: ↘.

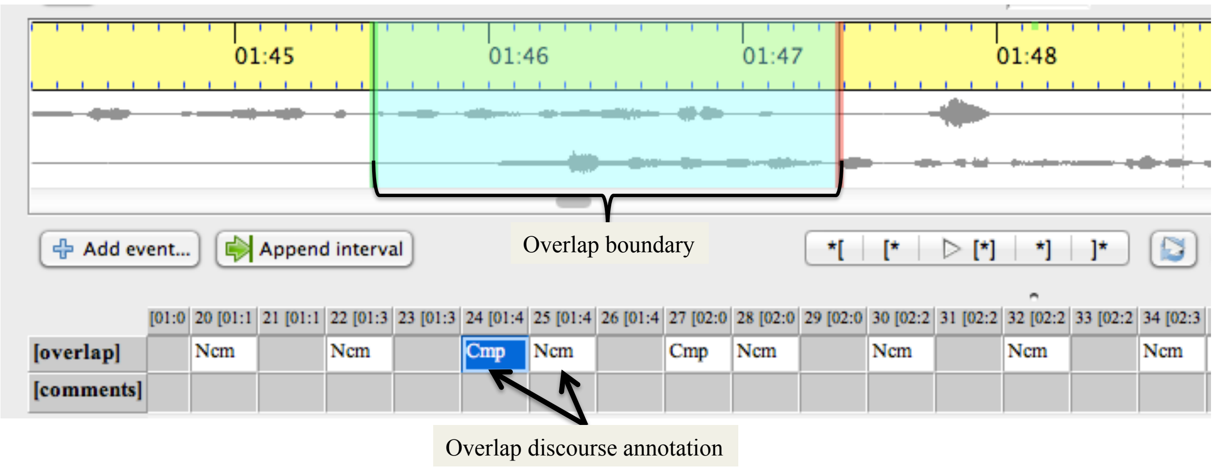
Non-Competitive Ncm
S1: è una piccola [**cosa però**] ↘ se (.)
S2: [**no signora ↘ ha**] fatto bene ↘
S1: it is a [***little thing*** *] ↘ if (.)*
S2: [***no madam ↘ have*** *] done well ↘*
Competitive Cmp
S1: perché questa [**è la vostra ultima**] che ho ↗↘
S2: [**no signora ↗ dal**] 31 marzo non è con noi ↗
S1: because this [***is the your latest*** *] that have ↗↘*
S2: [***no madam ↗ from*** *] march 31 you are not with us ↗*

Non-Competitive (Ncm): Scenarios where 1) another speaker starts in the middle of an ongoing turn, 2) both parties do not show any evidence for grabbing the turn for themselves, 3) speakers perceive the overlap as non-problematic and 4) the intervening speaker use it to signal the support for the current speaker's continuation of speech. In Table 3.1, we report two examples of overlap segments with their English translation. The overlap segments are represented in bold form between square brackets and reported tone direction, based on IPA notation [136]. In the first example, the overlap speech segments of speaker S1 and S2 have a falling intonation: S1 hesitates and S2 intervenes for reassuring her. The opposite occurs in the second example: S1 speech has a rising-fall intonation, whereas the tone of S2 speech is constantly rising. S1 is surprised and overwhelmed by the sharp tone of S2.

3.1.2.1 Annotation Procedure

Two expert annotators, Italian native speakers, performed the annotation task. As specified in the guidelines, they manually segmented the speech overlap occurrences and labeled each segment as competitive or non-competitive in 565 conversations of approximately 62 hours of spoken content with an average duration of 395 seconds.

The annotation of overlap discourse is carried out using the Partiture editor of EXMARaLDA [137]. EXMARaLDA is an acronym of "Extensible Markup Language for Discourse Annotation". It is a tool of concepts, data formats, and tools for the computer assisted transcriptions and annotation of spoken language, and for the construction and analysis of spoken language corpora. The annotation of the overlap requires a specific designed tiers of annotation that is added to the xml file by the annotator when there is presence of overlapping speech. A comment layer is also added in case of any atypical phenomena found or the annotator faced any confusion. For the annotation task only audio file is used thus decreasing the influence of transcription on the annotator. An example of annotation using Partitur-Editor is shown in Figure 3.1, which contains overlap and comment tiers.



	[01:0	20 [01:1	21 [01:1	22 [01:3	23 [01:3	24 [01:4	25 [01:4	26 [01:4	27 [02:0	28 [02:0	29 [02:0	30 [02:2	31 [02:2	32 [02:2	33 [02:2	34 [02:3
[overlap]		Ncm		Ncm		Cmp	Ncm		Cmp	Ncm		Ncm		Ncm		Ncm
[comments]																

Figure 3.1: Annotation example using Partitur-Editor, containing overlap discourse label and the comment tiers. For the annotation, only audio signal was available to the annotator. Cmp represents competitive overlap where Ncm represents non-competitive overlap discourse.

3.1.2.2 Evaluation of Annotation

To assess the reliability of the annotations we calculated inter-annotator agreement by using the kappa statistics. Equations 3.1 - 3.3 define Cohen's κ [138, 139] and its observed (P_o) and chance (P_e) agreements in terms of *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN). In the equations N = TP+TN+FP+FN.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3.1}$$

$$P_o = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{N}} \tag{3.2}$$

$$P_e = \frac{\frac{(\mathrm{TP}+\mathrm{FP})*(\mathrm{TP}+\mathrm{FN})}{\mathrm{N}} + \frac{(\mathrm{TN}+\mathrm{FP})*(\mathrm{TN}+\mathrm{FN})}{\mathrm{N}}}{\mathrm{N}} \tag{3.3}$$

For calculating the agreement two annotators worked independently over a set of 28 spoken conversations randomly extracted from the call center corpus. The amount of spontaneous speech annotated for the inter-annotator agreement test was around 3 hours 17 minutes. The Kappa statistics is frequently used to assess the degree of agreement among any number of annotators by excluding the hypothetical probability that they agree by chance. By evaluating our data we reported $\kappa = 0.7033$.

Additionally, to quantify the inter-annotator agreement as human-performance in categorization of overlaps, a Positive (Specific) Agreement (P_{pos}, Equation 3.4) [140], identical to the widely used F-measure (Equations 3.5 - 3.7) [141], was also used to obtain pair-wise F-measure as an evaluation to the annotator agreement. In this case we obtained an F_1 of 85%.

$$P_{pos} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}} \tag{3.4}$$

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} && (3.5) \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} && (3.6) \\ F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} &= \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}} && (3.7) \end{aligned}$$

The cases of disagreement were discussed in a consensus meeting by the annotators and the author of the guidelines. The most relevant disagreement between annotators concerned speech disfluencies, including false starts, repairs, and filled pauses. In most of the cases consensus was reached between the two annotators.

3.1.3 Affective Behavioral Annotation

In the corpus the annotation of the affective behavior include *empathy* on the agent channel, and *anger* and *frustration* on the customer channel. For the annotation, we adopted the *modal model* of emotion by [142] in order to define empathy and design annotation guidelines for the annotators. Gross's modal model is based on appraisal theory, which has been studied by many psychologists for the investigation of emotional states. Appraisal models of emotion suggest that organisms appraise (i.e., evaluate, interpret, explain) events/situations based on the appraisal process in order to determine the nature of ensuing emotion as discussed by [143].

According to the *modal model*, *"emotions involve person-situation transections that compel attention, have meaning to an individual in light of currently active goals, and give rise to coordinated yet flexible multisystem responses that modify the ongoing person-situation transection in crucial ways"* [144, 145]. The key idea of the modal model is that emotional states unfold over time, and their response may change the environmental stimuli, and that may alter the subsequent instances of that and other emotional states. It is a useful framework for describing the dynamics of the emotional states, which manifests over time, leads to the generation of an emotional sequence from the interlocutors' emotional manifestations. For example, the sequence of emotional states between an agent and a customer could be Frustration (C) $\rightarrow$ Empathy (A) $\rightarrow$ Satisfaction (C), here A for agent and C for customer.

To design the annotation guideline, we have done an extensive analysis of one hundred conversations (more than 11 hours), and selected dialog turns where the speech signal showed the emergence of empathy, basic emotion, such as anger, and complex emotion such as frustration. In our qualitative analysis, we investigated the relevant emotional speech segments, which were often characterized by some perceivable variation in the speech signal. We observed that such variations could co-occur with emotionally connoted words, but also with functional parts of speech, such as adverbs and interjections, which could play the role of lexical supports for the variations in emotional states. We hypothesized that perceivable variations in the speech are a possible signal of an appraisal process. On the basis of those observations, we have designed annotation guidelines whose critical principle was to focus annotators' attention on their own perception of the variations in the speech signal as well as the variations in the linguistic content of the utterances.

For example the annotation guidelines include the following recommendations for the annotators:

- annotating the onset of the signal variations that supports the perception of the manifestation of emotions,
- identifying the speech segments preceding and following the onset position, and
- annotating the context (left of the onset) and target (right of the onset) segments with a label of an emotional state (e.g., frustration, empathy, etc.).

In addition, the annotation guidelines include operational definitions of emotional states related to the given domain of application. For example, in this annotation task, the operational definition of empathy is defined as ‘‘*an emotional state triggered by another's emotional state or situation, in which one feels what the other feels or would normally be expected to feel in his situation*” [146].

The annotation task was performed by two expert annotators who worked on non-transcribed spoken conversations by following the annotation scheme reported above. In this task, the annotation unit is the speech segment. They annotated *Empathy* on the agent channel and *Frustration* and *Anger* on the customer channel. The annotators labeled *Neutral* on the segment that appeared before any emotional segment to define the context, as mentioned earlier. Finally, the annotated corpus includes 1894 customer-agent conversations (210 hours and 23 minutes in total). In order to evaluate the reliability of the annotation we measured inter-annotator agreement on the annotated segments, and obtained an average $\kappa = 0.74$. More details can be found in [147–149].

3.1.4 Corpus Analysis

3.1.4.1 Corpus Summary

Based on the annotation guidelines, 565 conversations were annotated with overlap discourse, containing manual transcription. Along with the overlap discourse, agent and customer channel and gender are also specified in the annotation information layer. The corpus consists of 62 hours and 23 minutes of conversations. Out of 62 hours and 23 minutes of the conversation, $\approx 35\%$ of conversational space are silences including pause, gaps, between- and within-speaker lapses, where as other $\approx 55.43\%$ of the conversation floor belongs to non-overlapping speakers turn. From this dataset, a total of ≈ 5 hours and 8 minutes, which is $\approx 8.2\%$[2].

In the corpus, the most frequent turn-taking signal is smooth-switch and gaps, includes turn-changes like no-silence-no-overlaps, silence-in-between, i.e., gaps as shown in Figure 1.2. Their duration distribution is presented in Figure 3.2.

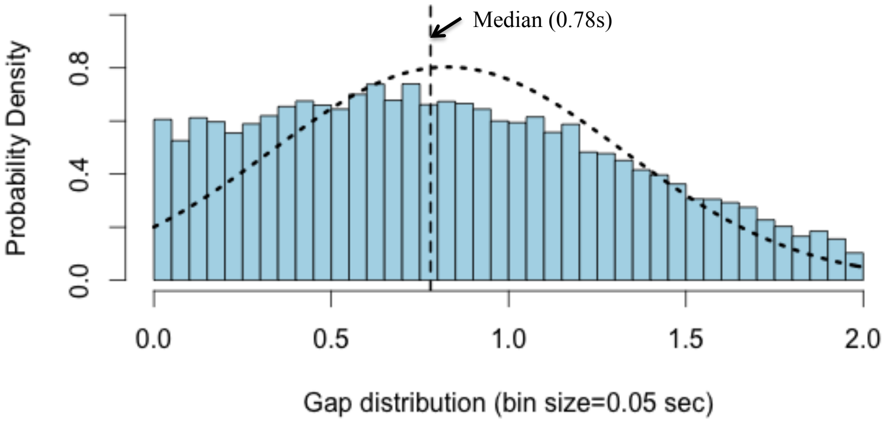


Figure 3.2: Duration distribution of gaps (smooth-switches) labels in the corpus.

Even though overlapping speech is a violation of turn-taking rule, it is observed that around $\approx 39.53\%$ ($\approx 40\%$) of the turn taking (speaker changes) occurs while overlapping with the interlocutors as shown in Figure 3.3.

After forced alignment and fixing overlap boundaries, a total 15,899 overlap segments, of a total duration of 5 hours and 8 minutes is obtained. Among the overlapping instances, $\approx 24\%$

[2] This overlapping speech duration is based on alignment and filter technique used in Chapter 5 and 7. The original manual annotation included extra silences surrounding the overlaps. Thus containing ≈ 7 hours and 57 minutes of overlap data.

of overlaps are of competitive nature as shown in Figure 3.4. When the overlapping speech are studied using unsupervised technique, in Chapter 4, a similar distribution of competitive *vs* non-competitive overlaps are found, thus indicating that this is the natural distribution of overlaps in the dataset. The duration distribution of the overlap discourse (*Cmp vs Ncm*) are given in Figure 3.5.

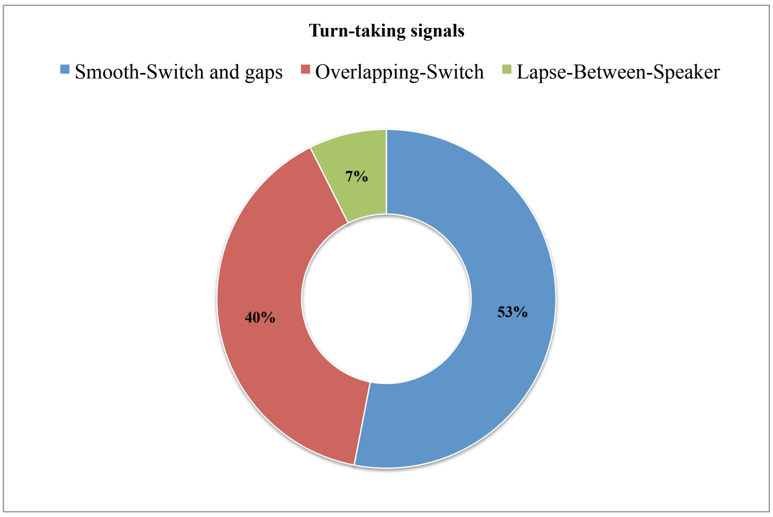


Figure 3.3: Distribution of turn-taking signals in the corpus.

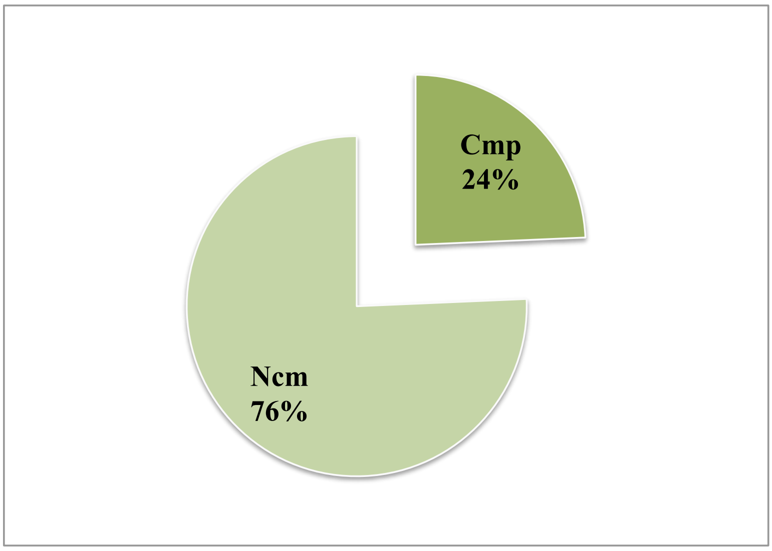


Figure 3.4: Distribution of competitive *vs* non-competitive labels in the corpus.

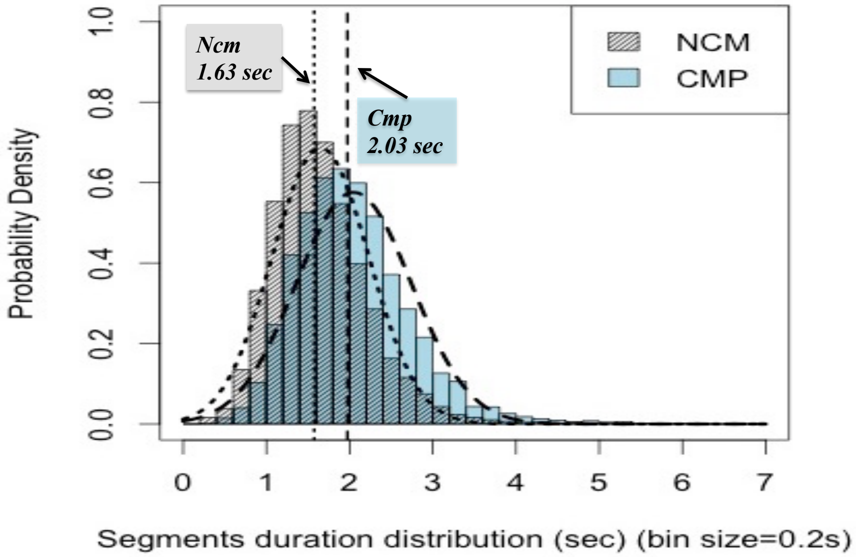


Figure 3.5: Duration distribution of competitive *vs* non-competitive labels in the corpus.

Another variation of switching speakers by silence in between (gaps) are between-speaker lapses (where the silence in between in $>=$ 2 seconds) shown in the Figure 3.3. Even though this is a very rare (only 7%), the long gaps has the potential functions to play in determining the conversational behavior. The distribution of these turn signal is presented in Figure 3.6.

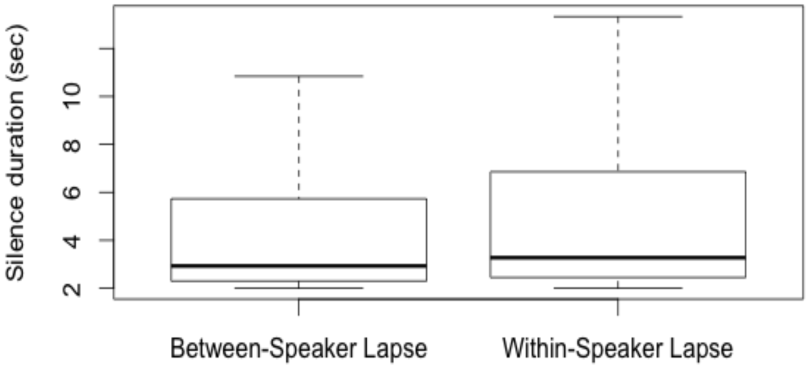


Figure 3.6: Duration distribution of lapse (between- and within-speaker) labels in the corpus.

The duration distribution of both agent and customer's non-overlapping turns in presented in the Figure 3.7. This turns are created using technique described in details in Section 7.2.2, using forced aligned transcription. For the forced aligning of the transcription, a domain specific ASR is used which is designed with manual transcription mentioned in Section 3.1.1.

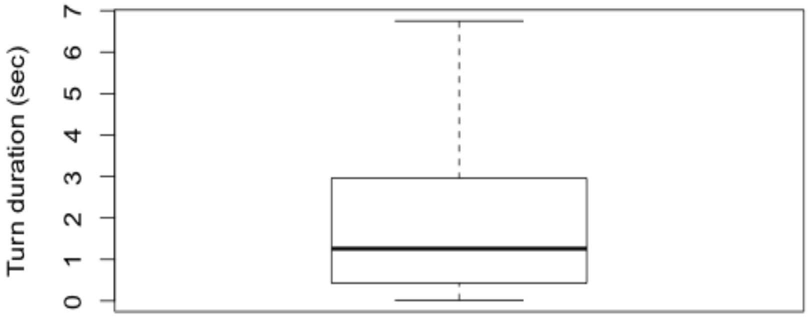


Figure 3.7: Distribution of non-overlapping turns in the dataset.

Silence instances duration, on the other hand, covers $\approx 35.21\%$ of the conversations. The instances includes within speaker silences (pauses, within-speaker lapses) and between speaker silences (gaps and between-speaker lapses).

Table 3.2 contains details of silence statistics and percentage of each type of instances present in the data. The duration distribution of pauses and within-speaker lapses are shown in Figure 3.8 and in Figure 3.6 respectively.

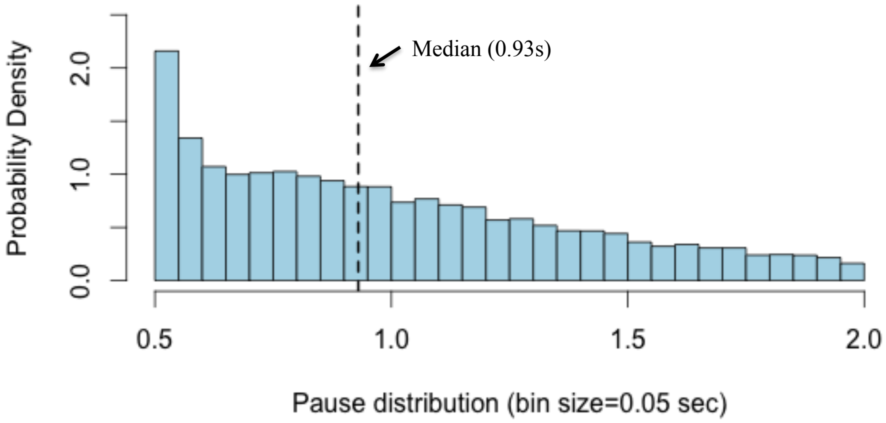


Figure 3.8: Duration distribution of pause (within speaker silences $< 2.0 seconds$) labels in the corpus.

Table 3.2: Duration and frequency statistics of different types of silence in the corpus

Stat.	Pause	Within-Speaker Lapse	Smooth-Switch	Between-Speaker Lapse
Minimum	0.500	2.000	0.010	2.000
1st Qu.	0.680	2.450	0.420	2.300
Median	0.930	3.275	0.780	2.930
Mean	1.009	8.003	0.827	7.981
3rd Qu.	1.270	6.860	1.190	5.728
Maximum	1.990	304.3	1.990	452.3
Total silence instances	39886			
Percentage	29.39	7.51	55.36	7.75

3.1.4.2 Linguistic Analysis

An analysis has been conducted using manual transcriptions to understand *what* has been said while unfolding competitiveness in overlapping speech. For the study, only the overlapper (speaker responsible for initiating the overlaps) turn is used. The study includes finding most frequent as well as syntactic (part-of-speech) categories.

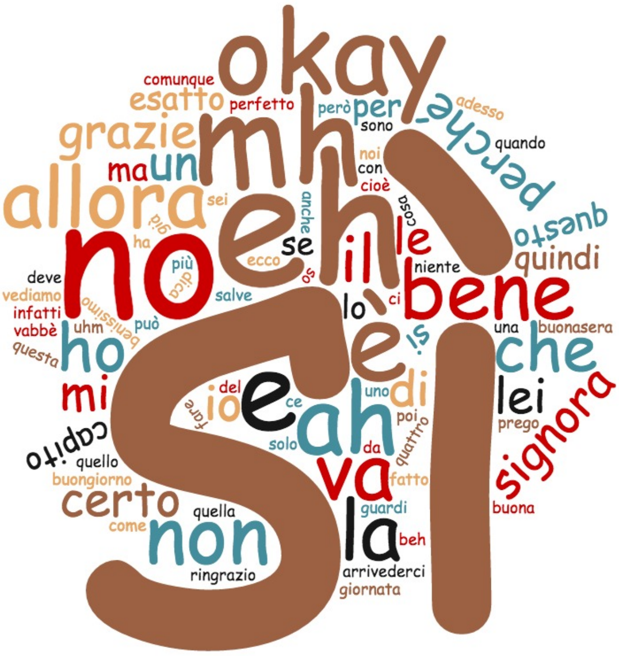


Figure 3.9: Word-cloud for Non-competitive class uni-gram feature only.

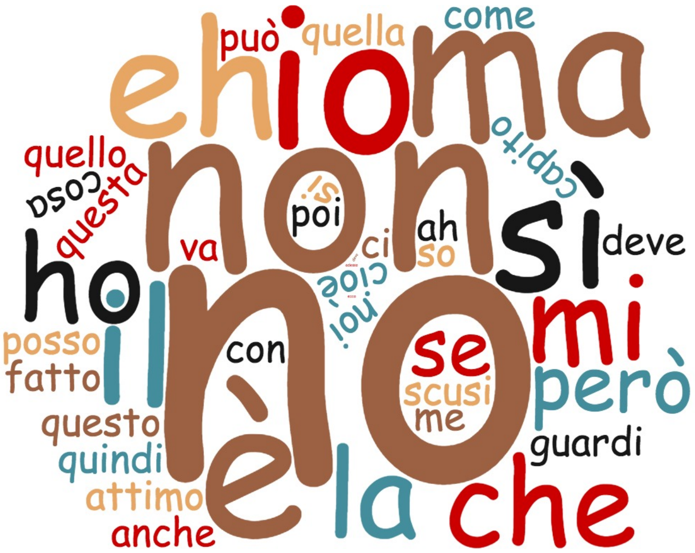


Figure 3.10: Word-cloud for Competitive class uni-gram feature only.

Table 3.3: Most frequent bi- and tri-grams for each overlap. English translations are inside parenthesis.

Overlaps	Examples of most frequent bi and tri-grams only
Non-competitive	sì sì (yes yes), sì sì sì, va bene (well), no no, ho capito (I have understood), no no no, lo so (I know), eh sì, grazie a (thanks to), la ringrazio (thank you), grazie a lei (thanks to you), no non, ah okay, ah ho capito, un attimo (just a moment), si figuri (never mind), mh mh, bene va (goes well), va bene va (alright), mi dica (tell me), non si, sì perché (yes why), sì no, non lo (not), eh eh, è stata (it was), sì infatti (yes indeed), questo è (this is), okay allora (ok then), ah ah, va benissimo (thats great), mi conferma (I confirmed), di nulla (nothing), non lo so (I don not know), conferma che (confirms that), sì esatto (yes right), ci mancherebbe (God forbid)
Competitive	no no, no no no, non è (it is not), c è (there is), ho capito (understood), un attimo (one moment), no non (not), io non (I do not), ma non (but not), eh ma (yeah but), sì sì (yes yes), ma io (but I), no ma (no but), mi dà (he gives me), sì ma (yes but), mi scusi(excuse me), non mi (I do not), no signora (no madam/lady), no perché (no because), mi dà il (gives me), io ho (I have), però io (but I), non è possibile (it is not possible)

Lexical Evidence *The token based* investigation includes n-gram and the word-cloud. Both approaches are frequency based analysis. In Table 3.3, the few top ranked bi- and tri-grams of each competitive and non-competitive classes are presented and the most frequent uni-grams are presented using word cloud as shown in Figure 3.10 and 3.9.

From the token based analysis, it is observed that lexical selection of the speaker may differ depending on their attitudes towards competitiveness. Comparing the frequencies of token for each class, the statistical significance over the observed differences with a two-tailed two-sample t-test and $p = 0.1$ are tested.

The findings suggests that in non-competitive instances, most frequent words indicates that the intervening speaker shares the opinions of the other speaker. For example, Italian words and phrases like "bene" ("well"), "ho capito" ("I have understood"), "certo" ("sure") are very frequent in *Ncm*.

On the contrary, in the competitive distribution, occurrences of words and phrases like "no", "ma" ("but"), "mi scusi" ("excuse me") in Italian may play the role of discourse markers usually used to emphasize a discordant point of view.

The findings implicates that for the non-competitive overlaps "sì" ("yes") is the most frequently used word to start an overlap, whereas the word "no", either alone or associated with adversative conjunctions like "ma" ("but"), is the most frequently used for competitive starts.

Part-of-speech (POS) Analysis of the Start token To observe the most frequent token to initiate the overlap along with what group of part-of-speech the token belongs, the lexical sequences are automatically annotated with Part-Of-Speech tags using Tree Tagger [150]. For the frequency, it is observed that for competitive overlaps the most frequent starting token with pos tag are: *ADV_no, CON_e, INT_eh, CON_ma, VER:pres_è, ADV_allora, ADV_non, ADV_sì, PRO:pers_io, CON_perché, PRO:pers_mi, NOM_signora*.

As for non-competitive overlaps, the most frequent starting tokens are: *CON_e, CON_sì, INT_eh, ADV_sì, ADV_no, INT_ah, VER:pres_è, VER:pres_va, ADV_non, ADV_allora, NOM_okay, DET:def_il, PRO:pers_mi, PRO:rela_che, PRO:pers_io, PRE_a, ADV_quindi, CON_perché*. The description of the tagset is given in Table 3.4.

Feature Ranking Since frequency based analysis does not entail that top ranked tokens or ngrams are important. Therefore, a feature selection followed ranking based approach has been investigated to find the tokens containing most important information. For this analysis, trigrams are extracted from manual transcriptions, in order to understand whether there are any linguistically relevant contextual manifestations for competitive expression. For the analysis of the lexical features, a Relief feature selection algorithm [151] has been used. Prior to the feature

Table 3.4: Most frequent POS tags found in starting token of the overlap, with its description.

POS	**Description**
ADV	adverb
CON	conjunction
INT	interjection
VER-pres	verb present
PRO-pers	personal pronoun
NOM	noun
DET-def	definite article
PRO-rela	relative pronoun

selection, the raw lexical features has been transformed into bag-of-words (vector space model). Then, Relief feature selection algorithm has been applied and ranked the features, based on the score computed by the algorithm.

It is observed that token sequence (tri-grams) which are unique for competitive (i.e. are not present in any non-competitive instances) that carries important information are: "attimo un attimo" *("wait one moment")*, "ma quello" *("but that")*,"scusi un attimo" *("Exceuse me for a moment")*, "scusi un" *("Exceuse me")*, "e cosa devo" *("and what should I")*, "e cosa" *("and what")*, "signora l" *("madam I ..")*, "non ho non" *("I have not ..")*, "ho non ho" *("I have not ..")*.

Similarly for non-competitive unique ranked features are : "grazie a lei" *("thanks to you")*, "grazie a" *("thanks to")*, "auguro una buona" *("have a nice ..")*, "le auguro" *("I wish you")*, "va bene va" *("all right all")*, "le auguro una" *("I wish you a")*, "che oggi è" *("it is today")*, "anche a lei" *("you too")*, "salve" *("Hello")*, "sì mi dica" *("Yes tell me")*, "bene okay" *("well okay")*, "di nulla" *("nothing")*.

3.1.4.3 Speaker Distribution

The **SISL** Human-Human Dyadic Conversation corpus consists of a total of 1403 agents providing service on the whole 10K dataset. However, the subset of the corpus contains 565 conversations from 408 speakers. The Figure 3.11 presents the distribution of calls received per agent. It is observed that only 30.4% of the speakers received more than one inbound call thus resulting a skewed distribution of the call received per agent.

The metadata of the corpus contains no information of the customer. Thus the lack of information about the customer identity prevent such analysis on customers repeated the call. Therefore, we do not present such information.

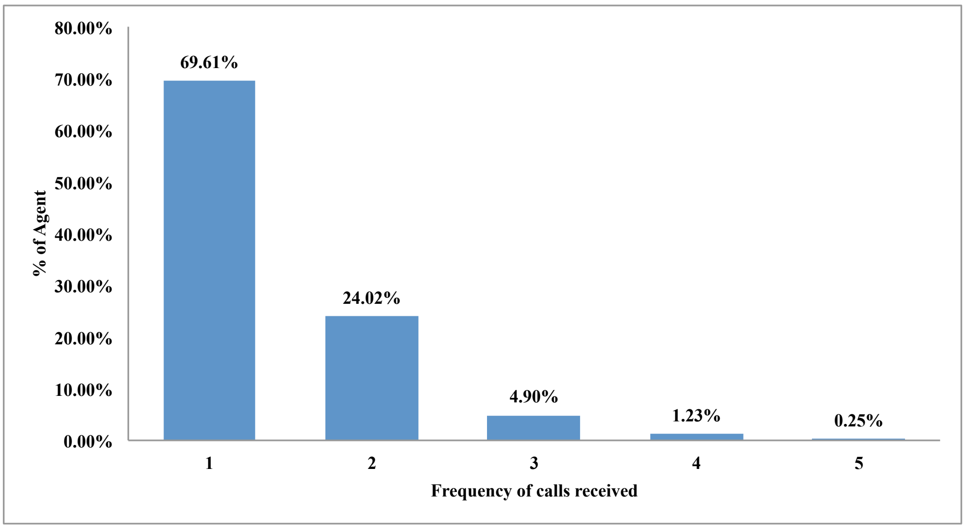


Figure 3.11: Distribution of the percentage of call received per agent in the corpus.

3.1.4.4 Gender Distribution

The distribution of male-female in the SISL conversational discourse corpus for both agent and customer channel are presented in Figure 3.12. Given this distribution of male and female conversations in the corpus, a general automatic gender identification model can be designed.

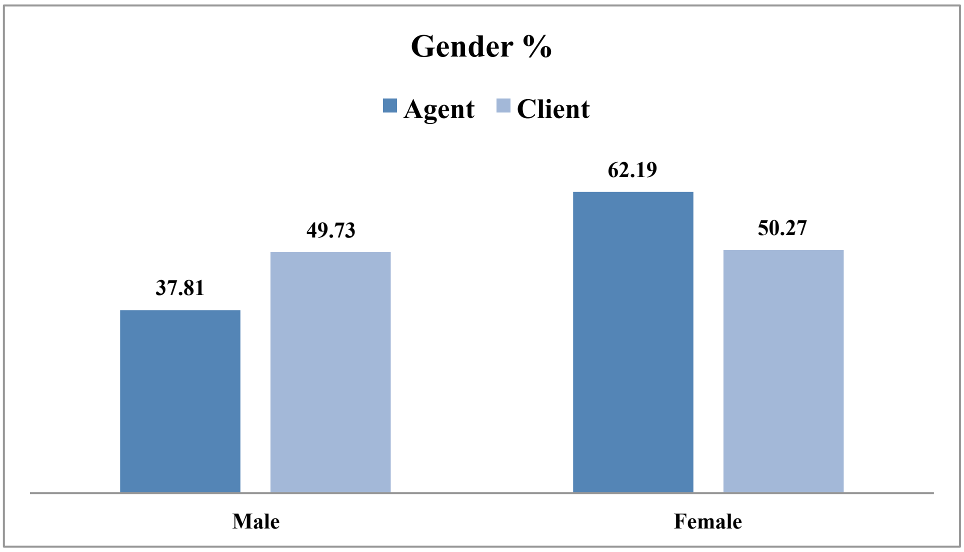


Figure 3.12: Gender distribution on the agent and customer side of the datasets.

3.2 LUNA Italian Corpus

The Italian LUNA Corpus [152] is a collection of 723 human-machine (approximately $4,000$ turns and 5 hours of speech) and 572 human-human (approximately $26,500$ turns and

30 hours of speech) spontaneous dialogs in the hardware/software help desk domain[3]. The dialogs are conversations of the users involved in problem solving.

While the human-human dialogs are recording of the real user-operator conversations, the human-machine dialogs are collected using Wizard of Oz (WOZ) technique: the human agent (wizard) reacting to user requests is following one of the ten scenarios identified as most common by the help desk service provider. Text-to-Speech Synthesis (TTS) was used to provide responses to the users. Through out this study, we used LUNA human-human corpus for modeling dialog act segmenter and classifier. For the task, a subset of 50 dialogs are annotated with dialogue acts using the annotation scheme given in the following sections.

3.2.1 LUNA DA annotation scheme

The LUNA DA annotation scheme [153] was inspired by DAMSL [154], TRAINS [155], and DIT++ [156]. The most common 15 dialog acts from these taxonomies are grouped into three categories [152]: *Core Dialog Acts* (8) are main actions in the dialog, such as request of information, response, or performing the task; *Conventional/Discourse Management Acts* (4) are utterances such as greetings, apologies, etc. whose function is to maintain general dialog cohesion; *Feedback/Grounding Acts* (3) are utterances whose function is to acknowledge, provide feedback, or just time fillers; and *Others* (1) to capture the rest. The unit of annotation for dialog acts in LUNA Corpus is an utterance. However, due to the overlapping turns (both speakers speaking), an utterance can span several turns.

3.2.1.1 Problem with current LUNA-DA annotation

In the absence of a single commonly accepted standard, spoken corpora, such as LUNA often adapt existing domain independent annotation schemes like DAMSL [154], TRAINS [155], DIT++ [156] to task-specific needs; thus, creating incompatible annotations. This limits the re-usability of the corpora and thus the models created using it.

Recently accepted international ISO standard for DA annotation – Dialogue Act Markup Language (DiAML) [1, 157], as shown in Figure 3.13 – could serve as a *lingua franca* for cross-corpora DA mapping. However, such mappings might require significant amount of manual re-annotation effort.

3.2.1.2 Mapping LUNA to ISO Standard

Full description of the DiAML annotation scheme [1] is out of the scope of this thesis. Rather we focus on the DA tag set and dimensions. The DiAML annotation scheme consists of 56 DA tags (communicative functions), organized into 9 dimensions: 26 general (applicable to any dimension) and 30 dimension specific (see Table 3.5, ISO column).

[3] The corpus is available for research purposes from http://sisl.disi.unitn.it

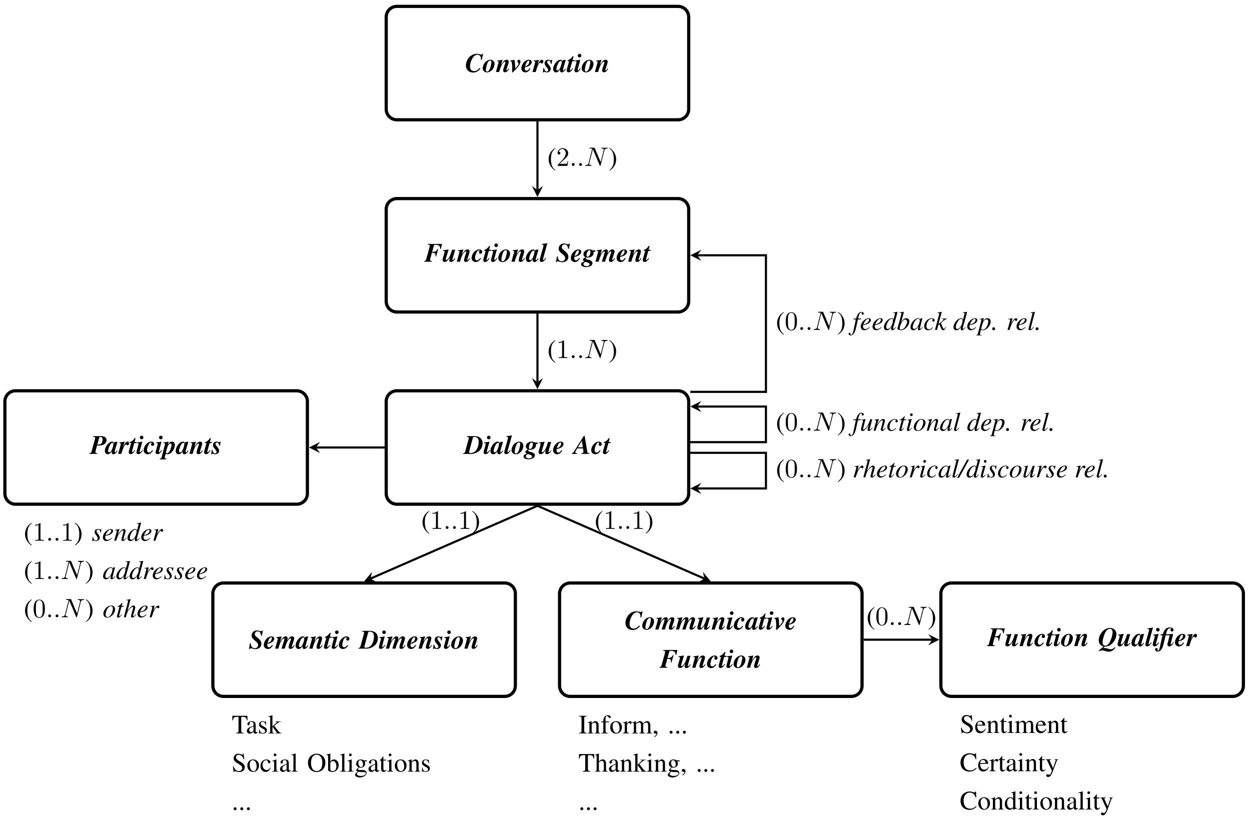


Figure 3.13: Conversation as a dialog act annotation model of ISO 24617-2. A conversation consists of several *functional segments* (marked as (2...N) for number) – minimal spans of behavior (verbal or not) that have a *communicative function* (56) – in multiple *semantic dimensions* (9) (segments are dimension specific and can overlap). Thus, a *dialog act* consists of a *communicative function* - *semantic dimension* pair and is defined as having *participants* such as *sender* and one or more *addressees*. *Function Qualifiers* are describing how *communicative function* is performed: e.g., with positive sentiment or uncertainty. *Functional* and *feedback dependency relations* connect a *dialog act* with previously identified conversation units. *Rhetorical/discourse relations* possibly relate *dialog acts* and *semantic content* to other *dialog acts* or *semantic content* units of a conversation.

Table 3.5: Mapping LUNA dialogue acts to DiAML ISO Standard 9 dialogue act dimensions and communicative functions with counts per dimension.

Dimension	**ABBR**	**ISO**	**LUNA**
General (Task)	G	26	8
Social Obligations Management	SOM	10	4
Auto-Feedback	AutoFb	2	3
Allo-Feedback	AlloFb	3	
Time Management	TimeM	2	
Turn Management	TurnM	6	–
Discourse Structuring	Disc	2	–
Own Speech Management	OSM	2	–
Partner Speech Management	PSM	3	–
Total		56	15

The issues of converting DAMSL-based corpus to the ISO standard were addressed by [158] and [159]. Following the re-annotation methodology outlined in [158] we mapped LUNA DAs to DiAML. LUNA contains only 15 tags compared to DiAML's 56, and most of the relations in the mapping are one-to-many. Even though, some of these relations can be disambiguated with respect to context [160] (e.g. if the DA in the previous turn is *Info-Request* and the current DA is *Yes-Answer*, there is a high chance that the former maps to *Propositional Question* and the latter to *Confirm*), since both relations are one-to-many, such mapping is error prone. Thus, automatic mapping is manually examined. Due to data distribution and for the consistency with the legacy annotation, we did not annotate all the dimensions: Discourse Structuring, Speech and Turn Management dimensions were mapped to *Other*.

3.2.1.3 Re-Annotation Methodology

In [159] the authors list segmentation differences as one of the issues of converting DAMSL-based annotation to ISO standard. While in the former the unit of annotation usually corresponds to a turn, in the latter it is a *functional segment* that can be shorter or longer than turn. In LUNA, on the other hand, the unit of annotation was considered to be an utterance, which is similar to turn, ignoring the other speaker barge-ins. Consequently, re-annotation procedure also included re-segmentation.

As the first step of the re-annotation effort, a linguist annotated a limited set of LUNA dialogs to get accustomed to the procedure. Since the legacy annotation was performed by a different person, to ensure the consistency, the annotator performed an **unsupervised** annotation (15 dialogs) of the LUNA corpus with new DiAML scheme in the dimensions selected previously. This set of 15 dialogs is used to compute the inter-annotator agreement between the

Table 3.6: Mapping from LUNA DA to ISO dimensions and communicative functions. Note that most of the relations are one-to-many and frequently are cross-dimension.

LUNA DA	**ISO DA**
Core Dialogue Acts → General/Task (**G**)	
Info-Request	Question, Set-Question, Choice-Question, Propositional-Question, Check-Question
Action-Request	Instruct, Suggest, Request
Yes-Answer	Confirm, Accept-Offer, Accept-Request, Accept-Suggest
No-Answer	Disconfirm, Decline-Offer, Decline-Suggest, Decline-Request
Answer	Address-Offer, Address-Request, Address-Suggest, Answer, Correction, Disagreement, Agreement
Offer	Offer, Promise
Report-On-Action	Inform
Inform	Inform, SOM:I-Self-Introduction, SOM:R-Self-Introduction
Conventional Dialogue Acts → Social Obligations Management (**SOM**)	
Greet	I-Greeting, R-Greeting
Quit	I-Goodbye, R-Goodbye
Apology	Apology, Accept-Apology
Thank	Thanking, Accept-Thanking
Feedback/Turn Management Dialogue Acts	
Clarif-Request	AlloFb:Positive, AlloFb:Negative
Ack	AutoFb:Positive, AutoFb:Negative
Filler	TimeM:Stalling, TimeM:Pausing
Non-Interpretable/Non-Classifiable Dialogue Acts	
Other	Other

‘legacy’ and the ‘ISO’ annotator.

For the agreement calculation ISO DAs are mapped to the ‘legacy’ DAs. Due to segmentation differences the two annotations are first aligned with respect to the Levenshtein distance and F-measure is computed with respect to alignment errors [161]. Since ‘legacy’ annotation unit covers several functional segments, insertion errors are ignored. The overall agreement between the ‘legacy’ and ISO annotators is $F_1 = 0.68$.

As the second step, we have annotated 10 dialogs from an SISL corpus described in Section 3.1. The activity has two goals: (1) to check the dimension and DA distributions cross-domain and (2) for later cross-domain evaluation on supervised classification task. The resulting annotation was compared to the random 10 dialogs from LUNA annotation, from the previous step. The dimension and communicative function distributions were observed to be similar.

As the third step, the remaining LUNA dialogs are automatically re-annotated using the mapping described in Section 3.2.1.2, which was refined through steps 1 and 2. The annotator's job at this step was to segment the turns into functional units and to disambiguate the labels. This step is a **supervised** annotation, and automatic mapping is provided to ensure consistency with the **unsupervised** annotation, while reducing the amount of the required effort. The distribution of the resulting annotation into dimensions is given in Table 3.7.

Table 3.7: Distribution of dialogue acts in LUNA corpus and the subset of SISL corpus (**SISL**). The counts are given per annotated dimension and in total.

Dimension	**LUNA** (50)		**SISL** (10)	
General (Task)	1,950	(59.7%)	911	(61.9%)
Social	250	(7.6%)	99	(6.7%)
Auto-Feedback	673	(20.6%)	278	(18.9%)
Allo-Feedback	44	(1.3%)	11	(0.75%)
Time Management	114	(3.5%)	68	(4.62%)
Other	237	(7.3%)	105	(7.13%)
Total	3,268	(100.0%)	1,472	(100.0%)

3.3 Summary

This chapter focuses on describing the SISL conversational discourse corpus in details, which is collected from naturally occurring conversations in call centers. As the experimental design aimed to collect ecologically valid data, therefore, no knowledge has been given to the subject regarding the experiment. To model automated systems, the chapter discussed a detailed schema for the design and evaluation of the annotation guideline for overlap discourse. The labels in the annotation includes overlap discourse label – competitive *vs* non-competitive tag among other affective behavioral labels. A detailed corpus analysis is then presented, which includes the patterns of turn-takings, and other turn types observed in the dataset, followed by durational distribution study of turn-taking signals, pauses, overlaps among others. The chapter also presented in depth study of linguistic analysis of competitive and non-competitive overlap segments.

In addition to the SISL corpus, the chapter also focused on the LUNA human-human italian corpus which is used later in the dissertation to create a dialog act segmenter and classifier. The 'legacy' dialog-act annotation of LUNA-HH was designed keeping in mind a task-specific needs; which limits the re-usability of the corpus and the model created with it. Thus the chapter describes the mapping of the ‘legacy’ dialog act annotation to the ISO standard DiAML scheme following the transfer of annotation and annotation of SISL corpus using the new DA annotation schema. The dataset described is later used for the experiments, discussed in the subsequent chapters of the dissertation.

Chapter 4

Unsupervised Study of Overlaps

In this chapter, we are interested in understanding speech overlaps and their function in human conversations. The characterization of overlaps based on timing, semantic and discourse function requires an analysis over a very large feature space. In this study, the corpus of overlapped speech segments was automatically extracted from human-human spoken conversations using a large vocabulary Automatic Speech Recognizer (ASR) and a turn segmenter. Each overlap instance is automatically projected onto a high dimensional space of acoustic and lexical features. Then, we used unsupervised clustering to find the distinct and well-separated clusters in terms of acoustic and lexical features. We have evaluated recognition and clustering algorithms over a large set of real human-human spoken conversations. The clusters have been comparatively evaluated in terms of feature distributions and their contribution to the automatic classification of the clusters.

4.1 Introduction

To understand and analyze different categories of speech overlaps and their function in human conversations, an unsupervised technique is being first applied. The motivation behind the approach is to observe what patterns of overlapping speech is present in the data set before involving any human judgments.

We have analyzed acoustic and lexical features that discriminate individual clusters and compared them with the characteristics of features mentioned in previous literature on distinguishing overlaps. In contrast with that of previous studies, the contribution of this study differs in a number of ways:

- Designing the corpus of speech overlaps using an automated approach.
- Extraction of a large set of acoustic and lexical features.
- Investigation of speech overlaps using unsupervised clustering.

Contents published in:
Shammur Absar Chowdhury, Giuseppe Riccardi, Firoj Alam, *Unsupervised Recognition and Clustering of Speech Overlaps in Spoken Conversations,* Workshop on Speech, Language and Audio in Multimedia (SLAM 2014), Penang, Malaysia.

- Analysis of discriminative characteristics of speech overlaps over acoustic and lexical features.

This chapter is organized as follows. A description of data preparation procedures in Section 4.2. In Section 4.3, we discuss the experimental methodology used in this study. We present an analysis of our findings in Section 4.4 and provide summary of the study in Section 4.5.

4.2 Data Preparation

The overlapped segments are detected using start and end time of each speaker's turn and for each word unit within that turn. To get each speaker's turns we passed the conversation to an automatic turn segmenter [134] followed by a large vocabulary Italian ASR, as described in Section 7.2.1.2 to get automatic transcription from the corresponding turns.

Then, the overlapping turns are detected, where each turn has an alignment between the automated word level transcriptions and the speech recording. Using the overlapping turns, the words that are within the overlaps are also detected. Then, the boundary of overlaps is extracted using the start time of the first word in the overlap to the end time of the last word. The details of the process is presented in Figure 4.1.

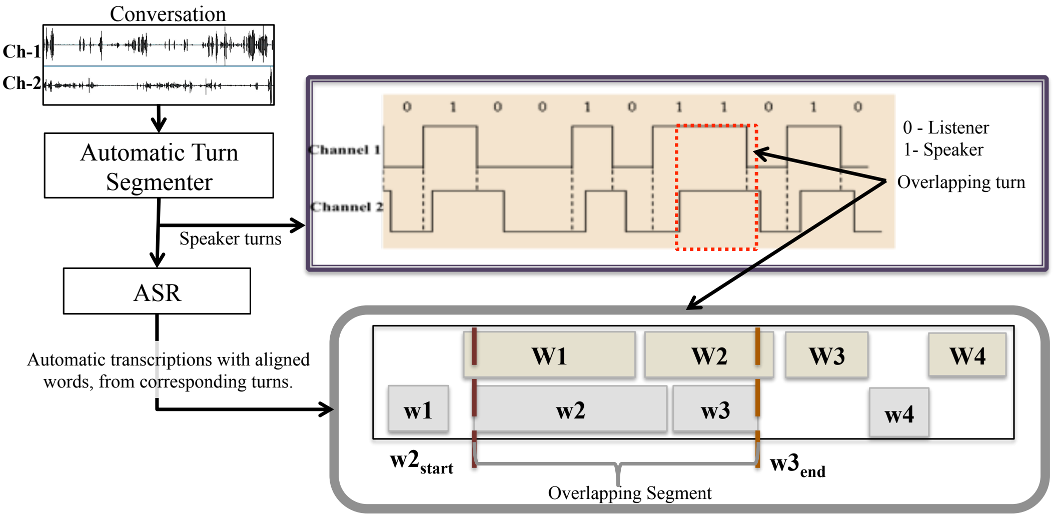


Figure 4.1: Unsupervised extraction of overlapping speech segments

Following this approach, 25132 instances of overlaps (average duration of 0.52s) are extracted from 515 conversations, where the duration of overlap is 3 hours and 38 minutes and total the duration of speaking time is 41 hours and 52 minutes.

4.3 Methodology

The work-flow of clustering and feature analysis is shown in Figure 4.2. The overlap segment's components are shown for each channel. We extracted acoustic and lexical features from both channels. We then merged the acoustic features from each channel to create a combined acoustic feature vector. We then followed the same procedure to create a combined lexical feature vector by merging the lexical features of each channel.

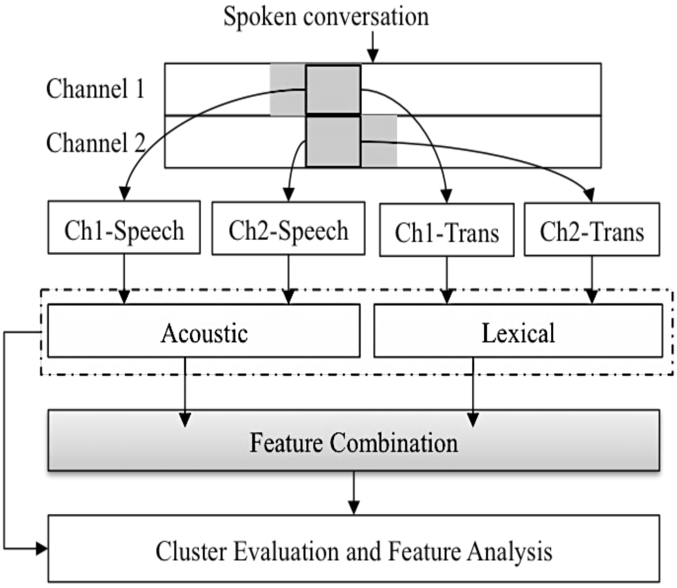


Figure 4.2: System architecture for overlap segment clustering. *-Trans: channel's transcription

4.3.1 Feature Extraction

4.3.1.1 Acoustic Features

We extracted a large number of acoustic features, motivated by their successful utilization in the paralinguistic task discussed in [162, 163]. The process extracts a large number of Low-Level Descriptors (LLD) and then projects them onto statistical functionals using openSMILE [164]. These low-level features were extracted with approximately 100 frames per second, with 25 milliseconds per frame. The 39 low-level features include frame energy, loudness, mel-frequency cepstral coefficients (MFCC1-12), voice quality, fundamental frequency (f0), exponentially smoothed f0-envelope, jitter-local, differential of jitter, shimmer-local, logarithmic harmonics-to-noise ratio (HNR) computed from auto-correlation, zero crossing rate of time signal, formant frequencies (f0-f3) and spectral features with different bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), roll-off points (25%, 50%, 70%, 90%), centroid, flux, max-position and min-position. Delta and acceleration coefficients of these features have also been extracted. These low-level acoustic features were then projected onto 24 statistical functionals, which included range, absolute position of max and min, linear and quadratic regression coefficients and their corresponding approximation errors, moments-centroid, variance, standard deviation, skewness, kurtosis, zero crossing rate, peaks, mean peak distance, mean peak, geometric mean of non-zero values, and number of non-zeros. As mentioned earlier, the overlap segment's components appear in two channels. Therefore, we extracted the same number of features from each channel. The size of the feature vector in a single channel is computed as: $(39 + \triangle 39 + \triangle \triangle 39) LLD$ × 24 functionals = 2808. A total of 5616 features were obtained after merging the feature vectors from both channels.

4.3.1.2 Lexical Features

Lexical features were extracted from automatic transcription using the ASR explained in Section 7.2.1.2. The lexical features were transformed into a bag-of-words (vector space model) [165]. The idea of this approach is to represent the words into numeric features. For this study, we extracted bigram features and selected the top 2000 frequent features to reduce features dimension. The frequencies in the feature vector were then transformed into tf-idf values – the product of the logarithmic term frequency (tf) and inverse document frequency (idf).

4.3.1.3 Feature Combination

Although feature combination has been widely used in other speech processing tasks, its relative contribution greatly varies depending on the data and experiments. For this study, we also analyzed the contribution of feature combination. As shown in Figure 4.2, after extracting acoustic and lexical features we merged the feature vectors into a single vector and then used that for clustering.

4.3.2 Dimensionality Reduction

Since the complexity of any pattern recognition algorithm also depends on the number of features, we reduced the feature space to limit the complexity and number of free parameters. A typical approach for feature reduction is to map higher dimensional feature spaces into lower dimensional spaces, while maintaining as much of the information as possible. In our study, we have used Principal Component Analysis (PCA), which is one of the fundamental feature reduction methods. After transforming the feature space using PCA, the usual approach is to take the leading p components that explain the data with 95% variance [166]. However, as a baseline study we took the leading p components with 99% variance. The value of p varies for different features sets. Hence, we reduced 63% acoustic, 11% lexical and 59% acoustic+lexical features. The reason for obtaining a minimal reduction with lexical features is the weak correlation with feature dimensions and sparseness.

4.3.3 Clustering Experiments and Results

To find well-separated clusters of speech overlaps in our dataset, we used K-means [167] where data points are classified as belonging to one of K groups. For reproducibility and transferability, we used Weka's implementation [168]. Members of the clusters are determined by comparing the data point with each group's centroid and assigning it to the nearest one. The reason for choosing K-means is that it is highly recommended for large datasets [169, 170] and is one of the simplest methods. However, one of main limitations of K-means is in choosing the value of K in prior. Therefore, we used cascaded K-means, which uses Calinski-Harabasz CH [171] criterion to determine the best value of K that represents the dataset.

For each value of K, cascaded K-means calculates the between-group dispersion, BGSS, within-cluster sum of squares, WGSS, and Calinski-Harabasz (CH) value or index, using Equations 4.1-4.3.

$$BGSS = \sum_{k=1}^{K} n_k \left|\left|G^k - G\right|\right|^2 \tag{4.1}$$

$$WGSS = \sum_{k=1}^{K} \sum_{k \in I_k} \left|\left|M_i^k - G^k\right|\right|^2 \tag{4.2}$$

$$CH = \frac{(N-K) * BGSS}{(K-1) * WGSS} \tag{4.3}$$

K is the number of clusters, N is number of observations, G^k is the barycenter of cluster C_k, G is the barycenter of the entire dataset, n_k is the number of elements in the C_k. I_k is the set of the indices of the observations belonging to C_k, and M_i is the ith observation of element in C_k.

Figure 4.3 shows the values of CH corresponding to the number of clusters K. The results of our experiments are shown in Table 4.1. The optimal number of K using acoustic and acoustic+lexical feature sets is 2, as can be seen in Figure 4.3. The clustering difference between the two feature sets is minimal. The optimal number for k, which we obtained using lexical features, is 4. The CH value for the cluster is significantly smaller compared to the CH values with acoustic features. The minimal separability of lexical features could be due to the sparseness and recognition error of the ASR.

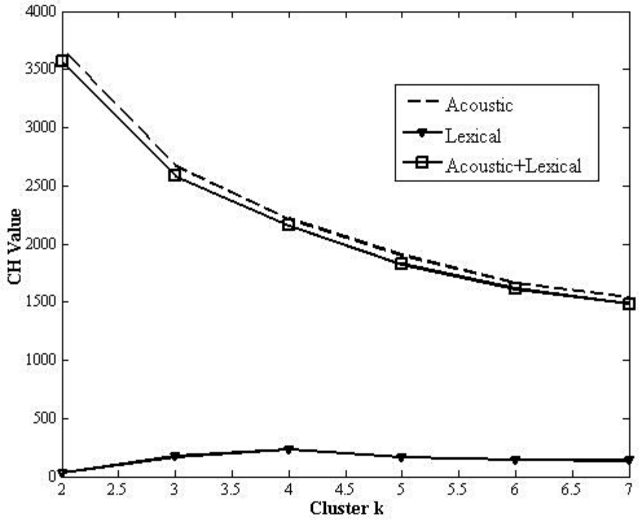


Figure 4.3: Calinski-Harabasz (CH) value for cluster decision

Table 4.1: Cluster evaluation using different feature sets: K - number of cluster, CH - Calinski-Harabasz value, W – weighted within-cluster sum of squares, B – between group dispersion

Feature Set	**K**	**CH**	**W**	**B**
Acoustic	2	3681.12	42.31	155749.22
Lexical	4	232.66	1.38	320.04
Acoustic +Lexical	2	3568.28	43.71	155985.56
Acoustic + lexical with PCA	2	2754.82	21.28	58631.95

We applied PCA feature reduction method to the acoustic and acoustic+lexical feature sets and with reduced dimensions we obtained 2 clusters for each set. We then calculated the cluster agreement between the feature sets using kappa statistics [172]. We found that the agreement between the original and reduced dimensions is fairly reasonable. The agreements for the acoustic and acoustic+lexical feature sets are 92% and 91%, respectively. This indicates that feature reduction is important in reducing computational cost with minimal loss of information. To check the validity of the clusters using cascaded K-means, we used another well-known clustering algorithm – Spectral Clustering [173, 174]. Using this algorithm, we also found 2 clusters. Then, we compared the clusters generated by these two clustering algorithms for acoustic and acoustic+lexical feature sets using kappa measure. The agreements between the two algorithms on acoustic and acoustic+lexical feature sets are 90% and 87%, respectively. Detailed results of Spectral Clustering algorithm have not been included in this dissertation, in favor of brevity.

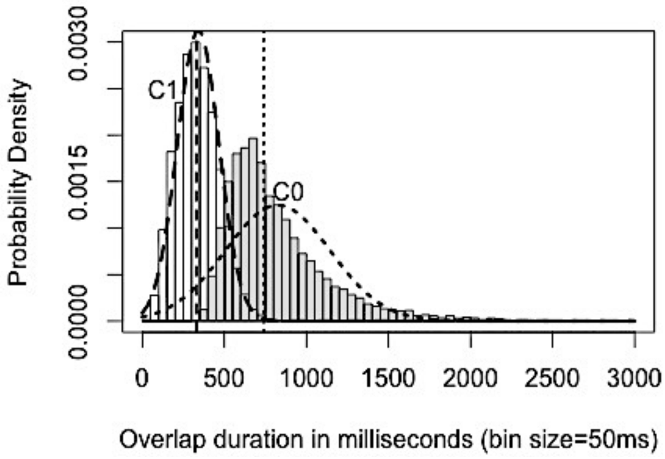


Figure 4.4: Overlap duration distribution of the two clusters

4.4 Findings

We analyzed different features based on the cluster decision of acoustic features, where cluster 0 (C0) and cluster 1 (C1) contain 37% and 63% of overlapping instances, respectively. The members of the clusters were analyzed using duration distribution of speech overlaps and top-ranked acoustic features. Based on this cluster decision, we extracted and analyzed lexical features. In doing so, we compared our observations with those of previous studies to determine whether our clusters resembled competitive or non-competitive overlaps.

Figure 4.4 shows the distribution of overlap durations for C0 and C1. It can be seen that C1 contains instances of overlaps with short durations whereas C0 has instances with comparatively long durations. The authors in [78] and [71] state that non-competitive overlaps tend to be shorter and resolved soon after the second speaker has recognized the overlap, whereas competitive overlaps are persistent because speakers keep on speaking despite overlapping. Therefore, it can be inferred that competitive overlaps have longer durations than non-competitive overlaps. With duration a key distinguishing feature, we observed that there is clear distinction between C0 and C1. We also observed that the median duration distribution of C1 is very close to the minimum distribution of C0. The minimum, median, third quartile and maximum durations, in milliseconds, of the clusters are C0 - 300, 740, 950, 3590 and C1 - 40, 330, 430, 850, in that respect. The vertical lines in the figure indicate the median of the respective distributions.

For the analysis of acoustic features we used Relief feature selection technique [151] to rank the features, which has been useful in paralinguistic task [175]. It calculates the weight of the features based on the nearest instances of the same and different classes. The top-ranked low-level acoustic features include logarithmic Harmonic to Noise Ratio (logHNR), f0 envelope, shimmer-local, jitter-local, and spectral features, with their delta and acceleration coefficients, whereas the statistical functionals include range, standard deviation, mean of peak, linear regression with error coefficients, and centroid. Figure 4.5 shows some of the top-ranked low-level features projected on statistical functionals as described in Table 4.2. From the Figure 4.5, it can be seen how the two clusters differ in their distributions – the mean values for C1 are always lower than those for C0.

Table 4.2: Acoustic features and their description

Feat.	Description
F1	Logarithmic harmonic to noise (logHNR) ratio with delta coefficient projected to statistical range
F2	logHNR projected to statistical range
F3	logHNR with delta coefficient projected to statistical mean of peak
F4	logHNR projected to statistical standard deviation
F5	logHNR with linear error computed as the difference of the linear approximation and the actual contour
F6	f0 envelope projected to statistical mean of peak
F7	Local shimmer with centroid
F8	Local jitter with centroid
F9	f0 envelope projected to geometric mean of non-zero values
F10	First formant with number of non-zero values
F11	Loudness with number of non-zero values
F12	log energy with delta coefficient projected to non-zero values

Some of the voice quality features show significant difference in their distribution between two clusters. This indicates that these features play an important role in detailing the patterns in each cluster. logHNR is a feature which is widely used to analyze disorders such as hoarseness and depression. However, we observed that this feature has not been applied before to the analysis of overlaps. Other commonly used features for categorizing overlaps are f0, loudness and energy. By observing the values of F10 in Figure 4.5, it can be inferred that the mean value of C0 is higher than that of C1. This inference is extended to apply to the values of F11 as well. This, coupled with observations from previous research, provides the grounds for the conclusion that our C0 exhibits patterns similar to competitive overlaps. By studying the most frequent lexical features, it can be noted that filler and affirmative words are present in both clusters but that C1 has higher frequencies than C0 has. For example, the token "sì/yes" is present in C1 with a frequency of 2506, three times as much as that of the same token in C0. It can also be noted that, in comparison with C0, C1 has a homogenous lexicon, as demonstrated by the long tail of C0 in Figure 4.6.

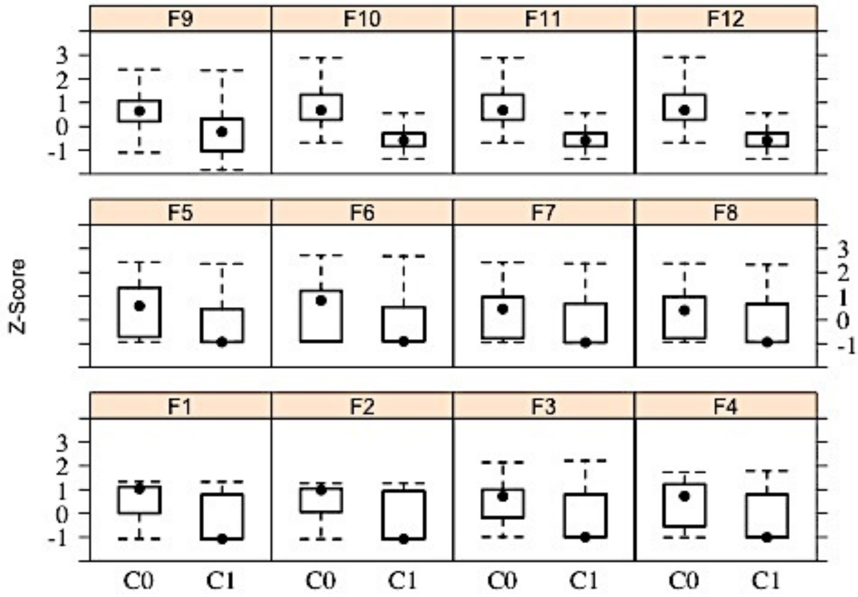


Figure 4.5: Selected acoustic features (F1 to F12) and their z-score distribution in C0 and C1. Box-plots, representing the mean, max, upper and lower inner fences of top ranked features. Outliers have been removed for readability.

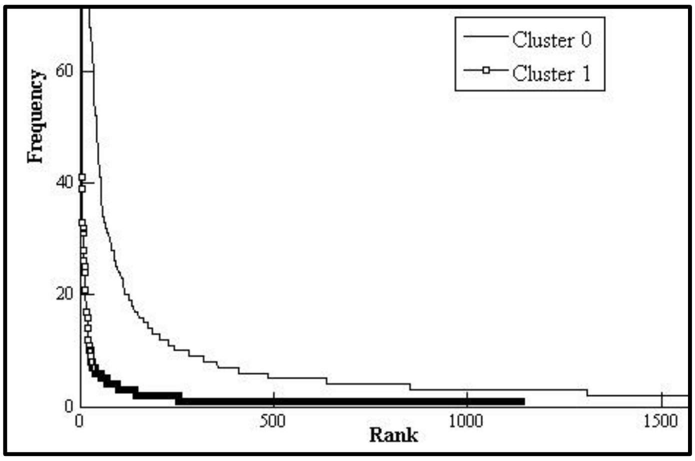


Figure 4.6: Zipf's plot with bigrams for C0 and C1 clusters. Frequency is plotted as a function of frequency rank.

4.5 Summary

In this chapter, we designed an automatic system that divides speech overlaps into two classes using unsupervised approach. We prepared our data with an automated manner, by cascading a turn segmenter and an ASR system. For clustering, we extracted a large number of acoustic features from overlapped segments, and lexical features from automatic transcriptions. Our findings suggest that acoustic features play a more important role than lexical features in discovering well-separated clusters. Voice quality features, especially logHNR, jitter and shimmer, are the most discriminative in clustering overlaps. Based on previous work and from our analysis, we found that instances of C0 have a higher chance of being competitive overlaps, while instances of C1 are more likely to be of non-competitive nature. Our observation of lexical features, which are obtained from the clustering decision of acoustic features, is that the frequencies of filler and affirmative words are higher in C1 than the frequencies of such words in C0. These features may help in supervised classification, which we investigate further in the next Chapter 5 with includes manual annotation and studying contextual features as well as understanding the roles of acoustic, lexical features and other contextual features and different classification techniques.

Chapter 5

Overlapping Speech Classification

Overlapping speech is one of the most frequently occurring events in the course of human-human conversations. Understanding the dynamics of overlapping speech is crucial for conversational analysis and for modeling human-machine dialog. Overlapping speech may signal the speaker's intention to grab the floor with a competitive *vs* non-competitive act. In this chapter, we propose and evaluate an annotation scheme for these two overlap categories in the context of spontaneous and *in-vivo* human conversations. We analyze the distinctive predictive characteristics of a very large set of high-dimensional acoustic, lexical such as Bag-of-ngrams and word embeddings and psycholinguistic features. In addition, we explored how they can be combined at the feature and decision level. In this chapter, we also studied the role of speakers, whether they initiate (*overlapper*) or not (*overlappee*) the overlap, and the context of the event. Using different feature sets and their combination we designed classifiers for overlap discourse using supervised linear and non-linear machine learning algorithms. The evaluation of the classifier has been carried out over call center human-human conversations. The results show that the complete knowledge of speakers' role and context highly contribute to the classification results, and performance increased when we combined the acoustic and lexical features. Our findings suggest that the lexical selections of the overlapper are good indicators of speaker's competitive or non-competitive intentions. We also observed that non-linear system, i.e., fully-connected feed-forward neural network, is best suited for different feature combination.

5.1 Introduction

Overlapping speech in spontaneous conversations is a naturally occurring phenomenon that may reveal speakers' attitudes and in particular their intentions with respect to the control of the turn-taking structure of the conversations. In order to design conversational competent spoken

Contents published in:
Shammur Absar Chowdhury and Giuseppe Riccardi, *A Deep Learning Approach to Modeling Competitiveness in Spoken Conversation*, in Proc. of ICASSP. IEEE, 2017, New Orleans, USA.
Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi, *The Role of Speakers and Context in Classifying Competition in Overlapping Speech*, in Proc. of Interspeech-2015, Dresden, Germany.
Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi, *Annotating and Categorizing Competition in Overlap Speech,* in Proc. of ICASSP. IEEE, 2015, Brisbane, Australia.

dialog systems, the understanding of the overlapping phenomena is crucial. Over the years many linguists, psycholinguists, and speech researchers have been studying these aspects of spoken interactions. In the conversational analysis tradition, overlaps have been considered as a violation of the fundamental rule [4, 35] of turn-taking, that is one person speaking at a time. Nevertheless, it has been shown that overlapping is pervasive in human conversations, for example, authors in [5] suggest that about 40% of all between-speaker intervals is overlap. Further studies focused on highlighting speaker's intentions behind the overlaps. For example, it has been proposed that speech overlaps are related to dominance or aggression towards the other speaker [7]. However, the picture is more complex. Not all the overlapping occurrences are related to competitiveness. They also support cooperativeness in the conversations, for example in providing the other speaker with cues about the mutual understanding [8].

In the computational literature, a widely accepted categorization of overlaps, over the years, is in between **Competitive (Cmp)**, *an attempt to grab the floor*, and **Non-Competitive (Ncm)**, *an attempt to assist the speaker for the continuation of the current turn*. Distinguishing the overlaps by the overlapper's intention is important for behavioral signal studies and for improving the quality of the spoken dialog system.

The aim of our study is to automatically classify competitive *vs* non-competitive overlaps. To classify overlaps, we focus on the followings research that:

- Design of a speech overlap annotation scheme of competitive *vs* non-competitive overlapping segments from the spoken conversation.

- Investigate different high dimensional features such as low-level acoustic features, lexical features (represented as Bag-of-ngrams, word-embeddings), psycholinguistic feature sets among others.

- Understands whether the competitiveness is best represented by the information from speakers' segments: overlapper, overlappee, context or their combination for different feature sets.

- Design computational models using Support Vector Machine (SVM) and also by exploiting many layers of the non-linear information processing (DNN) for high-dimensional features and their linear and deep space combinations.

Therefore for automatic classification, we investigated each speakers' segment enclosing overlaps and their combination for each feature set, followed by feature and decision level combination and use of different machine learning algorithm.

For the experiments, we analyzed a large dataset of Italian spoken conversations collected in call centers, with customers and agents engaged in problem-solving tasks. Unlike most of

Table 5.1: Dialog excerpts from the annotated corpus. Speech overlaps: bold form between [and], Hesitations: (.), Rising intonation: ↗, Falling intonation: ↘.

Ncm
S1: e quando [**cambiamo** ↘] (.)
S2: [**sì sì** ↘ **ho già detto**] di cambiare ↘
S1: and when [***we change*** ↘] (.)
S2: [***yes yes*** ↘ ***I have already told***] *to change* ↘
Cmp
S1: io non lo so [**io devo risparmiare**] ↘(.)
S2: [**ma no la tariffa**] è buona ↗
S1: I do not know [***I had to save***] ↘(.)
S2: [***but no the***] *rate is good* ↗

the previous studies, we investigate the role of speakers, context using acoustic, linguistic and psycholinguistic features on 15,899 instances of overlaps.

Examples of overlap instance of *Cmp* and *Ncm*, with their English translation, are shown in Table 5.1. As observed in the example, in the *Ncm* scenario, the intention of S2 (the agent) is to repeat something that was already mentioned in the previous turns of the dialog: S2 wants to reassure S1 that she agrees on something that was already on the floor of the conversation; both overlapping segments are uttered with falling intonation. On the contrary, in the *Cmp* scenario, S1 (the customer) is complaining about his problem, and he does not consider what S2 (the agent) claimed before. S2 has the intention to stop the complaints and to take the turn from the on-going conversation, S2's overlapping segment has rising intonation and pitch level.

The chapter is organized as follows. An overview of the description and preparation of the dataset in presented Section 5.2. In Section 5.3, we discuss the details of the different speakers' segments, and context that we extracted to investigate the role of speakers and the context in the classification task. Then, we discussed a detail description of the extracted features in Section 5.4. A brief description of the evaluation technique is presented in Section 5.5. Section 5.6 presents the experimental algorithms and techniques used for modeling the overlap discourse. After that, we presented the experiments, results, and analysis of our findings in Section 5.7. The chapter also includes details of the designed computational model for overlap discourse classification in mono channel scenarios, presented in Section 5.8. A summary of the chapter is then provided in Section 5.9.

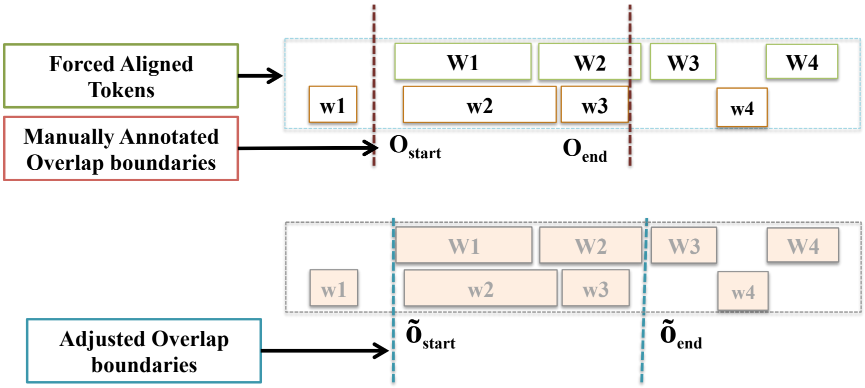

Forced Aligned Tokens
Manually Annotated Overlap boundaries
W1
W2
W3
W4
w1
w2
w3
w4
O start
O end
Adjusted Overlap boundaries
Õ start
Õ end

Figure 5.1: Forced alignment between the manual overlap boundary and word token inside the overlap. O represent the manual overlap boundary whereas Õ represent the adjusted overlap boundary.

Table 5.2: Description of the overlap classification data set and the distribution of competitive (Cmp) and non-competitive (Ncm) overlaps in training, development and test sets.

	Dialogs		Overlaps (Duration)		Cmp		Ncm	
Train	341	(60.35%)	9,537	(2h 55m)	2,379	(24.95%)	7,158	(75.06%)
Dev	109	(19.29%)	3,019	(1h 15m)	724	(23.98%)	2,295	(76.02%)
Test	115	(20.35%)	3,343	(0h 58m)	763	(22.82%)	2,580	(77.18%)
Total	565	(100.0%)	15,899	(5h 08m)	3,866	(24.32%)	12,033	(75.68%)

5.2 Data Preparation

For the study, we selected overlapping segments containing manual speech transcription. The exact boundary of the overlapping segments and their transcriptions are obtained using forced alignment, as shown in Figure 5.1, between the word-level transcriptions and the speech recording within the manual overlap segment boundary. For the alignment task, we used a domain-specific automatic speech recognizer [176], described in Chapter 7.

After forced alignment, we obtained 15,899 overlap segments, of a total duration of 5 hours and 8 minutes. For the experiments, we split our data into train, dev and test sets. Details of the dataset are shown in Table 5.2.

5.3 Role of Speakers and Contexts

In order to evaluate the roles of speakers and the context for classification of *Cmp vs Ncm*, we defined different speakers' segments enclosing overlaps, as shown in Figure 5.2, which are as follows:

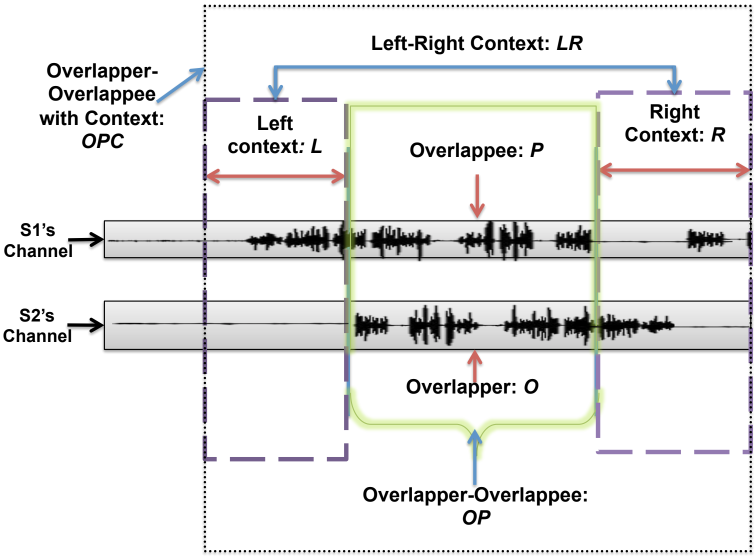


Figure 5.2: Example of different speakers' segments enclosing an overlap and their combination. S1-speaker 1; S2-speaker 2.

- Individual speakers' segments:
  - Overlapper ($\boldsymbol{O}$): overlap initiator
  - Overlappee ($\boldsymbol{P}$): current turn-holder
  - Left Context ($\boldsymbol{L}$): speakers' segment before the start of the overlap
  - Right Context ($\boldsymbol{R}$): speakers' segment after the completion of the overlap
- Combination of speakers' segments:
  - Overlapper-Overlappee ($\boldsymbol{OP}$)
  - Left-Right Context ($\boldsymbol{LR}$)
  - Overlapper-Overlappee with Left-Right Context ($\boldsymbol{OPC}$)

One of the main challenges of studying about context is to decide the window size, which gives us cues for classification/prediction. The author in [80], indicates that cues can be found in preceding segment ($\boldsymbol{L}$) of overlapping speech but they do not exceed a window of 0.2s. The study also showed that window of 0.3s is sufficient for the following context ($\boldsymbol{R}$).

From the manual annotation of the context of our data, we observed that the window size of the left context is $0.2s$ $\pm 0.15s$ and right context is $0.8s$ $\pm 0.5s$. We see that the window size of the right context varies a lot compared to the left context, which opens an avenue for further research. For this study, we used a window size of 0.2s and 0.3s, containing speech, for the left and the right context respectively, motivated by [80].

The left context ($\boldsymbol{L}$), in Equation 5.3, is defined by linearly merging speakers' channels where the information for speaker1's ($\boldsymbol{L_{s1}}$) and speaker2's ($\boldsymbol{L_{s2}}$) channels, are shown in Equations 5.1 and 5.2.

$$L_{s1} = \left\{l_{s1}^{1}, l_{s1}^{2}, ..., l_{s1}^{m}\right\} \tag{5.1}$$

$$L_{s2} = \left\{l_{s2}^{1}, l_{s2}^{2}, ..., l_{s2}^{m}\right\} \tag{5.2}$$

$$L = \left\{l_{s1}^{1}, l_{s1}^{2}, ..., l_{s1}^{m}, l_{s2}^{1}, l_{s2}^{2}, ..., l_{s2}^{m}\right\} \tag{5.3}$$

The same procedure is used to design the feature vector for the right context ($\boldsymbol{R}$).

The $\boldsymbol{OP}$ is designed by merging the overlapper ($\boldsymbol{O}$) and overlappee ($\boldsymbol{P}$) and the merged new feature vector is $\boldsymbol{OP = \{a_1, a_2, ..., a_m, b_1, b_2, ..., b_m\}}$. We used similar approach to merge the left ($\boldsymbol{L}$) and right ($\boldsymbol{R}$) context to form $\boldsymbol{LR}$.

In order to obtain the feature vector for Overlapper-Overlappee along with left and right context, we extracted features from both speakers' channels and merged them to obtain $\boldsymbol{OPC}$, as same as $\boldsymbol{OP}$. The boundary of the speaker channel is shown in Figure 5.2, which include overlap segments and contexts.

5.4 Features

5.4.1 Acoustic Features

The recent success of the use of low-level acoustic features and their projection onto statistical functionals has been applied to many paralinguistic tasks [148, 163, 175, 177]. The acoustic features are extracted using openSMILE [178] with frame size of 25 milliseconds and 100 frames per second. The low-level acoustic features include prosodic, spectral, voice quality, mfcc, and energy. These low-level features along with their derivatives are then projected onto 24 statistical functionals such as range, absolute position of max and min, linear and quadratic regression coefficients and their corresponding approximation errors, moments-centroid, variance, standard deviation, skewness, kurtosis, zero crossing rate, peaks, mean peak distance, mean peak, geometric mean of non-zero values and number of non-zeros [179]. The features are selected from the observation of a pilot study presented in Section 5.7.1.

The low-level features are extracted from both agent and customer channels. As shown in Equation 5.5, $CH1$ and $CH2$ represents the feature vectors of channel-1, and channel-2, respectively. We merge these feature vectors to create a new feature vector S that is used for the classification experiments.

$$\begin{aligned} CH1 &= \{a_1, a_2, ..., a_m\} \\ CH2 &= \{c_1, c_2, ..., c_m\} \end{aligned} \tag{5.4}$$

$$\begin{aligned} S &= \{CH1, CH2\} \\ S &= \{a_1, a_2, ..., a_m, c_1, c_2, ..., c_m\} \end{aligned} \tag{5.5}$$

Table 5.3: Low level acoustic features and Statistical functionals

Raw-Signal: Zero crossing rate
Energy: Root-mean-square signal frame energy
Pitch: F0 final, Voicing final unclipped, F0 final - nonzero
Voice quality: jitter-local, jitter-DDP, shimmer-local, log harmonics-to-noise ratio (HNR)
Spectral: Energy in bands 250-650Hz, 1-4kHz, roll-off-points (0.25, 0.50, 0.75, 0.90), flux, centroid, entropy, variance, skewness, kurtosis, slope band (0-500, 500-1500), harmonicity, psychoacoustic spectral sharpness, alpha-ratio, hammarberg-index
Auditory-spectrum: band 1-10, auditory spectra and rasta
Cepstral: Mel-frequency cepstral coefficitnts (mfcc 0-3)
Formant First 3 formants and first formant bandwidth
Statistical functionals
Relative position of max, min
Quartile (1-3) and inter-quartile (1-2, 2-3, 3-1) ranges
Percentile 1%, 99%
Std. deviation, skewness, kurtosis, centroid, range
Mean, max, min and Std. deviation of segment length
Uplevel time 25 and rise time
Linear predictive coding lpc-gain, lpc0-1
Arithmatic mean, flatness, quadratic mean
Mean dist. between peaks, peak dist. Std. deviation, absolute and relative range, mean and min of peaks, arithmatic mean of peaks, mean and Std. of rising and falling slope

5.4.2 Lexical Features

We extracted lexical features using the boundary of start and end of the corresponding speakers' segments with forced aligned reference transcription. The lexical features are transformed into a bag-of-words (vector space model) [180]. The idea of the approach is to represent the words into numeric features. For this study, we extracted trigram features, to use the contextual benefit of n-grams, and selected the top 5000 frequent features to reduce the feature dimension.

5.4.3 Word Embedding Features

Word embeddings, also known as context predictive model or neural language model, are new techniques to design distributional semantic models (DSMs), which differ from traditional DSMs where co-occurrence counts are used [181]. In word embedding, distributed vector representations are learned from a large corpus by neural network training, and represent them in a low dimensional continuous space. It has been proven that such representation better captures semantic and syntactic relationships [182].

For word embedding, we collected data from different sources such as PAISA[1], Republica[2], itwac[3], wikipedia dump [4] and automatic transcriptions from SISL behavioral corpus.

To design the word embeddings, we utilized gensim implementations [183], which is an implementation of Mikolov et al. [184, 185] word vector model. It contains both continuous bag-of-words (CBOW) and skip-gram algorithms. We designed our model using the CBOW approach with a size of the feature vector 500, a context window size 5, negative-sampling with a value of k=10. The resulting trained word-embedding model contains 6 billions words with a vocabulary of size 2.84 millions. We filtered the words with a frequency less than 5. The training file contains 2.4 billions of words.

5.4.4 POS Features

We automatically annotated Part-Of-Speech tags using Tree Tagger [150]. After that, we used similar approach of lexical features for the transformation and reduction of the POS feature set.

[1] The Paisà corpus is a large collection of Italian web texts. More details can be found on http://www.corpusitaliano.it/en/contents/description.html.

[2] The "la Repubblica" corpus is a very large corpus of Italian newspaper text (approximately 380M tokens)http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica

[3] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[4] Last accessed: June, 2015

5.4.5 Psycholinguistic Features

We automatically annotated Part-Of-Speech tags using Tree Tagger [150]. After that, we used a similar approach of lexical features for the transformation and reduction of the POS feature set.

5.4.6 Feature Combination

In addition to the individual feature set, we also evaluate the linear combination of acoustic and lexical features. Let $S = \{s_1, s_2, ..., s_m\}$ and $L = \{l_1, l_2, ..., l_n\}$ denote the acoustic and lexical feature vectors respectively. After the linear combination, the feature vector is represented by $Z = \{s_1, s_2, ..., s_m, l_1, l_2, ..., l_n\}$ with $Z \in R^{m+n}$.

5.5 Evaluation Methods

For the evaluation, there has not been any well-agreed metric for the task. Studies [78] used accuracy as an evaluation measure. It is evident that accuracy is not a good measure for imbalanced class distribution [186], therefore in our study, we considered to measure Precision (**P**), Recall (**R**) and (**F1**) in Equation 5.6-5.8.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5.6}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5.7}$$

$$F1 = 2 * \frac{\mathrm{P} * \mathrm{R}}{\mathrm{P}+\mathrm{R}} \tag{5.8}$$

where *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN).

As we want to evaluate our system considering both of the classes, we computed macro-averaged P_{avg} and R_{avg}, which is an average of P and R for both classes, respectively. Using P_{avg} and R_{avg} we calculated the $F1$ for the overall system. For the simplicity, we are only reporting the $F1$ measure. Statistical significance has been reported in Section 5.7 using McNemar's test.

5.6 Classification Experiments

Most signal and information processing studies, until recently, have focused on 'shallow' supervised machine learning algorithms such as Support Vector Machines (SVM), which use a shallow linear pattern separation model. Use of such architecture has been proved effective in solving many classification problems.

However, in a case of a natural speech, such a shallow representation can be problematic. The natural way of understanding human conversation suggest the need for a deep architecture. Due to the advancement of high-performance computing over the last years, such as modern graphics processing unit (GPU) [187], neural networks, containing several hierarchical layers, have been widely applied to all sorts of problems in Speech and Natural Language Processing (NLP) and Computer Vision with the huge success [188–190]. The approach is termed as "deep learning or deep neural networks (DNN)".

Therefore in this study, we first model, the overlap discourse classification system, using Support Vector Machines (SVM) to study the distinctive characteristics of acoustic features and role of speakers and context in classifying Cmp *vs* Ncm overlaps. Following the experimental study of the context, we linearly combined acoustic, AC, and lexical (bag-of-ngram representation, L-B) feature sets to model the overlaps. We then designed a Deep Neural Networks (DNN) to compare its performance with SVM. We further exploit the CNN architecture for designing overlap classification system using lexical word embedding (L-E) representation and a multimodal DNN architecture for combining the lexical (L-E) feature with AC in the hidden representation of the networks. A brief overview of the entire flow of designing the classification system is presented in Figure 5.3.

5.6.1 Support Vector Machines

We trained our classification systems using Sequential Minimal Optimization (SMO), a support vector machine implementation of weka [191]. Prior to classification, feature values are normalized within $[0, 1]$ intervals. Due to the high-dimentionality of the feature vector and a large number of instances, we used the linear kernel of SMO with its default value of the penalty parameter, $C = 1.0$, for training the model.

5.6.2 Deep Neural Network

5.6.2.1 Feed-Forward Neural Network

Figure 5.4 depicts the architecture of the fully-connected feed-forward neural network. This architecture is used to classify acoustic features (AC), lexical bag-of-ngram features (L-B) and their combination. In the architecture, the layers are densely connected, and each layer consists of a different number of units (u). The input to the DNN architecture is a vector x, which consists of individual feature sets or a linear combination of the acoustic and lexical features. The input is mapped to the output y, as shown in Equation 5.9,

$$y = f(x) = g(W.x) \tag{5.9}$$

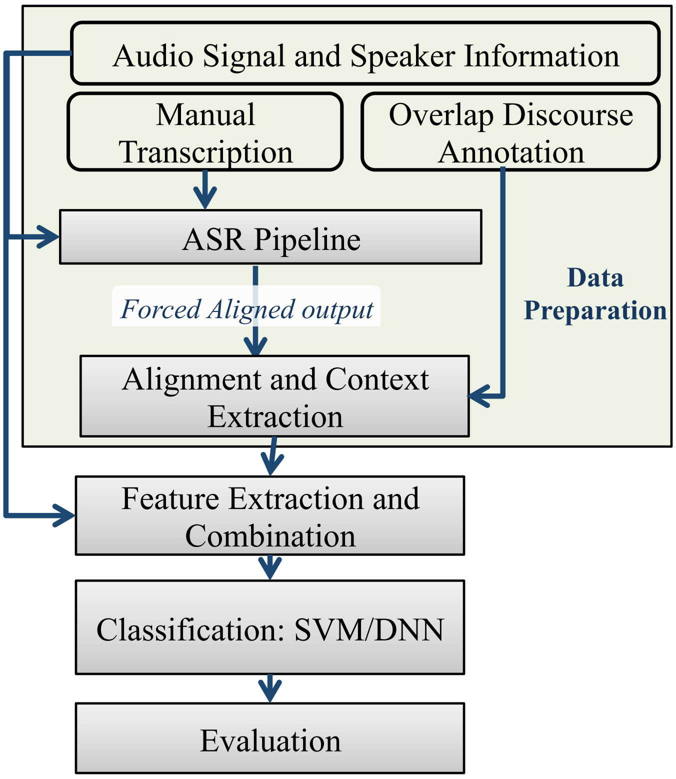

Audio Signal and Speaker Information
Manual Transcription
Overlap Discourse Annotation
ASR Pipeline
Forced Aligned output
Alignment and Context Extraction
Data Preparation
Feature Extraction and Combination
Classification: SVM/DNN
Evaluation

Figure 5.3: System architecture for modeling competitiveness in overlapping speech.

$$g(z) = \begin{cases} 0 & \text{for} \quad z < 0 \\ 1 & \text{for} \quad z \geq 0 \end{cases} \tag{5.10}$$

where the function $g(.)$ is some activation function, and $W \in R^2$ is a matrix of parameters. For the input, the feature values are scaled with zero mean and unit variance.

In the hidden layers of the DNN, we use rectified linear unit (ReLU) [192] as an activation function, in Equation 5.10. We experimented with ReLU function due to its linear, non-saturating form, which helps greatly to accelerates the convergence of stochastic gradient descent compared to the other functions, such as sigmoid or tanh.

For the output layer, we use the softmax function. The number of hidden units per layer is given in Figure 5.4. These optimal values are obtained empirically on the development set using Adagrad [193] optimization and a batch size of 100.

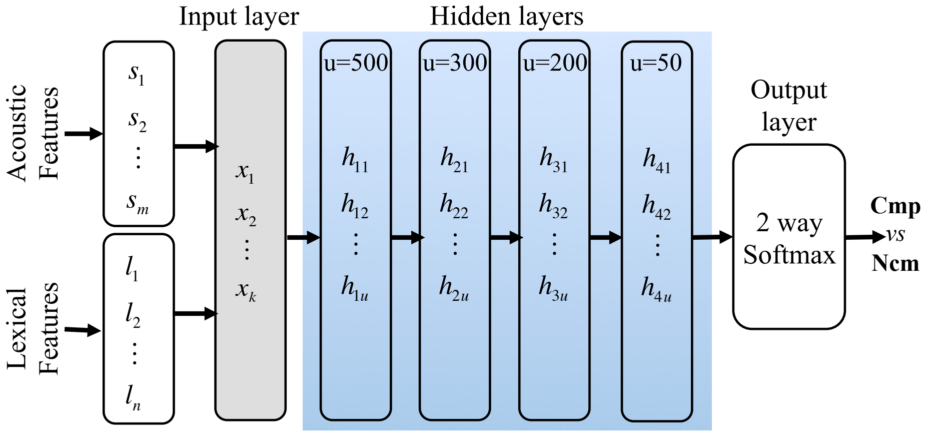


Figure 5.4: The DNN architecture for the classification of competitiveness in overlapping speech. u represents a number of units in each hidden layer. The input layer vector x can be acoustic feature vector S $(k = m)$ or lexical bag-of-ngram (L-B), feature vector representation $(k = n)$ or their linear combination $(k = m + n)$

.

5.6.2.2 Convolution Neural Network

In Figure 5.5, we present the architecture of our Convolution Neural Network (CNN). The input to the CNN is the D dimensional word vector for each word in vocabulary V in a shared look-up table $L \in \mathbb{R}^{|V| \times D}$, where L is the model parameter. We initialized L using the word embeddings discussed in Section 5.4.3. For an input transcription $s = \{w_1, w_2 ... w_n\}$, we design input vector $x_t \in \mathbb{R}^D$ for each word $w_t \in s$, which is an index in L. This input is then passed to the convolution layer. By applying max-pooling, we obtain a higher level feature representation, which is an equal sized feature vector for each instance. This representation is then passed to the one or more hidden layer(s), followed by an output layer. In each layer of this representation, we used different activation functions.

Since input transcriptions differ in length, i.e., number of words, therefore we padded them to make an equal length. It was required to perform convolution. The convolution operation involves applying a series of filter $u \in R^{L.D}$ to a window of L words to produce a new feature representation, shown in Equation 5.11.

$$h_t = f\left(u.x_{t:t+L-1} + b_t\right) \tag{5.11}$$

where $x_{t:t+L-1}$ is the concatenation of L input vectors, b_t is a bias term, and f is a nonlinear activation function. We used rectifier linear unit (ReLU).

We applied the filters with size 2, and 3 considering the fact that they can capture n-gram information. This filtering has been applied to generate a feature map $h_i = [h_1, h_2, \cdots, h_{T+L-1}]$.

This feature map has been designed for all filters. Then max-pooling operation (as shown is Equation 5.12) is applied to obtain an equal sized higher level feature representation.

$$m = [\mu_p(h_1), \mu_p(h_2), \cdots, \mu_p(h_N)] \tag{5.12}$$

where $\mu_p(h_i)$ is the max pooling operation. It is applied to each window of p features in the feature map h_i. We used the value of p as 2, and 3.

In the convolution and fully connected layers we used ReLU as an activation function, and in the output layer, we used softmax activation function.

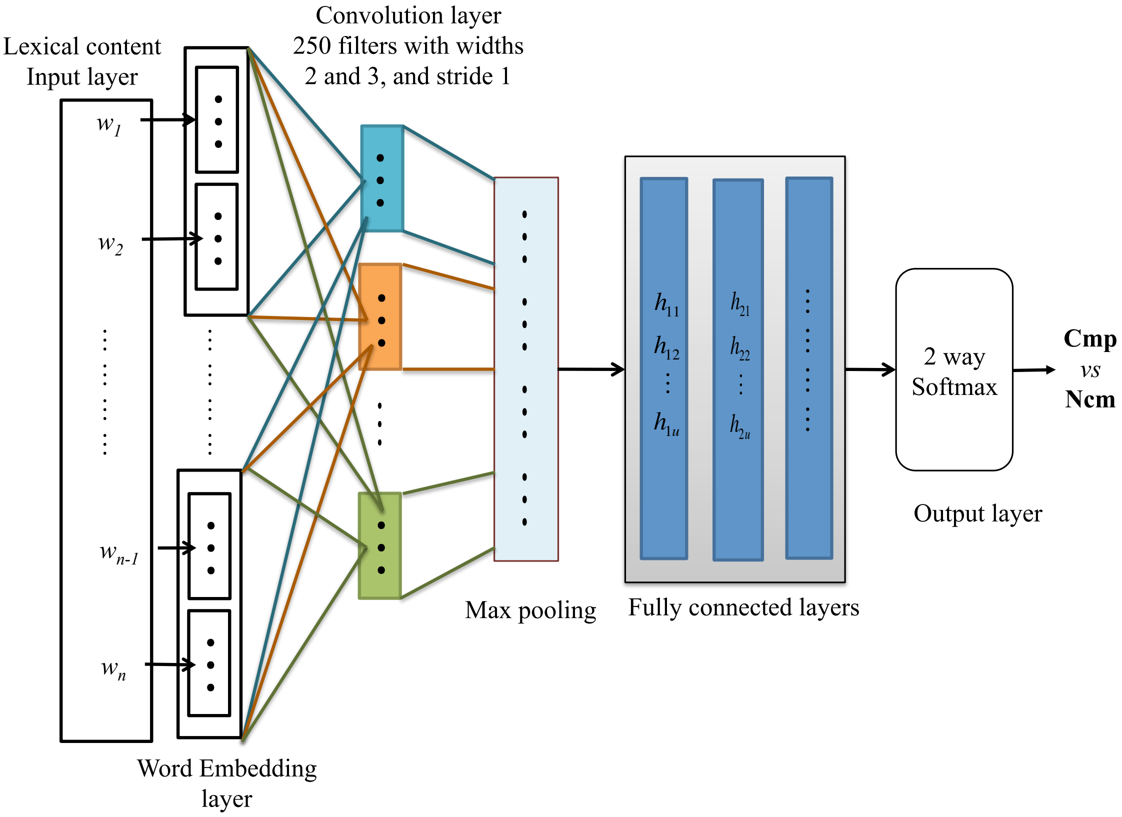


Figure 5.5: The CNN architecture for the classification of competitiveness in overlapping speech using lexical word embedding feature set (L-E).

5.6.2.3 Multimodal Deep Neural Network

In Figure 5.6, we present the architecture of the multimodal deep neural network to combine acoustic and lexical (L-E) information. The system takes audio signal and transcription as input, and for each input modality, we have different hidden representation followed by a layer, in which we combine the hidden representation. After the combined layer we can employ one or more hidden layer(s) before output layer. This architecture is heavily dependent on parameter tunning, which includes a number of layers, a number of hidden unit in each layer, choices of activation function such as ReLU, tanh, and optimization function such as SGD, Adadelta, Adagrad, and Adam. As for the activation function, we used ReLU in hidden layers and softmax in the output layer and used Adadelta as an optimization method.

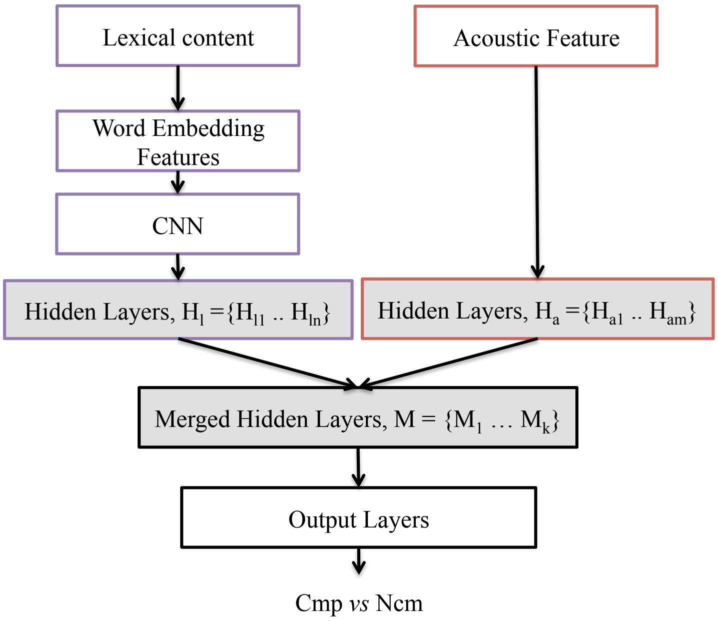

Lexical content
Acoustic Feature
Word Embedding Features
CNN
Hidden Layers, $H_l = \{H_{l1} .. H_{ln}\}$
Hidden Layers, $H_a = \{H_{a1} .. H_{am}\}$
Merged Hidden Layers, $M = \{M_1 \ldots M_k\}$
Output Layers
Cmp *vs* Ncm

Figure 5.6: The multimodal DNN architecture for the classification of competitiveness in overlapping speech using lexical word embedding feature set (L-E) and low-level acoustic features combined in deep space of the network.

5.7 Experiments, Results and Discussion

5.7.1 Experiments with Acoustic Features

The goal of this study is to understand the discriminative characteristics of each acoustic feature group in categorizing competitive *vs* non-competitive overlaps and to remove any unnecessary group. For the experiment, we selected a subset of data containing 253 conversations with approximately 27 hours, from which we obtained 9858 overlaps segments, for a total duration of 3 hours and 56 minutes. The low-level features are extracted as a group-wise and then projected into statistical functionals, presented in Table 5.3.

As mentioned in Equation 5.4-5.5, the acoustic features are extracted for the channels and then merged. This procedure is applied for each feature group such as Pitch, Voice quality, MFCC (Cepstral), Energy, Formants, Spectral among others. Hence, the representation of each group is same as S in Equation 5.5.

Table 5.4: Classification results on pilot-study test set to observe the contribution of each acoustic feature set. Precision, Recall and F1 are macro-averaged. Dim. : feature dimension.

Features	**Dim.**	**P(Avg)**	**R(Avg)**	**F1(Avg)**
Prosody (P)	576	67.7	68.1	67.8
VQ (V)	576	67.8	60.2	63.8
MFCC (M)	1872	66.5	68.4	67.4
Energy (E)	144	67.4	67.5	67.5
Spectral (Sp)	1728	68. 4	69.3	68.8

For classification, we used support vector machine with its linear kernel and its default parameters. To understand the relevance of each feature set for competitiveness and non-competitiveness binary classification task, we designed per-category classifier using SVM. The results of the pilot experiment are presented in Table 5.4.

From the experiment results, we observed that certain groups of acoustic features carry information regarding the discourse of overlaps. The results indicate that spectral and prosodic features are the key distinguishing feature groups. It is also worth noticing that some feature groups contribute more on a specific class decision rather than overall such as Voice quality does not at all provide any information.

Based on this study, we selected groups of acoustic features, presented in Table 5.5, to design the acoustic feature set (AC) for all classification tasks discussed in the following sections.

$$AC = P \cup V \cup M \cup E \cup Sp \tag{5.13}$$

5.7.2 Overlapper, Overlappee and Context

The performance of different speakers' segment and their associated feature set is reported in Table 5.6 for both dev and test set. For comparison, a SMO classifier has been designed using duration of overlapping segments as a feature for the baseline results. The baseline, **F1** for the dev and test set are 43.18 and 43.57, respectively. Results in Table 5.6 are significantly better compared to the baseline with $p < 0.001$.

Table 5.5: Selected Low-level acoustic features extracted using openSMILE for overlap classification, with the feature counts per channel.

Feature Group	**#**
Prosodic	288
pitch (fundamental frequency F0, F0-envelop) loudness, voice probability	
Voice Quality	288
jitter, shimmer logarithmic harmonics-to-noise ratio (logHNR)	
MFCC	936
Mel-Frequency Cepstral Coefficients (MFCC 0-12)	
Energy	72
Logarithmic signal energy from PCM frames	
Spectral	864
Energy in spectral bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz) roll-off points (25%, 50%, 70%, 90%) centroid, flux, max-position, min-position	
Total	2448

We obtained best results with decision combinations as shown in the Table 5.6, with F1 69.41 and 66.43 on the dev and test set respectively. The improved result on the test set is significantly better compared to all of individual systems with $p < 0.001$. To combine the decisions from the best classification models we used majority voting ensemble method as a combiner. We selected four best models based on the performance of the dev set and the models are: 1) overlapper with lexical features ($O : Lex$), 2) overlapper-overlappee and context with acoustic features ($OPC : AC$), 3) overlapper-overlappee and context with psycholinguistic features ($OPC : LIWC$), and 4) overlapper-overlappee with POS features ($OP : POS$).

The system designed with lexical features (in bag-of-ngram representation) from the overlapper channel performs better than any other individual system. The results on the dev set is 67.10 and on the test set is 64.99. The statistical significant test reveals that results using lexical features ($\boldsymbol{O : Lex}$) are highly significant with all other systems and their associated feature set ($p < 0.001$) except acoustic feature set in the context of $\boldsymbol{OPC}$. The $\boldsymbol{O : Lex}$ results are weakly significant compared to $\boldsymbol{OPC : AC}$ with a $p = 0.06$.

We are obtaining comparable results with acoustic features, and it is an ideal condition

Table 5.6: Classification Results for speakers' segments: overlapper O, overlappee P, left-context L, right-context R, along with the combination of overlapper-overlappe OP, left-right context LR and overlapper-overlappee with context OPC. Reported value is F1 measure of overall system. S.Seg: speaker's segment, Comb. Model: results for best model combination.

S.Seg	**Eval**	**Lex**	**AC**	**POS**	**LIWC**
$\boldsymbol{O}$	**Dev**	*67.10*	62.80	59.73	53.60
	Test	**64.99**	60.38	59.37	52.60
$\boldsymbol{P}$	**Dev**	59.50	58.43	55.73	50.10
	Test	58.87	57.01	55.35	50.05
$\boldsymbol{L}$	**Dev**	50.77	52.05	49.37	50.00
	Test	50.47	52.44	51.62	50.05
$\boldsymbol{R}$	**Dev**	52.04	62.34	51.24	50.05
	Test	51.27	62.33	49.35	49.85
$\boldsymbol{OP}$	**Dev**	64.09	62.47	*61.12*	57.75
	Test	63.00	60.14	58.52	56.35
$\boldsymbol{LR}$	**Dev**	52.07	59.82	48.26	50.00
	Test	51.50	59.40	48.31	49.75
$\boldsymbol{OPC}$	**Dev**	64.27	*65.31*	59.02	*62.25*
	Test	62.57	**64.36**	**58.94**	**59.55**
Comb. Model	**Dev**	69.41			
	Test	**66.43**			

when no transcriptions are available. The performance of the classifier designed with acoustic features extracted from overlapper-overlappee and context, $\boldsymbol{OPC}$, is F1 64.36 on the test set. In the case of acoustic features, we observed that performance improves when we include context along with $\boldsymbol{OP}$.

As for the importance of context alone, the authors in [56] claim that no cues can be found before the overlaps. Our results with acoustic features from left context, $\boldsymbol{L}$, shows a similar characteristics. We obtained lower classification results, 52.44 of F1 on the test set. The lack of the contextual evidence affects on recall in case of *Cmp*. Another reason is the size of the left context window, and in our case, it is 0.2s of speech.

For the right context, the authors in [56] and [80] agree that the effect of competitive overlap sometimes gets extended after the end of the overlap. A similar pattern is observed in our results using acoustic features of the right context, $\boldsymbol{R}$, where we obtained 9.89% improvement compared to the left context on the test set.

We observed that psycholinguistic features can distinguish competitive instances better when knowledge of the surrounding (***OPC***) overlap is provided. One of the possible reasons is the presence of the change of word usage before, inside and following an overlap. We computed correlation coefficients between LIWC features and class labels using Pearson's correlation. We found that the highly correlated features are a pronoun, cognitive processes, social processes among others.

The best performance with the POS features is observed in overlapper-overlappee and context (OPC) segment, giving a **F1** of 58.52 on the test set. However, with POS features extracted from overlapper-overlappee (OP) we obtained 61.12 on the dev set.

In summary, the competitiveness of the overlapping speech is best predicted using: 1) overlapper's lexical choice (***O***:Lex), 2) acoustic and psycholinguistic features while exploiting the complete knowledge, i.e., speaker's role and context (***OPC***:AC and ***OPC***:LIWC), 3) POS features when using overlapper information along with overlappee (***OP***:POS), and 4) decision combinations of the best classification models.

5.7.3 Comparative Study: SVM *vs* DNN

The results of the classification experiments are reported in Table 5.7. The performances of the SVM model for the acoustic and lexical (bag-of-ngram) features are considered as a baseline and are taken from Table 5.6. For lexical (L-B) feature we used OPC instead of O for taking into accounts of cases where it is not easy to find who initiated the overlaps.

Table 5.7: F1 measure for the individual classes and the macro-averaged F1 for the system as a whole on the development and test sets. AC – Acoustic, Lex (L-B) – Lexical features in bag-of-ngram representation, AC + Lex (L-B)– Feature combination of the acoustic and lexical feature sets.

F1		**Dev-set**			**Test-set**		
Classifier	**Feat.Set**	**Cmp**	**Ncm**	**Overall**	**Cmp**	**Ncm**	**Overall**
SVM	AC	0.46	0.85	0.65	0.44	0.85	0.64
	Lex (L-B)	0.46	0.83	0.64	0.43	0.82	0.63
	AC + Lex (L-B)	0.54	0.84	0.69	0.48	0.83	0.66
FeedForward	AC	0.54	0.84	0.69	0.50	0.83	0.67
	Lex (L-B)	0.37	0.84	0.61	0.32	0.84	0.58
	AC + Lex (L-B)	*0.57*	*0.87*	*0.72*	**0.51**	**0.86**	**0.68**

For the SVM, we observe a significant ($p < 0.05$) improvement in performance using linear feature combination, especially for competitive overlaps. A significant increase in F1 of 4.50% and 5.31% on the test set is observed compared to the individual SVM models using acoustic and lexical (L-B) features only. For the non-competitive overlap class, on the other hand, the feature combination outperforms the lexical model only. The model trained on acoustic feature outperforms both the lexical (L-B) and the linear combination models.

We used the same DNNs architecture for the feature combination and lexical feature, as shown in Figure 5.4. As for acoustic feature, we used four hidden layers with a different number of units, (500,400,200,50)[5]. DNN architecture for the acoustic feature set significantly outperforms both individual feature SVM models. An improvement of $\approx 6\%$ in F1 is observed for competitive overlap with respect to the acoustic feature set using SVM model. We do not observe a similar pattern for the non-competitive class where SVM with acoustic features yields F1 of 0.85 compared to F1 of 0.83 for the DNN with acoustic features only. The overall performance for the lexical features (L-B) is poor with respects to the rest of the experimental results. The weak performance of lexical features (L-B) has been observed especially for competitive overlaps. This can be due to the fact that lexical pattern describing non-competitive classes are closed set whereas for competitive they are very open, i.e., any words can be used to express the competitiveness intension. Moreover, from an experimental point of view, lexical feature design used here is very basic. So we have experimented with a more advanced feature extraction technique such as convolution based word embedding features (L-E), explained in Section 5.7.4.

For the combined feature set, DNN architecture not only improves the F-measure of the competitive overlap class, but also for the non-competitive class, and, consequently, the performance of the whole system. An improvement of 7.39% and 8.20% is observed for competitive overlaps when compared to individual feature SVM models. A similar pattern is observed for the non-competitive overlap class with DNNs using feature combination when compared to the individual feature SVM models.

Comparing SVM and DNN models using the feature combination, we observe an increase of 2.89% in F1 for competitive overlap class, 2.48% in non-competitive overlap class and 2.24% for the system overall.

5.7.4 Word Embedding with CNN

The result of CNN architecture with word embedding lexical feature is presented in Table 5.8. It is observed that word embedding features with CNN architecture using 3 hidden layers with units u (200,400,300) outperforms the bag-of-ngram feature with the DNN architecture by $\approx 4\%$. However, when we compared bag-of-ngram feature using SVM architecture, we

[5] Some of the results differs from the results published in [194], due to changes in some parameters.

observed that the results are not significantly different. The performance of CNN architecture is heavily dependent on the tuning of hidden layers with its units, and other parameters such as filter number, learning rate, etc. In this experiment, only the number of hidden layer, h is tuned for $h = \{2, 3\}$ and for a fixed list of neurons, leaving scope for further improvement using a different architecture and different set of parameters. Also as it is observed that bag-of-words are also providing important information regarding class discrimination, in future, we will also try to combine word embedding features with a bag-of-ngram feature to exploit their classification power together.

Table 5.8: F1 measure for the individual classes and the macro-averaged F1 for the system using Lexical bag-of-ngrams and word embedding feature representation. Lex (L-B) – Lexical features in bag-of-ngram representation, Lex (L-E)– Lexical features in word embedding representation.

F1		**Dev-set**			**Test-set**		
Architecture	**Feat.Set**	**Cmp**	**Ncm**	**Overall**	**Cmp**	**Ncm**	**Overall**
FeedForward	Lex (L-B)	0.37	0.84	0.61	0.32	0.84	0.58
CNN	Lex (L-E)	0.46	0.82	0.64	0.43	0.82	0.62

5.7.5 Multimodal DNN

For exploting the combination of word embedding features along with acoustic feature, we studied a multimodal DNN architecture described in Section 5.6.2.3. Similar to all DNN experiment, this system's performance is heavily dependent on the parameters used. Due to the complexity of the architecture tuning and resourses it needs, we fixed the hidden layer for individual feature set to three before merging the hidden layers with a fixed number of nuerons in each layer. We then tuned the hidden layers (M_h) containing the combined features to: $M_h = \{2, 3\}$. The result presented in the Table 5.9, used the following architecture: for lexical word embedding cnn architecture $H_l = 3$ where units in the layers are $u(H_{l1}) = 200$, $u(H_{l2}) = 400$, $u(H_{l3}) = 300$; for acoustic feature feed forward architecture $H_a = 3$ with units in layers are $u(H_{a1}) = 400$, $u(H_{a2}) = 500$, $u(H_{a3}) = 300$; for the combined (after merge) architecture we used $M_h = 3$ with hidden units of layers are $u(M_{h1}) = 500$, $u(M_{h2}) = 300$, $u(M_{h3}) = 50$. This experiment is done to see the capabilities of such architecture. It is observed that even though the dimension of the lexical features is different in both the experiments, the performance of the system is quite similar. This can be due to the presence of acoustic features showing how dominant it becomes when used with the representation power of the DNN architecture, as observed in Table 5.7, where a feed-forward network with acoustic performs similar to the acoustic + lexical (L-B) settings in SVM.

Even though due to time and resource limitations we did not tuned it properly, however, the result suggests one can utilize such an architecture to investigate further. Therefore, in future, we will investigate this architecture in order to understand the parameters and improve the performance of the current system.

Table 5.9: F1 measure for the individual classes and the macro-averaged F1 for the system using feature combination in two settings 1) Linear feature combination (acoustic + bag-of-ngram lexical features) followed by a Feed Forward DNN architecture and 2) A multimodal architecture with acoustic feature and word embedding feature merged in the hidden representation of the network.

F1		**Dev-set**			**Test-set**		
Architecture	**Feat.Set**	**Cmp**	**Ncm**	**Overall**	**Cmp**	**Ncm**	**Overall**
FeedForward	AC + Lex (L-B)	*0.57*	*0.87*	*0.72*	**0.51**	**0.86**	**0.68**
Multimodal	AC +Lex (L-E)	0.53	0.85	0.69	0.51	0.84	0.68

5.8 Overlap Detection and Classification in Mono Channel

The overlap classification model relies on the identification of the overlapping segments of speech. In case conversation speakers are recorded on separate channels, the detection of these segments is less complex. Unfortunately, sometimes conversational data are usually recorded on a single channel; thus, an overlap detection step from a single channel is also required. The task is known to be a hard one. For both tasks – overlap detection and classification – we train model by remixing channels of the annotated data. The data is described in Section 5.2 and the process is described in Section 5.8.1. Then, we describe overlap detection and classification experiments in Sections 5.8.2 and 5.8.3, respectively.

5.8.1 Training Data and Pre-Processing

The data used for training and testing the overlap detection and classification models is discussed in details in Section 5.2, and is the same set that we used to model our overlap discourse classifier for separate channels in this chapter.

Since some data can usually be recorded on a single channel; therefore to evaluate the performance of the overlap classification on such data, we apply channel remixing on both training and testing data using SoX (Sound eXchange[6]). The whole process is depicted in Figure 5.7 including training and testing stages, which are described next.

[6] http://sox.sourceforge.net/

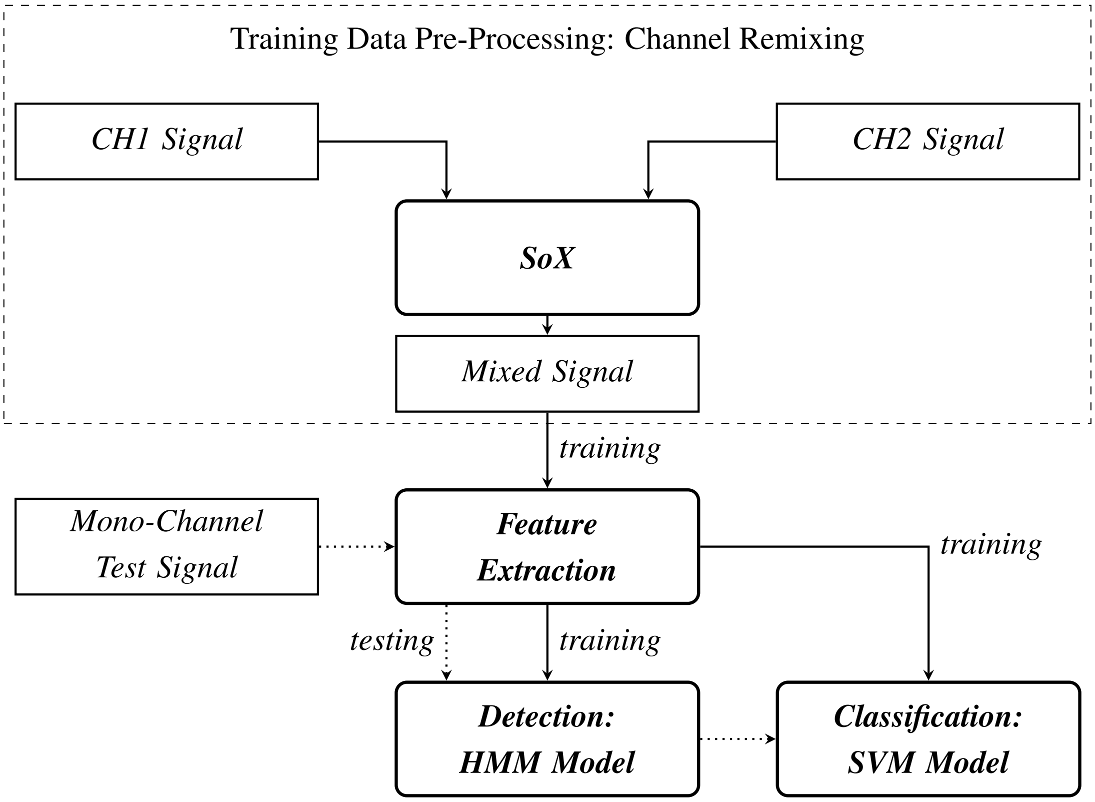


Figure 5.7: Overlap detection and classification system. Channel remixing (boxed), training (solid arrows) and testing (dotted arrows) pipelines.

5.8.2 Overlap Detection

Overlapping speech is detected using a Hidden Markov Model (HMM)-based overlap segmenter. In HMM, speech or overlap segment is represented with six-states and non-speech with a five-states model. The state emission probabilities are modeled with a multivariate Gaussian Mixture Model (GMM) with 32 components. The segmenter consists of three classes — non-speech, speech, and overlapped speech. Speech, non-speech, and overlap regions are identified in the training data using Automatic Speech Recognition (ASR) forced-alignment segment time, generated from manual transcriptions. The segmentation and labeling of the conversation are performed using a single Viterbi decoding pass on the full audio signal. The non-speech segments (mainly silence) are merged with their surrounding speech/overlap segment. The system is evaluated using NIST speaker diarization evaluation approach [195, 196].

The performance of the system on mono-channel signal is reported in Table 5.10 as recall and recall weighted by the duration of the overlap segment. The model can detect approximately 48% of overlaps. From the results, we can observe that it is harder to detect longer overlaps since duration weighted recall is lower (43.35). Overall, results are promising, and the task will

Table 5.10: Mono-channel overlap detection performance as duration weighted (WR) and unweighted recall (R).

Model	**R**	**WR**
HMM	48.05	43.35

Table 5.11: Macro- and micro- average F_1 for overlap classification using mono-channel model and a majority baseline.

Model	**Macro-F_1**	**Micro-F_1**
Baseline	43.6	77.2
Dual-Channel	64.4	76.0
Mono-Channel	61.8	76.0

be addressed in the future study.

5.8.3 Overlap Classification

The overlap classification model is trained using Sequential Minimal Optimisation (SMO), a support vector machine implementation of weka [191] using a linear kernel with default parameter settings. The models is an adaptation of [177] system to mono-channel data.

Models are trained using low-level acoustic features extracted using openSMILE [178] with the FrameSize = 25 ms and FrameStep = 10 ms, which yields approximately 100 frames per second. The groups of these low-level features such as prosodic, energy, etc. with counts are given in Table 5.5. The extracted low-level features and their derivatives are projected onto statistical functionals such as range, absolute position of max and min, linear and quadratic regression coefficients and their corresponding approximation errors, moments-centroid, variance, standard deviation, skewness, kurtosis, zero crossing rate, peaks, mean peak distance, mean peak, geometric mean of non-zero values and number of non-zeros.

The overlap classification results are given in Table 5.11. The reported numbers are without error propagation from the overlap detection step. Due to the high ratio of non-competitive overlaps in the test set (77.2%), the micro-averaged F_1 of the majority baseline is high. However, we are interested in both classes; thus, we also report macro-averaged F_1. The described overlap classification system significantly outperforms the baseline considering the macro-averages in both settings – dual channel and single channel.

5.9 Summary

This chapter illustrated automatic classification of competitiveness and non-competitiveness of overlap segments in real-world call center data. To classify competitiveness in the overlap, this chapter introduced and evaluated an annotation scheme for the overlap categories. The study also focused on studying different high-dimensional acoustic feature groups to create the acoustic feature set followed by an investigation of the role of speakers and context using different speakers' segments, such as overlapper, overlappee, left and right context and their combinations. To understand their role, the study employed high dimensional low-level acoustic, linguistic, and psycholinguistic features. Additionally, the study also implemented different combination, both decision, and feature-level, techniques. The purpose is to study is to develop an architecture to combine different information in classifying overlaps. It is observed, that the feature level combination of lexical information (Bag-of-ngrams or word embedding features) along with acoustic information outperforms any individual feature models and their decision level combination. In addition, to explore different features and their combination, the chapter also focused on exploiting the power of linear (SVM) and different architecture of non-linear (DNN) algorithms to classify the overlap discourse. Experimental results indicate that by exploiting many layers of a non-linear information processing for high-dimensional features yields significant improvement over all the individual feature sets using both SVM and DNN architecture along with feature combination with SVM.

Chapter 6

Functions of Long Silences in Dyadic Conversations

6.1 Introduction

Silence is a multifaceted natural phenomenon in human conversations that carries information rich in meaning and function. Even though “silence” is generally defined as the absence of speech [99] or a break in a conversation flow, its occurrence has the power to deliver a message, as well as trigger human response similar to any other conversational behavior. Silence in human conversations provides insights into the thought process, emotion, and attitude [197] among others. At the same time, silence is used to convey power (dominance) and respect, and manage conflicts (see Section 2.3).

Along with speech, silence is an integral part of human interaction, and the two complement and provide information about each other. In the words of Bruneau [10]:

> “*Silence is to speech as the white of this paper is to this print*” – Thomas J Bruneau.

Given that the reasons for silence are limitless, it also has many functions. One function is “eloquent silences” that includes the use of silence in the funeral, at religious ceremonies, as a legal privilege ,or in response to a rhetorical question [107]. Apart from this, silence can be used to indicate topic avoidance, lack of information to response, agreement, disagreement, anger, frustration, uncertainty, hesitancy and others.

The research on function of silence in human interaction is limited. Most of the studies have focused on the location of silence in a conversation [108, 197], or as a non-verbal communication [90] and its practices in different cultures [197] or in different contexts.

Unlike research on speech, the studies on silence are either definitional (theoretical) or descriptive. Even within speech research community, there are very few studies that analyzed function of silence in methodological manner. Generally silence is not acknowledged as a form of interaction, but rather its function in a conversation is viewed as a “pause” or a “gap”. Whereas speech is viewed as the primary carrier of information. Thus, a further study of silence and its functions is important as silence *often does serve as a message*, or at least as a *means that offers contextual cues to the surrounding speech*. Moreover, silence is one of the most common phenomena in human interaction.

The goal of this chapter is to analyze the function of long silences [1] occurring between- and within- speakers in dyadic spoken conversations. Our focus was to understand the perceived

[1] In our study, we defined long silences based on the duration, which is greater than 1 second.

reasons of such functions towards the information flow in conversations. In this study we utilize the sequences of dialog acts present in the turns surrounding the silence itself and design feature vectors for individual long silences. The designed feature vectors are used to cluster silences using a well-known hierarchical concept formation algorithm, which is designed to model different aspects of human concept learning. Followed by the clustering, the resulting clusters are manually grouped into functional categories. The significance of these functional categories of long silences is analyzed with respect to the duration distribution and their utility for automatic classification using decision trees.

The methodology followed for grouping functions of the long silences is shown in Figure 6.1. The steps followed in the methodology are:

1. data preparation – this step involves extraction and selecting the long silences from a set of conversation in Section 6.2;

2. modeling silence into feature space – that includes feature design for silence instances in Section 6.3;

3. unsupervised clustering of the selected silence instancesin Section 6.4;

4. conceptualize the silence cluster based on their functional similarity in the information flow of the conversation, as presented in Section 6.5.

6.2 Data Preparation

To study the role of silence in information flow of the conversation, we have selected 10 conversations from the data set described in Chapter 3. The conversations are manually transcribed and annotated for overlap discourse and dialog acts. The dialog act annotation follows Dialogue Act Markup Language (DiAML) [1, 157] annotation scheme. The details of the annotation – such as considered core dimensions and communicative functions – is described in Chapter 3, Section 3.2.1.3.

6.2.1 Extraction of Silence

Silence positions as well as turn types are extracted using the *turn segmentation and labeling system* depicted in Figure 6.2. The input to the system is the audio of the conversation, the forced-aligned transcription and speaker information.

The forced-aligned transcription is obtained using an in-domain Automatic Speech Recognition (ASR) [176]. Lexical information from these forced-aligned transcripts is used to extract turn-taking sequences. The pipeline uses the time aligned output as tokens to create Inter-Pausal

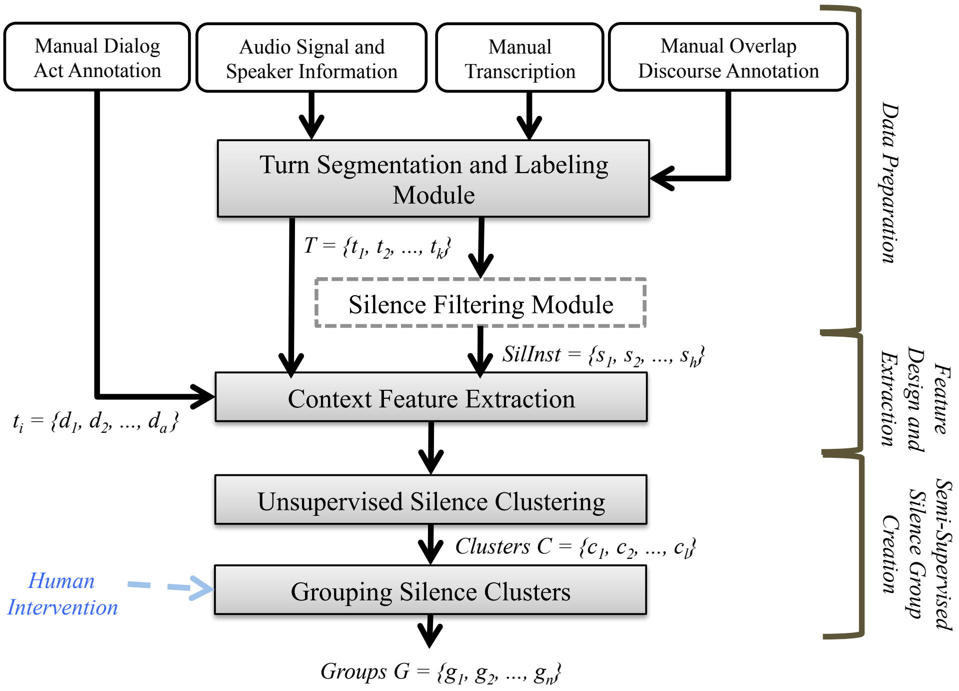

Manual Dialog Act Annotation
Audio Signal and Speaker Information
Manual Transcription
Manual Overlap Discourse Annotation
Turn Segmentation and Labeling Module
$T = \{t_1, t_2, ..., t_k\}$
Silence Filtering Module
$SilInst = \{s_1, s_2, ..., s_h\}$
Context Feature Extraction
$t_i = \{d_1, d_2, ..., d_a\}$
Unsupervised Silence Clustering
Clusters $C = \{c_1, c_2, ..., c_l\}$
Human Intervention
Grouping Silence Clusters
Groups $G = \{g_1, g_2, ..., g_n\}$
Data Preparation
Feature Design and Extraction
Semi-Supervised Silence Group Creation

Figure 6.1: System framework for categorizing functions of long silences.

Units (IPUs) for each input channel. IPUs are defined as segments of consecutive tokens with no less that 50 ms gaps in between. Using the time information of inter-IPUs and intra-IPUs, we then define **steady conversation segments** where each segment maintains a steady time-line for both interlocutors channel. The labels of each silence segment are then defined by a set of rules as follows:

- Pause (P): Gaps between the turns of the same speaker with no less than 0.5 second. P_A and P_C represent agent and customer's pauses respectively.

- Lapse between speakers (L_B): Floor switches between the speakers with a silence duration of 2 seconds or more.

- Lapse within speaker (L_W): Gaps between a speakers' turns with a silence duration of 2 seconds or more.

- Switch (SS): Floor switches between the speakers with silence less than 2 seconds or with overlapping frames not more than 20 milliseconds. This category is also know as *gaps*.

The labeled turn sequences are then used to select silence instances for the analysis.

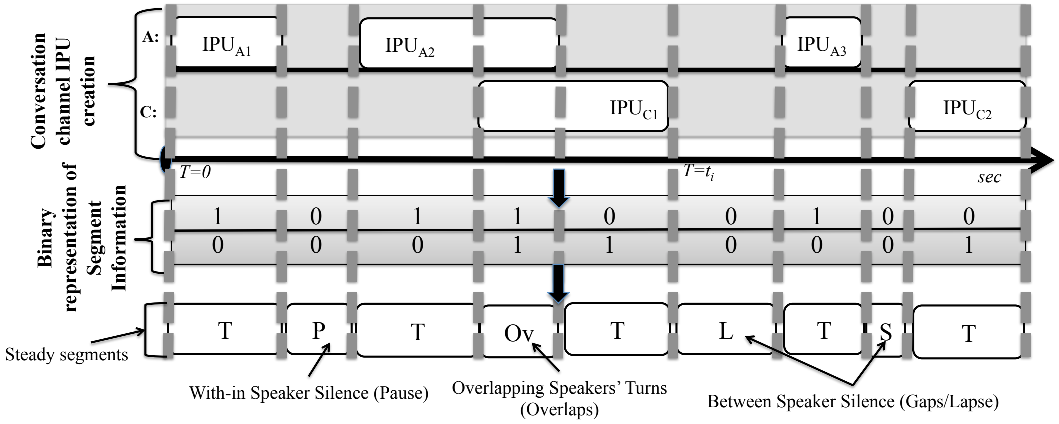

Conversation channel IPU creation
A:
C:
IPU$_{A1}$
IPU$_{A2}$
IPU$_{A3}$
IPU$_{C1}$
IPU$_{C2}$
T=0
T=t$_i$
sec
Binary representation of Segment Information
1 0 1 1 0 0 1 0 0
0 0 0 1 1 0 0 0 1
Steady segments
T P T Ov T L T S T
With-in Speaker Silence (Pause)
Overlapping Speakers' Turns (Overlaps)
Between Speaker Silence (Gaps/Lapse)

Figure 6.2: System architecture for extracting turn types and silences.

6.2.2 Silence Filtering

From the 10 conversations, we extracted 433 instances of silence with duration greater or equal to 1 second. The instances are categorized into two groups:

- Between-Speaker Silences (B): These instances of silence include gaps between different speaker turns that are greater or equals to 1 second. $B = \{S_l,\ L_B\}$, where S_l stands for gaps longer than 1 second and shorter than 2 seconds and L_B are gaps longer than or equal to 2 seconds.

- Within-Speaker Silences (W): These instances of silence include pauses between the same speaker's turns that are greater or equals to 1 second. $W = \{P_l,\ L_W\}$, where P_l stands for pauses longer that 1 second and L_W are pauses longer than or equal to 2 seconds.

For the initial analysis, the instances of long silence that occur after or before overlapping speech are ignored. As a result, the analysis is performed on 372 instances.

6.3 Modeling Silence

Even though silence is inherently valueless phenomena that possesses no function on its own, individual instances of silence gain its meaning and function from the surrounding context. Consequently, modeling function of silence requires conceptualization of the context and features that capture it.

As it is mentioned earlier in the dissertation, dialog acts carry specific communicative functions such as question, answer, expression of agreement, disagreement, etc. Since dialog acts are assigned to the speech segments (turns) that surround the long silences, they provide the information that could be used to model the context of silence instances.

Table 6.1: Core dimensions and communicative functions from ISO 24617-2 standard considered for dialog act annotation.

Dimension	**Comm.Function**	**Group**
General (Task)	*Information Transfer Functions*	
	Question Set Question Choice Question Propositional Question Check Question	Information Seeking
	Inform Answer Confirm Disconfirm Agreement Disagreement Correction	Information Providing
	Action Discussion Functions	
	Offer Promise Address Request Accept Request Decline Request Address Suggest Accept Suggest Decline Suggest	Commissives
	Suggest Request Instruct Address Offer Accept Offer Decline Offer	Directives
Time Management	Stalling, Pausing	
Auto-Feedback	Positive, Negative	
Allo-Feedback	Positive, Negative, Feedback Elicitation	
Social Obligations Management	Initial-Greeting, Return-Greeting Initial-Self-Intro, Return-Self-Intro Apology, Accept-Apology Thanking, Accept-Thanking Initial-Goodbye, Return-Goodbye	

Feature Design

The dialog act dimensions and communicative functions listed in Table 6.1 are used as features for the analysis of between and within speaker silence instances. Each turn preceding or following a silence is transformed into a feature vector using one-hot representation for dialog acts.

The vectors encode information as follows. Feedback, a joined dimension of auto-feedback and allo-feedback, (fb) = $\{0, 1\}$, where fb=0 represent the absence of feedback dialog acts in the turn and vice-versa. Similarly, the vector also includes other dialog act dimensions like Time Management (tm), and Social Obligations Management (s). The General dimension is split into two: (a) information seeking (q) and (b) information providing and action discussion functions (ac).

Since according to the DiaML annotation standard a turn can contain several dialog acts, the vector representation specifically encodes the last dialog act of the preceding turn ($lact$) and the first dialog act of the turn following the long silence ($fact$), according to the Equations 6.1 and 6.2.

$$lact = \{Ac, Q, F, TimeM, Ap, Thank, Int, Other, None\} \tag{6.1}$$

$$fact = \{Ac, Q, F, TimeM, Ap, Thank, Int, Other, None\} \tag{6.2}$$

In the equations, *Ac* represents communicative functions from *information providing and action discussion* functions; *Q* represents *Information Seeking* functions; *F* represents *Feedback (auto-feedback and allo-feedback)* functions; *Apo* represents *apology and accept-apology* functions; *Thank* represents *thanking and accept-thanking*; *Int* represents *initial and return greetings, self-introductions, and goodbyes*; *Other* represents all the dialog acts not used for the analysis. *None*, on the other hand, indicates absence of dialog acts.

The feature vectors of preceding, pr ($|pr|$=6) turn, and succeeding (following), su ($|su|$=6) turn, are merged to represent a silence instance for categorization ($|sil| = 6 * 2 = 12$), as shown in Figure 6.3.

6.4 Unsupervised Annotation of Silence Function

The designed representation of silence instances are used for clustering using Cobweb clustering algorithm [198] – a well-known concept formation system designed to model human concept learning.

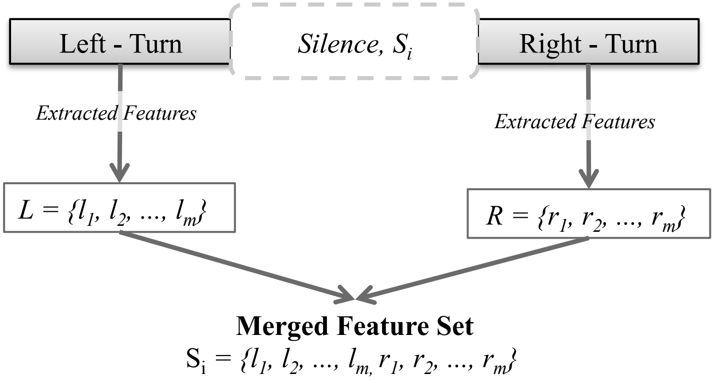

Left - Turn
Silence, S_i
Right - Turn
Extracted Features
Extracted Features
$L = \{l_1, l_2, ..., l_m\}$
$R = \{r_1, r_2, ..., r_m\}$
Merged Feature Set
$S_i = \{l_1, l_2, ..., l_m, r_1, r_2, ..., r_m\}$

Figure 6.3: Feature extraction and merging from preceding and succeeding turns.

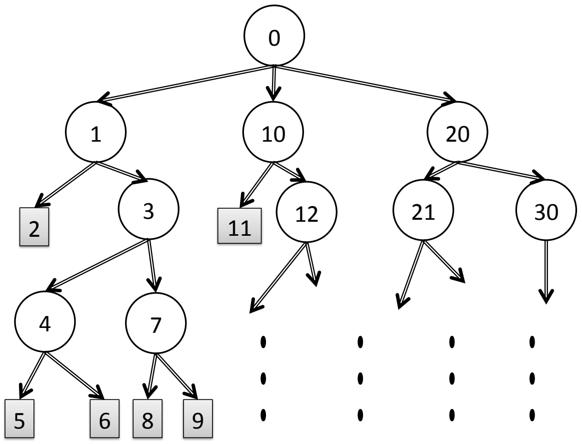


Figure 6.4: Cobweb classification tree example.

6.4.1 Clustering Algorithm

Cobweb constructs clusters using “concept hierarchy” that optimally and incrementally accounts for the observed regularities on a set of instances. In other words, given a set of silence instances, Cobweb discovers a classification scheme that covers the patterns with respect to provided feature vectors.

Instead of forming concepts at a single level of abstraction, Cobweb groups instances into a classification tree where leaves represent similar instances, and internal nodes represent broad concepts. The generality of a broader concept increases as the nodes approach the root of the tree (see Figure 6.4). Each cluster is characterized with a probabilistic description.

The classification tree is constructed incrementally by inserting the instances into the tree one by one. When adding an instance, the algorithm traverses the tree top-down starting from the root of the tree. At each node, there are four possible operations: (a) insert (b) create (c) merge and (d) split. These operations are selected with respect to the highest category utility (CU) metric [199]. The metric is derived from the categorization studies in cognitive psychology and is shown in Equation 6.3.

Category utility, CU, attempts to maximize both (a) the probability of the instances in the same category to have feature values in common; and (b) the probability of the instances in different categories to have different feature values.

$$CU(C_l) = \sum_i \sum_j (Pr[f_i = v_{ij}|C_l]^2 - Pr[f_i = v_{ij}]^2) \tag{6.3}$$

In the equation, $Pr[f_i = v_{ij}]$ represents the marginal probability that feature f_i has value v_{ij}, whereas $Pr[f_i = v_{ij}|C_l]$ represents the conditional probability that feature f_i has value v_{ij}, given the instance belongs in cluster C_l. $CU(C_l)$ estimates the quality of individual cluster.

To measure the quality of overall clustering of the silences, we calculate the average category utility function $CU(C_1, C_2, .., C_k)$, as shown in Equation 6.4.

$$CU(C_1, C_2, .., C_k) = \frac{1}{k}(\sum_l Pr[C_l]) \tag{6.4}$$

In the equation, k is the total number of categories. The overfitting is controlled by $\frac{1}{k}$.

6.4.2 Clustering Parameters

As it is mentioned in Section 6.2, the long silences are divided into two groups – B and W – with respect to the speakers of the turns preceding and succeeding silence. Therefore, for each set (B,W), we applied Cobweb clustering algorithm implemented in [191] with acuity $A = 1.0$ and cutoff threshold of $C = 0.0028$.

6.4.3 Resulting Clusters

For between-speakers silence (B), we obtained 24 leave clusters, whereas for within-speaker silence (W), we obtained 26 cluster leaves. Details of the clusters with the number of instances in each cluster is shown in Figures 6.5 and 6.6.

To understand contributions of each feature (f_i) and its values (v_{ij}) to cluster formation, we designed decision trees for between speaker and within speaker silence clusters as shown in Figures 6.7 and 6.8, respectively.

The distribution of dialog act sequences in each cluster is given in Tables 6.2 and 6.3.

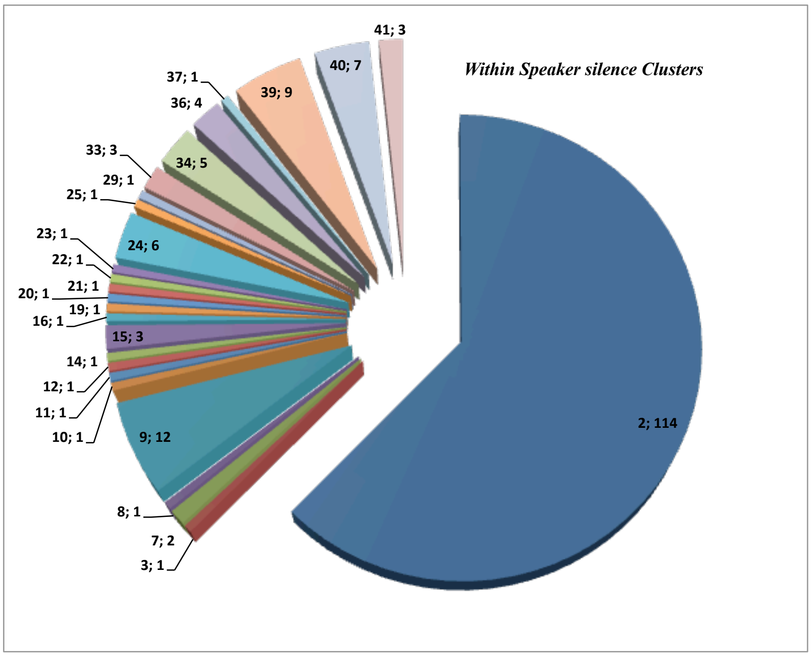


Figure 6.5: Pie chart presenting *cluster id* and *number of instances* in each cluster (c;n) of within-speaker instances where c is cluster id and n is number of instances.

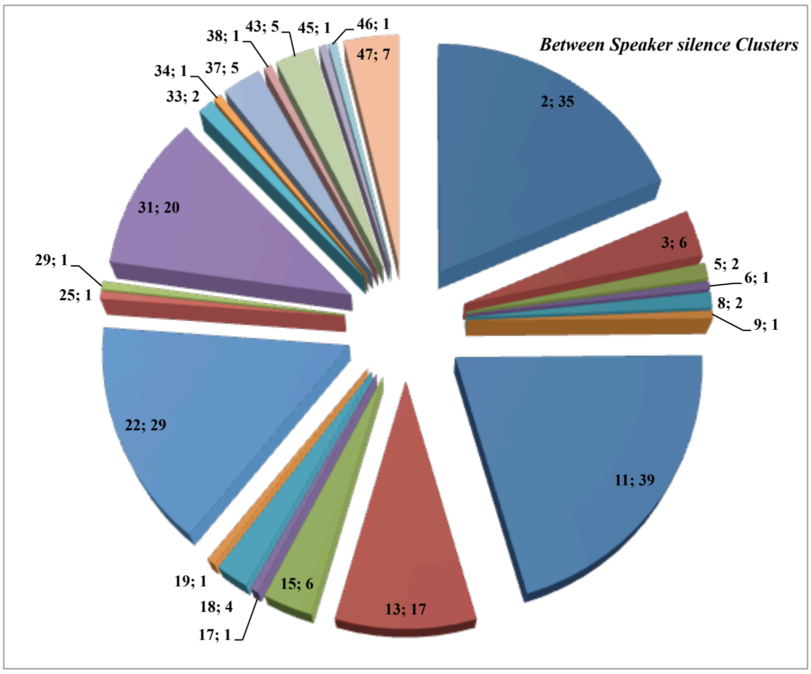


Figure 6.6: Pie chart presenting *cluster id* and *number of instances* in each cluster (c;n) of between-speaker instances, where c is cluster id and n is number of instances.

Table 6.2: Preceding and succeeding turn communicative function sequences for each clusters for between speaker silences.

Id	Preceding turn dialog acts	Succeeding turn dialog acts
2	question(19); checkquestion(9); inform question(2); inform checkquestion(2); inform autopositive question(1); choicequestion(1); autopositive checkquestion(1)	answer(12); confirm(11); inform(3); answer inform(3); disconfirm(2); confirm inform(2); disconfirm answer(1); answer request(1)
3	question(2); initialselfintroduction initialgreeting returnselfintroduction question(1); initialselfintroduction initialgreeting initialselfintroduction question(1); inform checkquestion(1); choicequestion(1)	other(2); autopositive(2); autopositive returngreeting stalling inform(1); allopositive(1)
5	question(2)	stalling answer(2)
6	question(1)	stalling checkquestion(1)
8	initialgreeting initialselfintroduction question(2)	returngreeting returnselfintroduction answer inform(1); returngreeting inform(1)
9	initialselfintroduction question(1)	returngreeting returnselfintroduction(1)
11	inform(20); request(6); confirm(2); answer(2); suggest(1); stalling request(1); offer(1); initialgreeting initialselfintroduction request(1); inform none inform(1); answer request(1); answer autopositive inform(1); agreement(1); addressrequest(1)	inform(22); acceptrequest inform(4); inform inform(3); confirm(3); acceptrequest(2); inform question(1); answer request(1); agreement(1); addressrequest(1); acceptoffer inform stalling(1)
13	autopositive(16); allopositive(1)	inform(15); inform request(1); correction(1)
15	other(6)	inform(5); suggest(1)
17	answer thanking(1)	inform(1)
18	pausing(2); stalling(1); inform stalling(1)	inform(2); confirm(1); answer(1)
19	allopositive none(1)	inform inform(1)
22	inform(23); answer(2); request(1); correction(1); confirm(1); acceptrequest inform(1)	autopositive(19); autopositive inform(4); autopositive question(3); autopositive checkquestion(2); allopositive(1)
25	allopositive(1)	autopositive(1)
29	pausing(1)	autopositive(1)
31	inform(10); answer(4); confirm(3); request(1); disconfirm(1); correction(1)	question(12); checkquestion(6); question inform(1); question checkquestion(1)
33	autopositive(2)	question(2)
34	autopositive(1)	question acceptthanking(1)
37	inform(2); confirm(2); offer(1)	pausing(4); stalling(1)
38	inform(1)	none(1)
43	other(5)	other(5)
45	other(1)	returnselfintroduction(1)
46	initialgreeting initialselfintroduction question other(1)	returngreeting(1)
47	inform(2); request(1); other inform(1); declinerequest(1); answer(1); acceptrequest(1)	other(5); other stalling(1); other other question(1)

Table 6.3: Preceding and succeeding turn communicative function sequences for each clusters for within speakers silences

Id	Preceding turn dialog acts	Succeeding turn dialog acts
2	inform(95); answer(6); request(3); stalling inform(2); inform inform(2); correction(2); question request(1); offer(1); inform request(1); confirm(1)	inform(90); request(5); answer(5); inform inform(4); offer(2); inform stalling(2); inform question(2); suggest(1); inform stalling inform stalling(1); correction(1); addressrequest(1)
3	none(1)	inform(1)
7	pausing(2)	question(1); checkquestion(1)
8	autopositive(1)	question(1)
9	question(8); checkquestion(3); inform question(1)	question(8); checkquestion(3); question inform(1)
10	question(1)	other(1)
11	question(1)	pausing(1)
12	question(1)	autopositive autopositive(1)
14	other(1)	apology inform(1)
15	other(3)	other(3)
16	other(1)	autopositive inform(1)
19	pausing(1)	pausing(1)
20	inform stalling(1)	stalling(1)
21	autopositive pausing(1)	pausing autopositive inform(1)
22	stalling(1)	other inform(1)
23	autopositive(1)	other(1)
24	autopositive(5); autopositive autopositive(1)	autopositive(4); autopositive thanking(1); autopositive question(1)
25	autopositive(1)	stalling inform(1)
29	inform none(1)	none inform(1)
33	stalling(1); pausing(1); other stalling(1)	inform(3)
34	autopositive(5)	inform(4); inform autopositive question(1)
36	question(4)	inform(3); inform inform(1)
37	other(1)	inform(1)
39	inform(9)	stalling inform(7); stalling(2)
40	inform(7)	question(4); question inform(2); choicequestion(1)
41	inform(2); agreement Null inform(1)	autopositive(3)

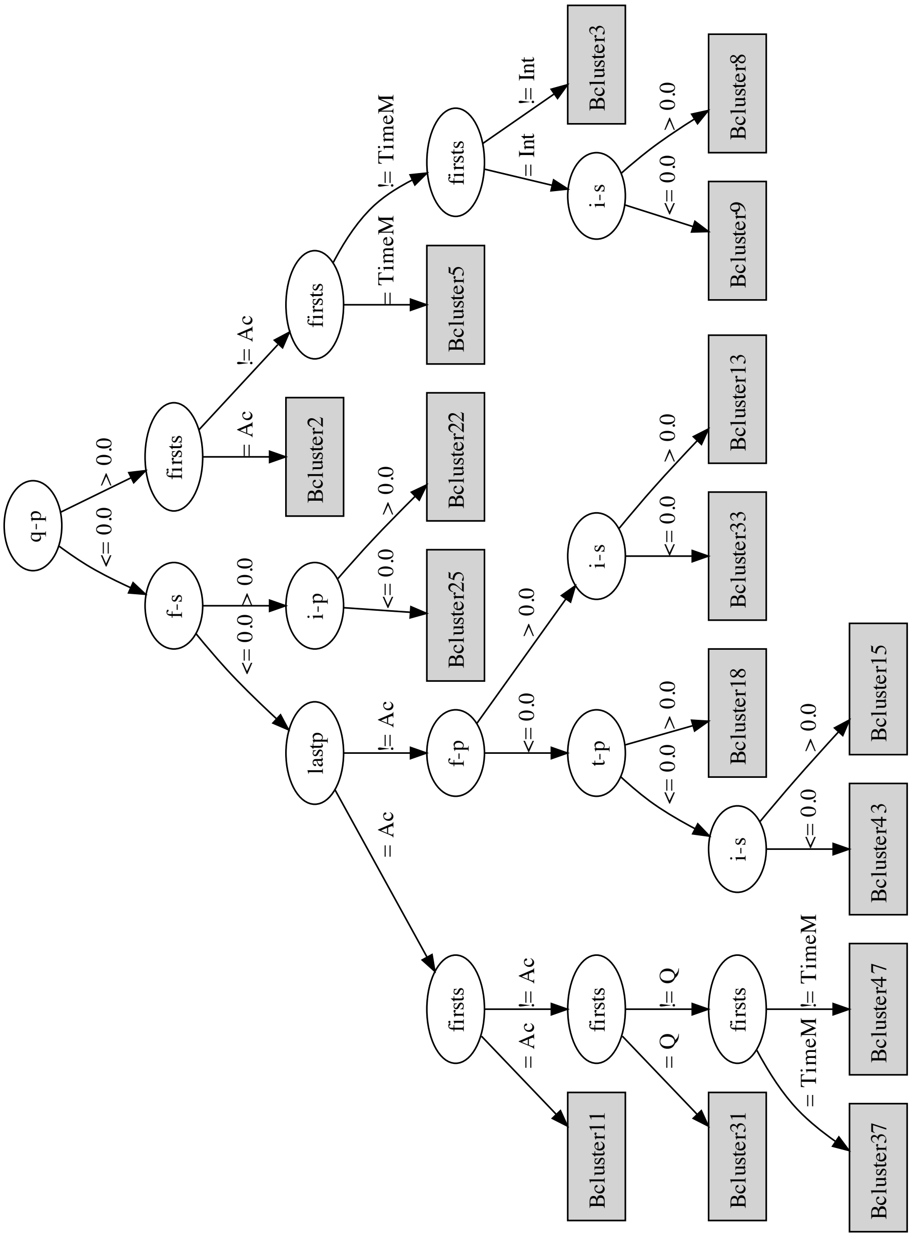


Figure 6.7: Decision Tree presenting the features and their values for between speaker silence. $-p$ represents preceding segment; $-s$ represents succeeding segment

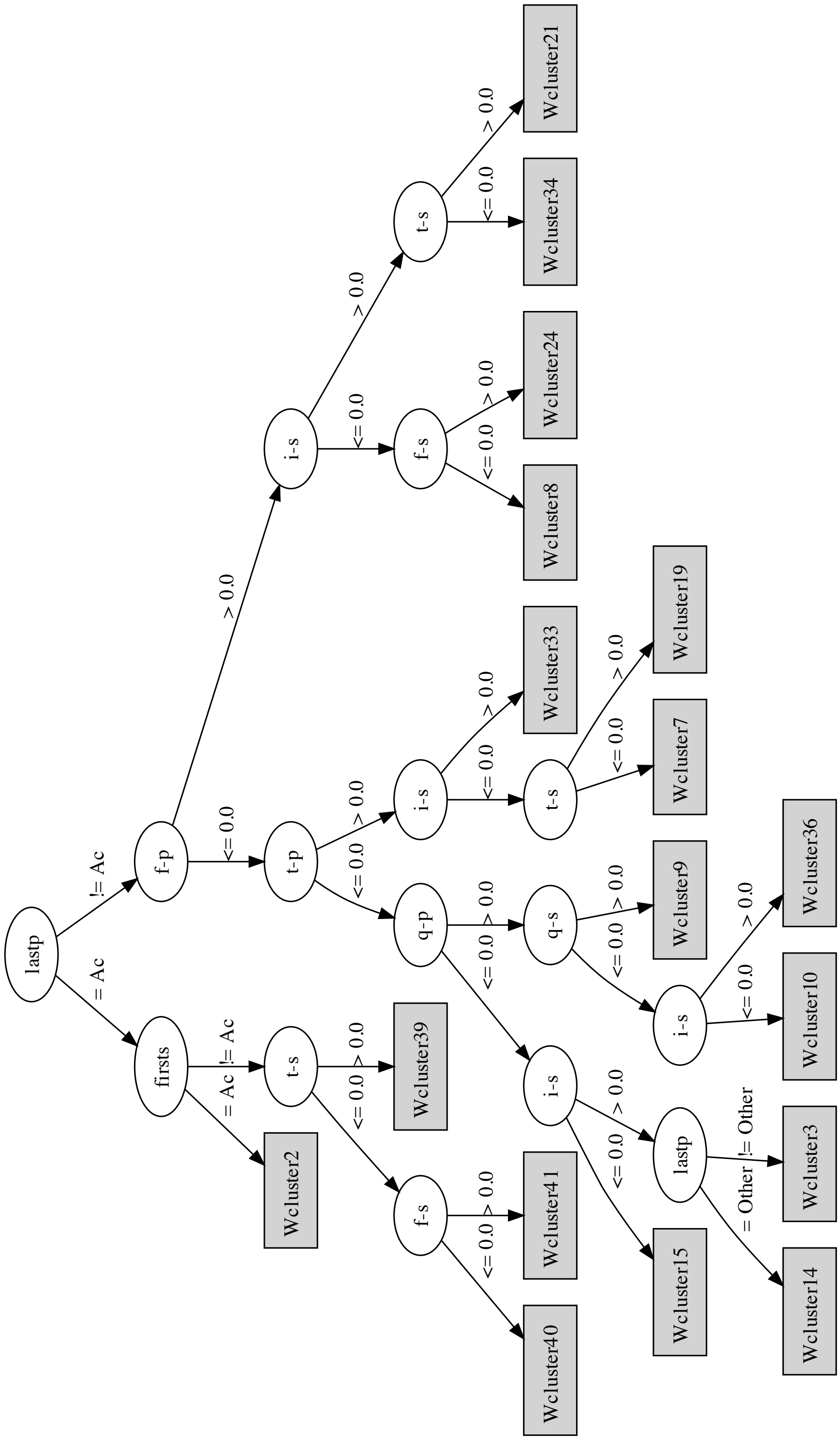


Figure 6.8: Decision Tree presenting the features and their values for within speaker silence. $-p$ represents preceding segment; $-s$ represents succeeding segment

6.5 Categorization of Silence Functions

Assuming that each cluster represents a function of a silence, the clusters are manually grouped with respect to their parents in the classification tree. The manual grouping of silence clusters is performed considering conversation scenarios. For instance, in a conversation a participant may expect an answer to a question or a contribution from another speaker that might yield a long silence due to the time required to prepare an answer. It might take long to get the information to the query or simply be an act of noncompliance. Strategies that follow a long silence are often target to repair the failure to contribute and are either repeating the query, changing the topic, or asking for more time to respond. Below we give example scenarios observed in the silence cluster groups:

The Between-Speaker Silence cluster groups are:

- A mode of response preparation (RP): In this group, there can be two subcategories based on the type of response given by the speaker before the silence. The subcategories are:
  - Response to the previous turn's question in the form of information that includes an answer to the question, a feedback, or asking for more time to answer. This category includes clusters **RP1**={2, 3, 5, 6, 8, 9}.
  - A response can also be a question to the information/feedback provided in the previous turn. This category includes clusters **RP2**={31, 33, 34}.
- A mode of information flow (IF): These silences can either be a: 1) conversational silences, where both speakers are exchanging information or feedback 2) forced silences (deliberate[2]), where the current speaker is using the silence as a tool to force the interlocutor to respond. The member clusters of this group are **IF**={11, 13, 15, 17, 18, 19, 22, 25, 29, 37, 38}.
- Silences in Other categories ($B - Oth$): These are the silences which are motivated by factors not considered in the dissertation. This group includes clusters **B-Oth**={43, 45, 46, 47}.

The above-mentioned categories are presented in Examples 4 and 5. In Example 4, we observe that the caller is asking the call center operator a reason behind an action, and the act is followed by a long silence of 1.41 seconds. After the interval, the operator is passing some information regarding the earlier query by the caller. From the operational point of view, the

[2] These silence instances are usually longer. For this study the threshold of this type of silences is $>= 2$ seconds.

interval might have been used to either acquire information or to structure it. Similarly, in $RP2$ scenario in Example 4, after the operator informs that the 'electric power' will not be activated, the caller is taking a long silence of 1.38 seconds to respond to the given information, asking another question. This silence could have been again used for preparing the answer, or it might be the time taken by the responding speaker to compose the next action. In Example 5, we present a scenario where the silence category IF is used deliberately to force another speaker to reply.

The silence in both examples may have other cognitive functions such as controlling emotional attitudes. However, as the focus of this study is to understand the function of long silences in information flow, these cognitive functions are not considered.

Example 4. Example of silence category RP : $RP1$

```
caller: al distacco perfetto ora eh eh su che base mi perdoni
caller: the complete interruption ... perfect! now ehm ehm due to what reason, excuse me?
     (1.14) Category - RP1
operator: ah ascolti qui ci sono una serie di fatture malgrado
operator: Listen (please) we have here a number of unpaied bill in spite of
```

Example of silence category RP : $RP2$

```
operator: la luce non gliela riprist non viene ripristinata
operator: the electric power will not be reactiv will not be reactivated
     (1.38) Category - RP2
caller: ma cosa devo pagare se io ho già conguagliato tutto con
     trecentoquarantacinque euro mi perdoni cosa devo pagare la
caller: but what do I have to pay if I have already paid 345 euros I beg you pardon but what do I
     need to pay the
```

Example 5. Example of silence category IF

```
caller:   [lei deve fare una cosina lei ha un delle]
caller:   [You have to do a small thing you have some]
operator:   [però e se]
operator:   [but and if]
caller: belle schermate a disposizione mi deve aprire la mia ehe il
     mio fax inviato il ventitrè zero otto duemiladodici
     cortesemente
caller: beautiful screens available you have to open my own and you will find my fax sent on 23rd
     of August 2016
     (2.12) Category - IF (deliberate silence)
operator: vediamo subito
operator: let us see immediately
```

The Within-Speaker Silence cluster groups are:

- Organizational silence (CS): The long pause used for the purpose of organizing the information flow in the conversation This group contains clusters of silences where a speaker is providing information and the silence between turns can either be a time taken to think, find information, or to compose and plan the next turn. **CS**={2, 3, 19, 20, 21, 22, 23, 24, 25, 29, 33, 34, 39}.

- Indecision or Hesitation silence (H): In this groups of silences, speaker is either confused about some information, needs clarification, or have some queries. The member clusters of this groups are **H**={7, 8, 9, 10, 11, 12, 36, 40}.

- Silences in Other categories ($W - Oth$): These are the silences which are motivated by other factors, not considered for the present study. This group includes clusters **W-Oth**={14, 15, 16, 37, 41}.

Example 6. Example of silence category CS and H

caller: non riesco a parl devo parlarle ho parlato con cinque suoi colleghi e mi hanno chiamato due consulenti
caller: *I cannot tal ... I need to talk ... I talked with five colleagues of you and two consultants called me*
(1.16) Category - CS
caller: io oggi pomeriggio devo andare dall avvocato per denunziarvi per diecimila euro al giorno di danni che mi avete arrecato da stamattina
caller: *this afternoon I will go my lawayer for sueing you due to ten thousand euros in damage per day due to this morning (power)interruption*
(1.65) Category - CS
caller: ehe perché io ho già pagato tutto nel
caller: *ehm because I already paid all what I due*
caller: senso che tutte queste bollette sono state conguagliate con una di trecentoquarantacinque euro incluso
caller: *because all these bills were paied with another one of 345 euros including*
caller: il mese di luglio e agosto
caller: *the months of July and August (as well)*
(1.57) Category - CS
caller: ehe avevo già chiarito il (.) primo distacco l abbiamo sospeso mi hanno richiamato perché non trovate una vostra lettera di risposta
caller: *and I already told this when (.) there was the first interruption (that) was suspended they called me because you are not able to find a reply letter from you*
(1.01) Category - H
caller: ora devo (.) parlare con lei o devo parlare con qualcuno sopra di lei mi perdoni se sono abbastanza
caller: *now (I) have (.) to call with you or have (I) to call with you boss? sorry but (I) am enough*

In Example 6, we present a dialog scenarios with assigned categories. It is observed that the first three long silence intervals are used either to plan the next turn or to take the time to think. Whereas in the last silence of 1.01 seconds, before threatening the operator, the caller either hesitates, feels bad, or is not sure whether a threat is going to work.

The duration distribution statistics for each category of silence functions are presented in Tables 6.4 and 6.5. For between-speaker silence categories, in Table 6.4, it is observed that median duration of silence category $RP2$ along with $B - Oth$ are longer compared to $RP1$ and IF. As for within-speaker silence categories, it is observed that median duration of H categories is longer than CS. The observation is explained as the speakers might need more time to take the next turn when s/he is facing indecision, hesitation, or need clarification about something.

To understand the utility of designed categories for prediction using the feature vector designed in Section 6.3, we train a decision tree classifier (J48 implementation in [191]) for both between and within silences using 10-fold cross validation. The resulting decision trees are presented in Figures 6.9 and 6.10.

Table 6.4: Statistics of between-speaker long silences categories.

Between-Speaker Silence	**RP1**	**RP2**	**IF**	**B-Oth**
Min.	1.01	1.02	1	1.1
1st Qu.	1.205	1.33	1.27	1.357
Median	1.37	1.76	1.59	1.96
Mean	1.473	3.063	3.093	2.562
3rd Qu.	1.615	2.665	2.125	2.93
Max.	3.7	19.21	84.37	8.15
No. Instances	47	23	107	12
Total	*189*			

Table 6.5: Statistics of within-speaker long silences categories.

Within-Speaker Silence	**CS**	**H**	**W-Oth**
Min.	1	1.01	1.02
1st Qu.	1.13	1.1	1.32
Median	1.36	1.42	1.63
Mean	2.018	2.916	1.638
3rd Qu.	1.76	2.63	2.06
Max.	53.21	27.92	2.22
No. Instances	145	29	9
Total	*183*		

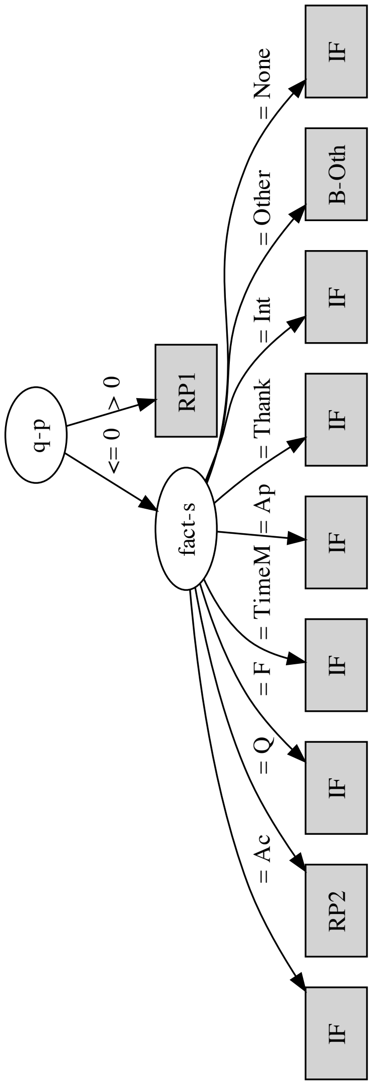


Figure 6.9: *Decision Tree* presenting the features and its values to categorize silence function in *Between speaker silence*.$-p$ represents preceding segment; $-s$ represents succeeding segment

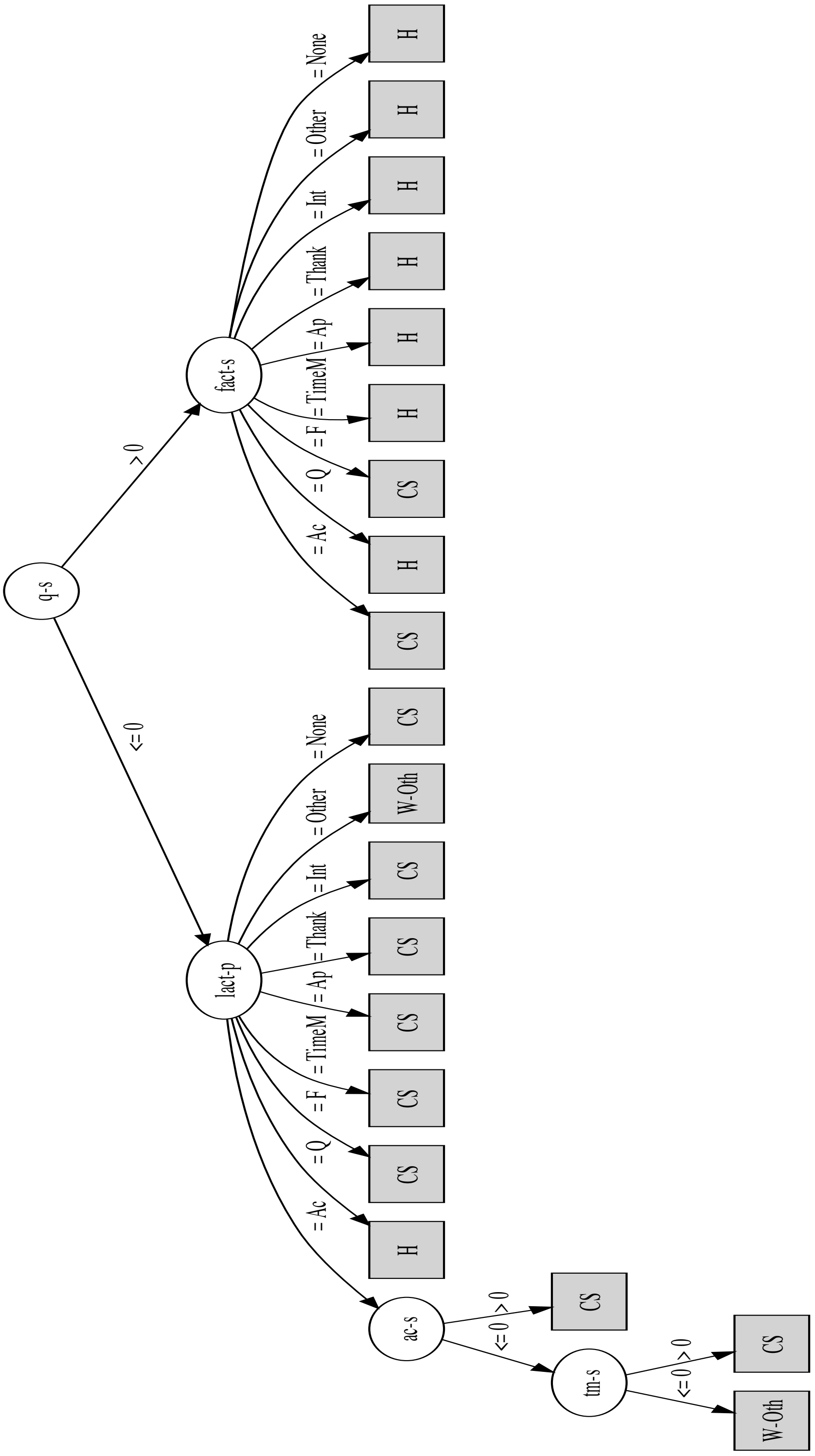


Figure 6.10: *Decision Tree* presenting the features and its values to categorize silence function in *Within speaker silence*.$-p$ represents preceding segment; $-s$ represents succeeding segment

6.6 Summary

The main focus of this chapter is to understand the functions of long silence in within and between-speaker, towards the information flow in a conversation. In an attempt to find such functions, this chapter utilize the sequences of dialog acts present in the left and right context (concerning speaker turns) surrounding the silence itself and design feature vector for individual long silence. These designed feature vectors are later used to cluster silences using a well-known hierarchical concept formation system (Cobweb), which is designed to model different aspects of human concept learning. Followed by the clustering, we grouped the clusters into functional categories of long silence and studied their significance, and duration distribution while classifying using a decision tree. From the study, the observed functions of silence varies from response preparation to hesitation to ask about some queries. It is also observed that sometimes this long silences are used deliberately to extract a forced response from another speaker. It can also indicate the indecisiveness of the current speaker. Even though most of the research from speech communities ignore the silences but our observation shows that by considering the function of long silences, we can better understand the information flow in the conversation, as silence do contribute in explaining the information presented by the speech signals. Silence also has the potential to explain long term behavioral traits and short term states. This study is our first attempt to understand and categorize functions of long silence in a dyadic conversation and there is still more research needed to be done.

Chapter 7

Turn Segmentation and Discourse Labeling Systems

7.1 Introduction

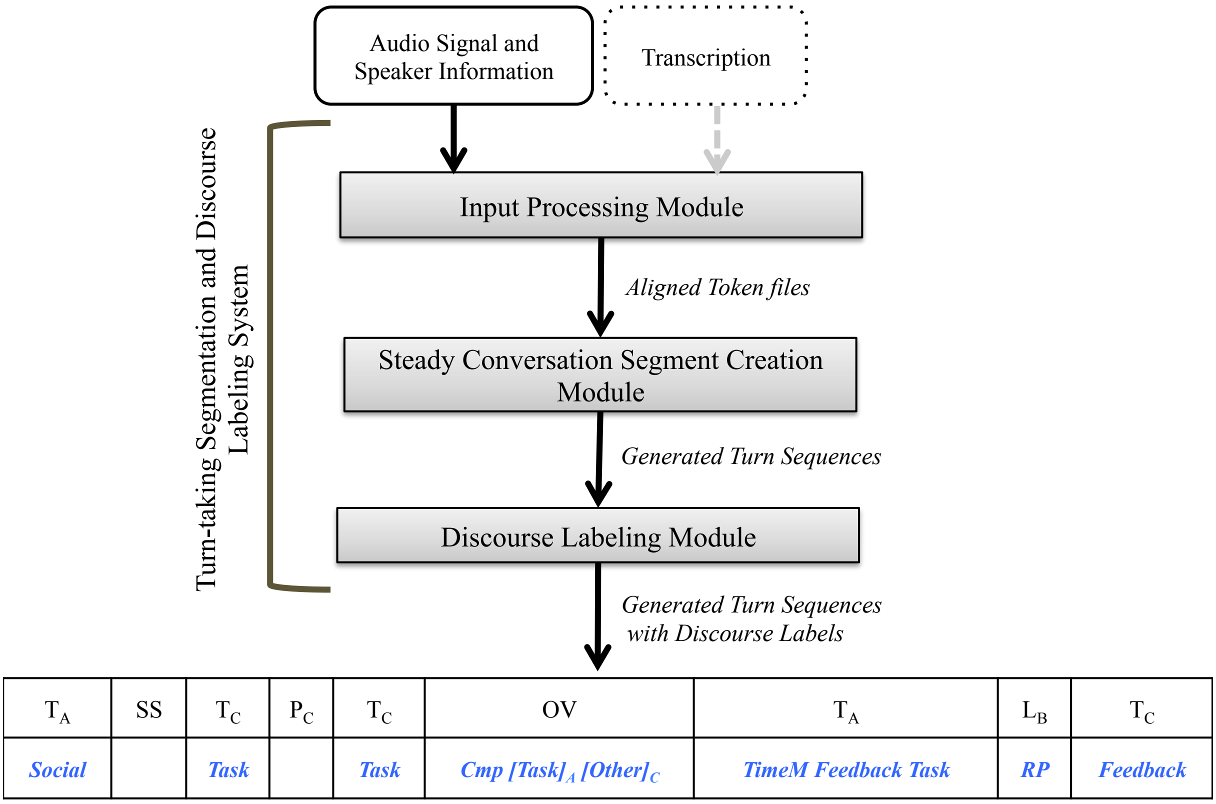
Speech is the primary medium of human communication. With the expansion of the call center industry, spoken conversation data is being generated in overwhelming amounts. Large corporations often outsource their customer support and hosting call centers either monitor the calls in real time or record them for later review. Human reviewers can evaluate only a small random portion of the data (much less that 1%). However, they are required to produce reports addressing various aspects of the service they are providing. These manual evaluation and analysis services are very expensive and do not scale to the quantity of data generated by call centers.

Therefore, to understand and summarize different behavioral aspects unfolding in the conversation, we need to look deep into the flow and dynamics of turn-taking and intent behind each action. To do so, we first need to align each speaker channel, followed by segmentation of the conversation into individual speaker's turn and then creating the steady (uniform) timeline for both the interlocutors to find out who is speaking when, i.e., *segmentation and labeling* of the basic turn-taking scenario in the conversation. To summarize the intent of the speakers (both agent and the customer) we need information regarding the actions each corresponding turn are performing, i.e., *discourse of the event*.

Using the generated information about the turn dynamics we can design descriptive summaries of behavior for each speaker and their contribution to the conversation.

As mentioned earlier that manual evaluation or preparation of data is very expensive, therefore the input of the system has to be designed in such way that reduces human effort.

Contents published in:
Shammur Absar Chowdhury, Evgeny A. Stepanov and Giuseppe Riccardi, *Transfer of Corpus-Specific Dialogue Act Annotation to ISO Standard: Is it worth it?*, in Proc. of 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portorož (Slovenia)
E. A. Stepanov, B. Favre, F. Alam, S. A. Chowdhury, K. Singla, J. Trione, F. B'echet, G. Riccardi, *Automatic Summarization of Call-center Conversations, IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, Scottsdale, Arizona, USA, 2015


Audio Signal and Speaker Information
Transcription
Turn-taking Segmentation and Discourse Labeling System
Input Processing Module
Aligned Token files
Steady Conversation Segment Creation Module
Generated Turn Sequences
Discourse Labeling Module
Generated Turn Sequences with Discourse Labels

T_A	SS	T_C	P_C	T_C	OV	T_A	L_B	T_C
Social		***Task***		***Task***	***Cmp [Task]$_A$ [Other]$_C$***	***TimeM Feedback Task***	***RP***	***Feedback***

Figure 7.1: System Architecture for Automated Turn-Taking Segmentation and Discourse Labeling System with audio signal and speaker information only (and trascription – if available).

Therefore, in this chapter, we present a framework of an automated pipeline, which can align, segment a conversation into low-level turn-taking behavior and extract the discourses behind the events. The resulted framework gives us the output of the segmented turns with discourse label and a descriptive statistic of the conversation.

7.2 System Architecture

In Figure 7.1, we present the system architecture for segmenting and labeling turn discourse of a conversation. As an input to the system, an audio and speaker information[1] of a conversation is sufficient in the absence of any transcription. The system then process the input using *Input Processing Module*, as described in Section 7.2.1; followed by creating aligned speaker channels and turn sequences using *Steady Conversation Segment Creation Module*, outlined in section 7.2.2; this sequence is then passed through *Discourse Module*, in Section 7.2.3, to label the function of each event.

[1] Includes which channel belongs to the agent and which one is the customer.

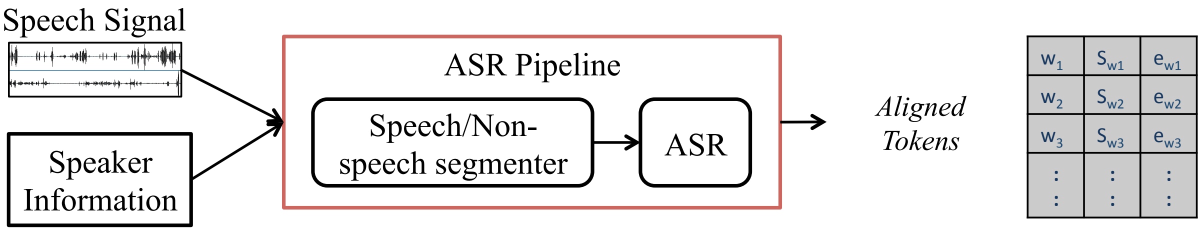


Figure 7.2: Input Processing Module pipeline for scenarios where only audio signal is available.

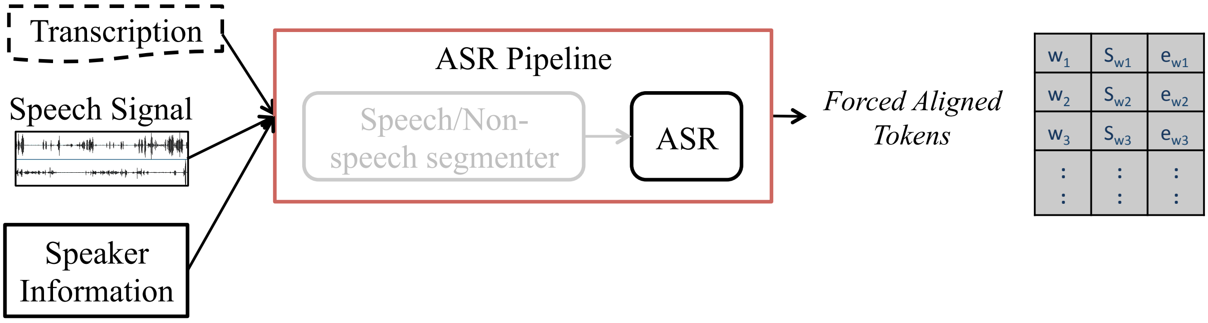


Figure 7.3: Input Processing Module pipeline for scenarios where both audio signal and transcription are available.

7.2.1 Input Processing Module

The input to the system can contain:

- Audio Signal and speaker information only
- Audio Signal along with transcription and speaker information

In the absence of transcription, the input audio signal is first passed to a Speech/Non-speech segmenter, described in Section 7.2.1.1 to extract the segments containing only speech signal from both the channel. Then these speech segments are passed into an Automatic Speech Recognition (ASR), details in 7.2.1.2, to extract the time aligned automatic transcription, containing start end pointers for each word present in the segments. The architecture of such input processing module is presented in Figure 7.2.

In the presence of transcription of the audio, the system does not use the speech/non-speech segmenter but only use the ASR to create forced aligned token files of the conversation, as shown in Figure 7.3.

7.2.1.1 Speech/Non-Speech Segmentation

An in-house speech *vs* non-speech segmenter has been designed using a set of 150 conversations, 300 wave files, containing approximately 100 hours of spoken content and used

Kaldi [200] for the training and decoding process. Training data has been prepared using force-aligned transcriptions. Mel Frequency Cepstral Coefficient (MFCC) and their derivatives have been used as features. Number of Gaussian and beam width have been optimized using a development set of 50 conversations, 100 wave files. The final model has been designed using 64 Gaussians and a beam width of 50, which has been tested using a test set of 50 conversations, 100 wave files. As a part of the post-processing, three rules has been applied: 1) removed non-speech segments, which are between speech segments and are less than 1 second, 2) added an non-speech segment between speech segments if there is a gap greater than 3 seconds, 3) concatenated the consecutive speech and non-speech segments, respectively. The F-measure of the system was 66.0% on the test set. We use the term *SISL speech segmenter* to refer to the segmenter mentioned here.

7.2.1.2 Automatic Speech Recognition System

The automatic speech recognizer was designed using the dataset described in Section 3.1.1 containing approximately 100 hours of spoken content and a lexicon of $\sim$15000 words. Mel Frequency Cepstral Coefficient (MFCC) features have extracted from the conversations and then spliced by taking three frames from each side of the current frame. It was followed by Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature-space transformations to reduce the feature space. Then, the acoustic model was trained using speaker adaptive training (SAT). In order to achieve a better accuracy Maximum Mutual Information (MMI) was also used. The Word Error Rate (WER) of the ASR system was 30.86% (with 40.08% in customer channel and 22.97% in agent channel) on the test set and 20.87% on the training set, using a trigram language model of perplexity 86.90. For the training and decoding process, an open-source implementation system, Kaldi [200], was used.

7.2.2 Steady Conversation Segment Creation

The system then uses aligned lexical information to create the sequence of turn-taking. The pipeline uses the time aligned output (tokens), to create Inter-Pausal Units (IPUs) for each input channel. IPUs are defined as the consecutive tokens with no less that 50 ms gaps in between.

Using the time information of inter-IPUs and intra-IPUs, we then defined **steady conversation segments** where each segment maintain a steady timeline for both interlocutors channel. This step is done by creating a sorted list of all the start and end time of IPUs from both channels. This sorted list is then further used to create segments, also referred as *steady segments*, of the conversation with a uniform timeline in both channels. Then each IPUs in both channels are aligned with the steady segments to form a binary representation of the information in each steady segment for each channel as shown in Figure 7.4. The binary value for each segment

s_i, in agent-channel x (Equation 7.1), and in customer-channel y (Equation 7.2), along with the duration of each segment d in second, is then passed to through a function, $f(x, y, d, t)$ as shown in Equation 7.3, to define the label of the steady segments.

$$x = \{0, 1\} \tag{7.1}$$

$$y = \{0, 1\} \tag{7.2}$$

$$f(x, y, d, t) = \left\{\begin{array}{llllllll}
T_A & x = 1 & y = 0 & & & & & \\
T_C & x = 0 & y = 1 & & & & & \\
Ov & x = 1 & y = 1 & d >= 0.2 & & & & \\
P_A & x = 0 & y = 0 & x_{t-1} = 1 & x_{t+1} = 1 & y_{t-1} = 0 & y_{t+1} = 0 & d >= 0.5 \\
T_A & x = 0 & y = 0 & x_{t-1} = 1 & x_{t+1} = 1 & y_{t-1} = 0 & y_{t+1} = 0 & d < 0.5 \\
P_C & x = 0 & y = 0 & x_{t-1} = 0 & x_{t+1} = 0 & y_{t-1} = 1 & y_{t+1} = 1 & d >= 0.5 \\
T_C & x = 0 & y = 0 & x_{t-1} = 0 & x_{t+1} = 0 & y_{t-1} = 1 & y_{t+1} = 1 & d < 0.5 \\
L_W & x = 0 & y = 0 & x_{t-1} = 0 & x_{t+1} = 0 & y_{t-1} = 1 & y_{t+1} = 1 & d >= 2.0 \\
L_W & x = 0 & y = 0 & x_{t-1} = 1 & x_{t+1} = 1 & y_{t-1} = 0 & y_{t+1} = 0 & d >= 2.0 \\
L_B & x = 0 & y = 0 & x_{t-1} = 1 & x_{t+1} = 0 & y_{t-1} = 0 & y_{t+1} = 1 & d >= 2.0 \\
L_B & x = 0 & y = 0 & x_{t-1} = 0 & x_{t+1} = 1 & y_{t-1} = 1 & y_{t+1} = 0 & d >= 2.0
\end{array}\right. \tag{7.3}$$

The labels of each segment are then defined by a set of rules, shown in Equation 7.3. Labels of the segments are as follows:

- Turn (T): Maximal sequences of IPUs where one single speaker has the floor, and none of the IPUs from the interlocutor are present [201]. T_A and T_C represent agent and customer's turns respectively.

- Pause (P): Gaps between the turns of the same speaker with no less than 0.5 seconds, as shown in Figure 7.5. P_A and P_C represent agent and customer's pauses respectively.

- Overlaps (Ov): Overlapping turns between the two interlocutors.

- Lapse between speakers (L_B): Floor Switches between the speakers with a silence duration of 2 seconds or more.

- Lapse within speaker (L_W): Gaps between a speakers' turns with a silence duration of 2 seconds or more.

- Switch (S): Floor switches (including gaps) between the speakers with silence less than 2 seconds or with overlapping frames not more than 20 ms.

The generated steady turn sequences along with the speech signals are then passed to Discourse Labeling Module (DLM) for extracting intent of action behind each turn event.

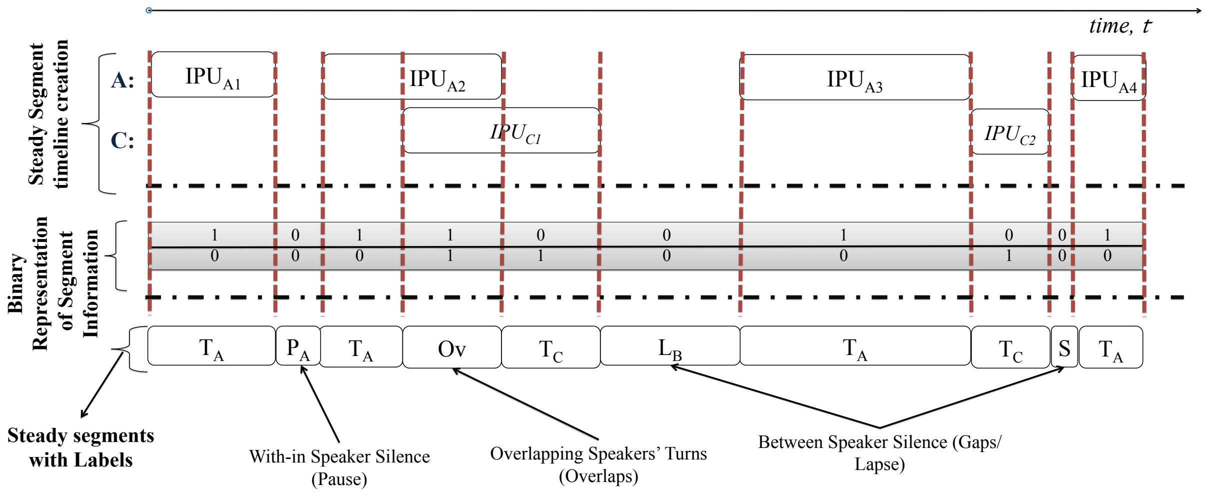

time, t
Steady Segment timeline creation
A:
C:
IPU_A1
IPU_A2
IPU_A3
IPU_A4
IPU_C1
IPU_C2
Binary Representation of Segment Information
1 0 1 1 0 0 1 0 0 1
0 0 0 1 1 0 0 1 0 0
T_A P_A T_A Ov T_C L_B T_A T_C S T_A
Steady segments with Labels
With-in Speaker Silence (Pause)
Overlapping Speakers' Turns (Overlaps)
Between Speaker Silence (Gaps/ Lapse)

Figure 7.4: System architecture for *steady conversation segment creation* module.

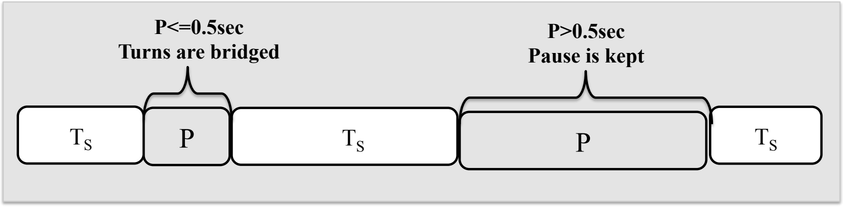


Figure 7.5: Illustration of bridging pauses within a conversation

7.2.3 Discourse Labeling Models

In conversation, turn-taking events by the speakers, are designed to convey something. To find the motivation or intention behind each action, the automatic pipeline passes the steady segment labels to its corresponding computation model. The models included in Discourse Labeling Models, DLM are: 1) Dialog Act model – including segmenter and classifiers for both dimension and communicative functions present in a speaker turn. 2) Overlap Discourse Categorization model – automatic overlap labeling system includes **Competitive (Cmp)** and **Non-Competitive (Ncm)** categories and 3) Functional module for long silence. The next sections explain each model that are integrated into the DLM.

7.2.3.1 Dialog Act Model

Dialogue Acts (DA) are fundamental for the analysis of conversations: they carry communicative functions such as question, answer, expression of agreement and disagreement, etc.. Consequently, the range of applications of DA analysis is quite wide and includes conversation summarization (both spoken and written), dialog systems, etc.; and DAs have been extensively studied in both theoretical and computational linguistics. The supervised and unsupervised annotation and classification of DAs (e.g. [202]) and cross-domain and cross-media classification (e.g., forums, email, and spoken conversations [202, 203]) have been shown to yield good results.

A subset of 50 dialogs from Italian LUNA Human-Human corpus [152], described in Section 3.2 was annotated with dialog acts. The LUNA DA annotation scheme was inspired by DAMSL [154], TRAINS [155], and DIT++ [156]. The most common 15 dialog acts from these taxonomies are grouped into three categories [152]: *Core Dialog Acts* (8) are main actions in the dialog, such as request of information, response, or performing the task; *Conventional/Discourse Management Acts* (4) are utterances such as greetings, apologies, etc. whose function is to maintain general dialog cohesion; *Feedback/Grounding Acts* (3) are utterances whose function is to acknowledge, provide feedback, or just time fillers; and *Others* (1) to capture the rest. The unit of annotation for dialog acts in LUNA Corpus is an utterance. However, due to the overlapping turns (both speakers speaking), an utterance can span several turns. Thus, the dialog act annotation was preceded by additional utterance segmentation. The already annotated dialog act was semi-automatically re-annotated with the recently accepted international ISO standard for DA annotation – Dialog Act Markup Language (DiAML) [1, 157] (see Section 3.2.1.3 for details).

The DiAML annotation scheme [1] is illustrated in Figure 3.13. The DiAML annotation scheme consists of 56 core DA tags[2] (communicative functions), organized into 9 dimensions: 26 general (applicable to any dimension) and 30 dimension specific [158] (see Table 6.1 in Chapter 6 for a set of dimensions and communicative functions considered for LUNA Corpus re-annotation). In the following section we present our approach to dialog act segmentation and classification and the results obtained on both LUNA Corpus and SISL corpus.

Segmentation

The designed automatic dialog act segmenter takes speaker turns as an input. The automated segmenter then extracts token/word with a context of ±2 as the feature and uses a discriminative approach, namely Conditional Random Fields (CRFs), [204] to simultaneously segment the turn into its DA boundaries, in IOB format. The results of such system are shown in Table 7.1.

Table 7.1: Precision (P), recall (R) and F1 of dialog act segmentation

DA Segmenter	**P**	**R**	**F1**
Overall	0.73	0.59	0.65

[2] In the literature the number of dimensions and dimension specific communicative functions varies.

Table 7.2: Comparative evaluation of ‘legacy’ and ISO annotations at dimension level. F_1 for in-domain data using LUNA corpus (**LUNA**), cross-domain SISL corpus (**SISL**) evaluation settings.

	Legacy		**ISO**	
Dimension	**LUNA**	**SISL**	**LUNA**	**SISL**
Task	0.79	0.72	0.78	0.74
Social	0.86	0.66	0.84	0.78
Time+Fb	0.71	0.61	0.73	0.64
Other	0.18	0.15	0.24	0.22
Micro	0.72	0.60	0.72	0.67

Classification

For the dialog act classification, we use Sequential Minimal Optimisation (SMO), a support vector machine implementation, with its linear kernel and default parameters [191]. As for the features, the experiment only utilized bag-of-words representation extracted from dialog act span tokens.

As it was already mentioned, the ‘legacy’ and ‘ISO’ annotations are evaluated in two settings: (1) in-domain, i.e., using LUNA corpus, (2) cross-domain, i.e., using SISL corpus. We perform classification into dimensions and into communicative functions, using bag-of-words representation for features. The distribution of labels in each layer (dimensions and communicative functions) is unbalanced (see Table 3.7); however, we do not address balancing issues. For consistency with the ‘legacy’ annotation and for comparing the results, we merged *Feedback* and *Time Management* dimensions. The *Social Obligations Management* dimension was kept separate. Performance is evaluated using standard precision, recall and F_1.

The results of the experiments on the dialog act classification at dimension level are reported in Table 7.2 as F_1. In dimension level classification, the number of classes (dimensions) is the same for the ‘legacy’ and ISO annotated data. Due to the segmentation differences, the number of instances, however, is different. The results illustrate that in-domain performances of the two annotation schemes are comparable; however, ISO annotation scheme has better performance in the cross-domain.

Communication function level classification settings are different for the ‘legacy’ and ISO annotated data: for the former it is classification into 16 classes, and for the latter into 41 class. To evaluate the ISO scheme in more comparable settings, we additionally evaluate it after mapping to the ‘legacy’ annotation (i.e., to 16 ‘legacy’ classes) using mappings in Table 3.6 in Chapter 3. The results of the experiments on the dialog act classification at communication

Table 7.3: Comparative evaluation of ‘legacy’ and ISO annotations at communicative function level. F_1 for in-domain (**LUNA**), cross-domain SISL corpus (**SISL**) evaluation settings. For each annotation scheme, the number of communicative functions is reported in parentheses. **ISO Mapped** reports performance of the ISO annotations after mapping to the ‘legacy’ annotation.

	Legacy (16)		**ISO** (41)		**ISO Mapped** (16)	
Dimension	**LUNA**	**SISL**	**LUNA**	**SISL**	**LUNA**	**SISL**
Task	0.31	0.20	0.24	0.27	0.35	0.39
Social	0.64	0.39	0.60	0.41	0.70	0.52
Time+Fb	0.68	0.55	0.84	0.63	0.84	0.62
Other	0.25	0.26	0.22	0.26	0.24	0.36
Micro	0.44	0.30	0.40	0.37	0.47	0.45

function level are reported in Table 7.3 as F_1. Individual communication function performances are aggregated to dimension level and reported numbers are micro-averaged F_1s.

7.2.3.2 Overlap Discourse Categorization Model

To automatically label overlap instances these two categories of overlaps, **Competitive (Cmp)** and **Non-Competitive (Ncm)** categories, we use an in-domain overlap categorization model [177]. The model was trained using acoustic features with the left and right context of 0.2 and 0.3 seconds of speech. The overall F-measure of the system using acoustic features is 64.36% on the test set as reported in [177]. Details of the techniques and evaluation are given in Chapter 5.

7.2.3.3 Silence Function Module

For assigning functions of silence instances, both between and within-speaker silence, we used J48 decision tree discussed in Chapter 6. The tree takes silence instance as input which is $>= 1$ second. For between speaker silences, the function labels are: response preparation function (**RP1** and **RP2**), information flow functions (**IF**) such as conversational and deliberate silence and the other (**B-Oth**). Similarly for within-speaker long silence the functional categories includes: indecision or hesitation silence (**H**), organizational silence (**CS**) and other categories (**W-Oth**). Details of functions can be found in Chapter 6.

7.3 Summary

In this chapter, we presented a designed a framework that can automatically segment the turns and turn-taking events, such as silence, and categorizes its discourse labels. Such a system can work even when an audio signal is available as the only input to the system. The system uses state-of-the-art ASR pipeline with other discourse module researched in this dissertation (including dialog act segmenter and classifier, and overlap discourse model) to label the turn-taking behavior in the conversation. This framework can also be used to predict different characteristics of conversation on the real-time scenario. Such usage of the pipeline are presented in the next following chapters.

Chapter 8

Modeling User Satisfaction with Turn-Taking

User satisfaction is an important aspect of the user experience while interacting with objects, systems or people. Traditionally user satisfaction is evaluated a-posteriori via spoken or written questionnaires or interviews. In automatic behavioral analysis we aim at measuring the user emotional states and its descriptions as they unfold during the interaction. In our approach, *user satisfaction* is modeled as the final state of a sequence of emotional states and given ternary values `positive, negative, neutral`. In this chapter, we investigate the discriminating power of turn-taking in predicting user satisfaction in spoken conversations. Turn-taking is used for discourse organization of a conversation by means of explicit phrasing, intonation, and pausing. In this study, we train different characterization of turn-taking, such as competitiveness of the speech overlaps. To extract turn-taking features we design a turn segmentation and labeling system that incorporates lexical and acoustic information. Given a human-human spoken dialog, our system automatically infers any of the three values of the state of the user satisfaction. We evaluate the classification system on real-life call-center human-human dialogs. The comparative performance analysis shows that the contribution of the turn-taking features outperforms both prosodic and lexical features.

8.1 Introduction

A satisfying communication plays an important role in social interaction such as multiparty and dyadic conversations in call-center, doctor-patient, and student-teacher scenarios. Over the years, user satisfaction has been evaluated using spoken or written questionnaires and interviews. In such an evaluation, users are usually asked to fill up questionnaires or rate certain aspects of a conversation that address users' satisfaction, as reported in [15]. User satisfaction has been addressed in other research fields as well – consumer satisfaction with products [205] and Spoken Dialog Systems (SDS) such as problem-solving [16] and tutoring [17]. In SDS, user satisfaction is used as one of the metrics to assess the quality of a dialog system [18, 19]. Thus,

Contents published in:
Shammur Absar Chowdhury, Evgeny A. Stepanov and Giuseppe Riccardi, *Predicting User Satisfaction from Turn-Taking in Spoken Conversations*, in Proc. of Interspeech-2016, San Francisco, USA.

the increasing importance of user experience as a quality assessment demands a computational model for observed user satisfaction rather than self-reported measure.

In a natural conversation, parallel to the exchange of information, there is also a flow of speakers' emotional states, unfolding with or without any intent. A sequence of emotional states manifested during a conversation is a strong cue for predicting user experience. The goal of this study is to exploit these sequences of emotional states, specifically the final state, to model user satisfaction. For the automatic prediction of the user satisfaction, the final emotional states are categorized into three labels as **Positive** (Pos), **Negative** (Neg), and **Neutral** (Neu). We investigate how the organizational structure of a conversation, such as turn-taking, contributes to the prediction of user satisfaction along with other more common levels of conversation description such as lexical and prosodic.

Turn-taking is a remarkable phenomenon that is fundamental for human communication [206]. Over decades the intriguing cues of turn-taking attracted researchers from conversational analysis, linguistics, psycholinguistics, and speech. One of the first studies on turn-taking was conducted by [4], where turn-taking is defined as a way to signal and perceive cues for Transition Relevance Place (TRP). The authors also suggest that the transition from the current speaker to the next should occur very frequently with minimum gap or overlap in speech. In [4,35], overlaps have been considered as a violation of the fundamental rule, but the authors in [5] suggest that about 40% of all between-speaker intervals are overlaps. It has been proposed that speech overlaps relate to the dominance, aggression, competitiveness or cooperativeness towards the other speaker [7, 8, 177]. Other relevant studies include overlap detection [207, 208] (including word-level as overlap vs. clean-speech [209]), interruption detection [82], and studies on types of turn-taking and their correlation with speakers' turn-taking behavior [206].

Considering the literature on overlaps and turn-taking in spoken conversations, competitiveness and non-competitiveness of the speaker turns did not receive enough attention. Among the few, [210] demonstrate the importance of the onset position of the overlap along with the temporal features. On the other hand, in [56], the author argue that overlap is better described by the phonetic design rather than its precise location; which is later supported by [46, 77].

Previous work on incorporating turn-taking with social signals have mainly focused on group dynamics or task-oriented dialogs, like modeling participant's affects from turn-taking with post-meeting ratings [211] or studies about participant's involvement or interest [212,213].

To the best of our knowledge, turn-taking has not been utilized for predicting user satisfaction as emotional manifestation. Hence, in this study, we focus on turn-taking features for predicting user satisfaction; to achieve this goal we are:

- modeling the state of the user satisfaction in terms of the final emotional manifestation of the customer.

- automatically predicting the state of the user satisfaction as Pos, Neg, Neu, using the lexical, prosodic and turn-taking feature sets.
- designing a turn segmentation and labeling system by utilizing automatic transcriptions and acoustic features, to extract turn-taking features.
- comparatively evaluating and analyzing the turn-taking features to understand their discriminative power.

For the study, we analyzed a large dataset of Italian call-center spoken conversations where customers and agents are engaged in problem-solving tasks, as described in Chapter 3, Section 3.1.2.

The chapter is organized as follows. An overview of dataset preparation is given in Section 8.2. Followed by details of the system framework, extracted features and classification experiments in Section 8.3. Section 8.4 presents the results and analysis of the observations. Summary of the chapter is provided in Section 8.5.

8.2 Data Description

In this study, we consider a corpus of 1894 call-center conversations [147], collected over the course of six-months (210 hours of speech, with an average length of 406 seconds per conversation). The conversations were recorded on two separate channels with 16 bits and 8kHz sampling rate.

The corpus was annotated for basic and complex emotions following the *modal model* of emotions developed by Gross [142, 144]. The model emphasizes the attentional and appraisal acts underlying the emotion-arousing process. For the annotation, the considered basic emotion was *anger*; and the complex social emotions were *satisfaction, dissatisfaction, frustration* and *empathy*. Empathy was annotated for the agent channel only; the rest of emotions for the customer channel. The inter-annotator agreement of the annotation has kappa = 0.74 (additional details of the annotation process can be found in [148]).

A subset of 739 conversations ($\approx$ 86 hours) was selected such that conversations annotated with customer emotion has also been annotated with empathy in the agent channel.

With respect to the annotation, the final manifested emotional state can be satisfaction, anger or frustration, or there might be no emotional manifestation. As shown in Figure 8.1, we define three labels for modeling *user satisfaction* concerning the final emotional state in the conversations. *Positive*, **Pos** is used for the conversations where the final emotional manifestation of the customer is satisfaction. Satisfaction may be the only manifested emotion in the customer channel ($S1$) or it may come as a results of a change from anger or frustration due to agent's manifestation of empathy ($S2$); thus, yielding a sequence Customer: Anger/Frustration

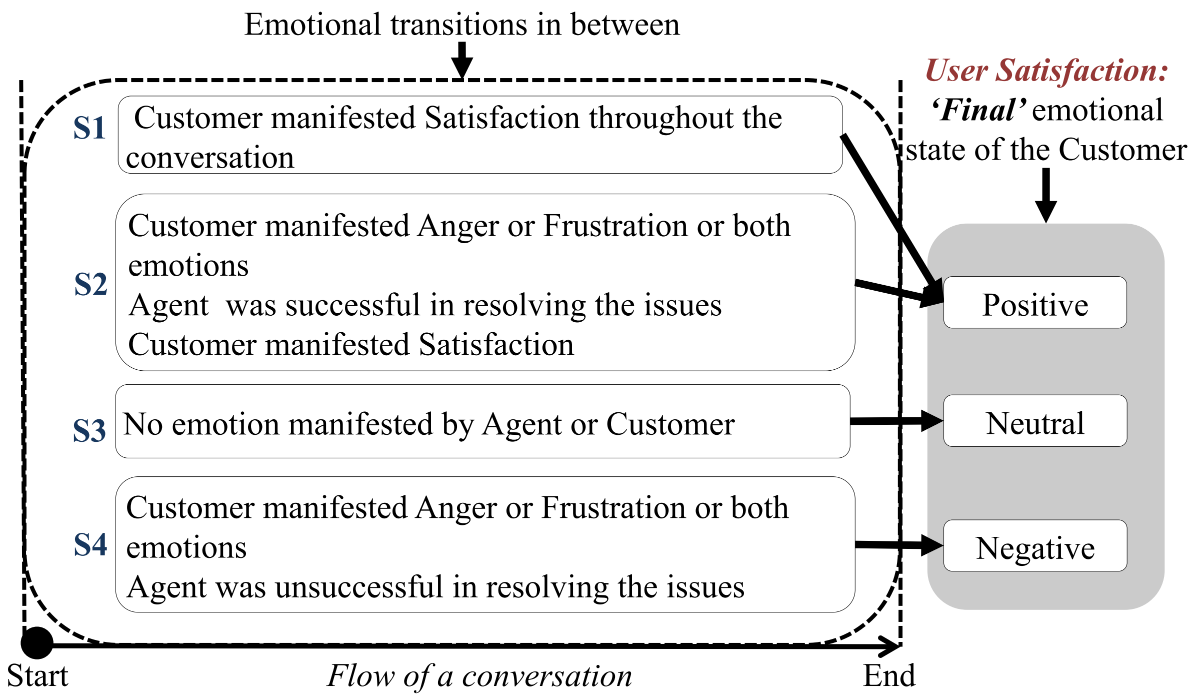

Emotional transitions in between
User Satisfaction:
'Final' emotional state of the Customer
S1
Customer manifested Satisfaction throughout the conversation
S2
Customer manifested Anger or Frustration or both emotions
Agent was successful in resolving the issues
Customer manifested Satisfaction
S3
No emotion manifested by Agent or Customer
S4
Customer manifested Anger or Frustration or both emotions
Agent was unsuccessful in resolving the issues
Positive
Neutral
Negative
Start
Flow of a conversation
End

Figure 8.1: Different scenarios of emotional manifestation with associated class labels representing user satisfaction.

Table 8.1: Train, Dev and Test set split and their distribution for the prediction task.

Sets	**Pos (%)**	**Neg (%)**	**Neu (%)**	**Total(%)**
Train	205 (34.0%)	198 (32.84%)	200 (33.17%)	603 (100%)
Dev	21 (30.43%)	22 (31.88%)	26 (37.68%)	69 (100%)
Test	19 (28.36%)	25 (37.31%)	23(34.33%)	67 (100%)

→ Agent: Empathy → Customer: Satisfaction. *Negative*, **Neg** is used for the conversations where the final emotional manifestation of the customer is either anger, frustration or both ($S4$). The conversations that do not have any emotional manifestations are labeled as *Neutral*, **Neu** ($S3$). The split of the data into training, development and test sets are given in Table 8.1.

8.3 System Framework

In Figure 8.2, we present a pipeline for predicting the state of the user satisfaction, which takes an audio and speaker information of a conversation as an input. The speech signals are then passed through Automatic Speech Recognition (ASR) pipeline, which consists of a speech vs. non-speech detector and domain-specific ASR. Each detected speech segment is passed to the ASR [176]. The time aligned output of the ASR along with speech signal is then used to

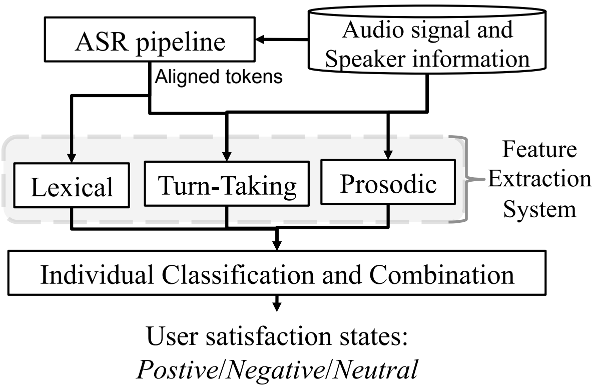

ASR pipeline
Audio signal and Speaker information
Aligned tokens
Lexical
Turn-Taking
Prosodic
Feature Extraction System
Individual Classification and Combination
User satisfaction states:
Postive/*Negative*/*Neutral*

Figure 8.2: Computational system for classifying the state of user satisfaction.

extract turn-taking, lexical and prosodic features.

The individual feature sets – lexical, prosodic, and turn-taking – are then used to train and evaluate classifiers. Additionally, we perform feature-level and decision-level fusion. For decision-level fusion, we are using weighted majority voting, where the weight of each classifier is the overall F1 of the system on dev set. Moreover, to understand the discriminative characteristics of the turn-taking features, they are analyzed using logistic regression model.

8.3.1 Feature Extraction

8.3.1.1 Turn-Taking Features

The Turn-Taking Feature Extraction System, described in Figure 8.3, consists of a *turn segmentation and labeling system* and the *feature generation* step. The system uses lexical and acoustic information to extract the features. The pipeline uses the time aligned ASR output as tokens to create Inter-Pausal Units (IPUs) for each input channel. IPUs are defined as the consecutive tokens with no less that 50 ms gaps in between. Using the time information of inter-IPUs and intra-IPUs, we defined **steady conversation segments** where each segments maintain a steady timeline in both interlocuters channel. The labels of each segment are then defined by a set of rules. Labels of the segments are as follows:

- Turn (T): Maximal sequences of IPUs where one single speaker has the floor, and none of IPUs from the interlocutor are present [201]. T_A and T_C represent agent and customer's turns respectively.
- Pause (P): Gaps between the turns of the same speaker with no less than 0.5 sec. P_A and P_C represent agent and customer's pauses respectively.
- Overlaps (Ov): Overlapping turns between the two interlocutors.

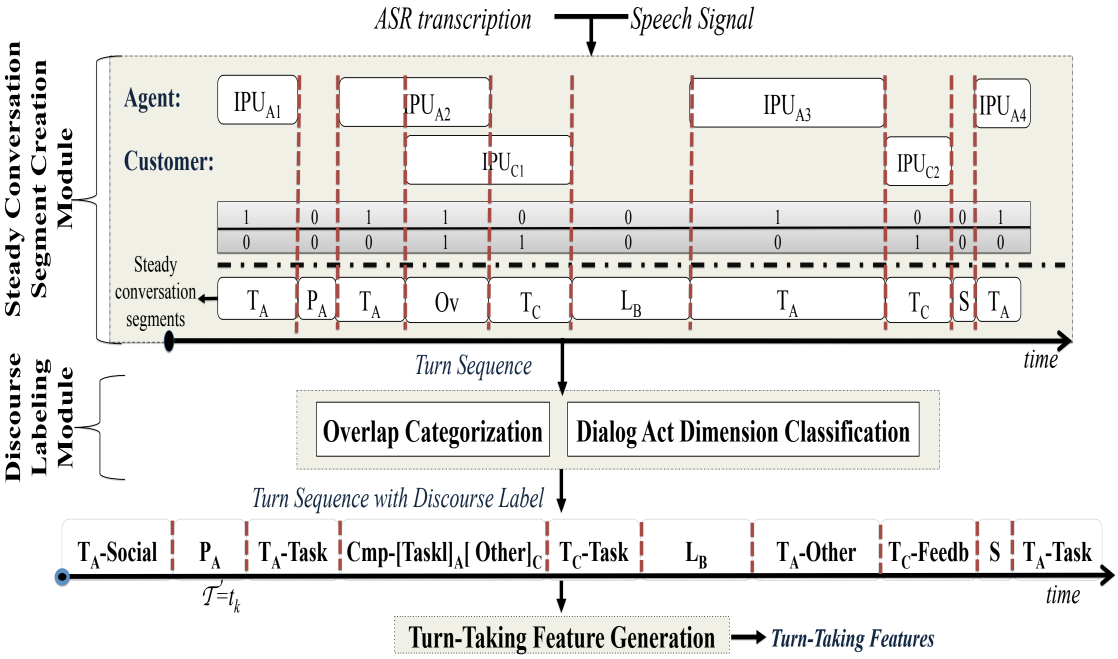


Figure 8.3: Schematic diagram of automated Turn-Taking Feature Extraction System with speech signal and asr transcription as input. $\boldsymbol{T_A}$, $\boldsymbol{T_C}$, $\boldsymbol{P_A}$: agent and customer's turn and Pause, $\boldsymbol{Ov}$:overlaps, $\boldsymbol{L_B}$:Lapse between speakers, $\boldsymbol{S}$:Smooth switch, $\boldsymbol{T_{A/C} - DA}$: $T_{A/C}$ with DA, Dialog Act dimension, where $DA \in \{Social, Task, Feedb, Other\}$, **Cmp**: Competitive overlap.

- Lapse between speakers (L_B): Floor Switches between the speakers with a silence duration of 2 sec or more.
- Lapse within speaker (L_W): Gaps between a speakers' turns with a silence duration of 2 sec or more.
- Switch (S): Floor Switches between the speakers with silence less than 2 secs or with overlapping frames not more than 20 ms.

The generated turn sequences along with the speech signals are then passed to Discourse Labeling Module (DLM) followed by the Turn-Taking Feature Generation module for extracting turn-taking features.

Discourse Labeling Module: The DLM module includes Overlap Categorization and Dialog Act Dimension Classification systems as described below.

Overlap Categorization: The automatic overlap labeling includes **Competitive (Cmp)** and **Non-Competitive (Ncm)** categories. In Cmp scenario, the intervening speaker starts prior to the completion of the current speaker and both the speakers perceive the overlap as problematic and display interest in the turn for themselves. In Ncm scenario, the intervening speaker starts at the middle of an ongoing turn with no evidence for the intent to grab the turn.

To automatically label these two categories of overlaps we use an in-domain overlap categorization model [177]. The model was trained using acoustic features with the left and right context of 0.2 and 0.3 seconds of speech. The overall F-measure of the system using acoustic features is 64.36% on the test set as reported in [177].

Dialog Act Dimension Classification: To get an overview of the function of each turn in the conversation, we use an in-house developed *dialog act segmenter* and *dialog act dimension classifier* [214]. The labels of output turns are the dimensions of the dialog acts from DiaML ISO specification [1] including dimensions such as: Task (e.g., question, instruct, suggest), Social (e.g., greeting, apology), TimeManagement and Feedback (e.g., stalling, positive-negative feedback), Others or None. The overall F-measure of the system, using bag-of-word features, is 72% (in-domain test set) and 60% (out-of-domain test set).

Turn-Taking Feature Generation: The turn-taking features are generated using the turn sequence output from the DLM module (see Figure 8.3). To understand the impact of overlaps – Cmp vs. Ncm, silence and other predictability factors of turn-taking structure are extracted as turn-taking features at both conversation and individual speaker levels. A brief description of extracted features are as follows:

- Participation equality [215], P_{eq}:

  $$P_{eq} = 1 - (\frac{\sum_i^N (T_i - T)^2/T}{E}) \tag{8.1}$$

  where T is the average speech duration of the speakers. T_i is the total speech duration for each speaker. E represents the total speech duration. $N = 2$, represents two speakers as agent and customer.

- Turn-taking Freedom, as defined in [211], F_{cond}:

  $$F_{cond} = 1 - \frac{H_{max}(Y|X) - H(Y|X)}{H_{max}(Y|X)} \tag{8.2}$$

  where we calculate $H(Y|X)$, the conditional entropy of speaker Y being the next speaker after X begins the turn, $H_{max}(Y|X)$ being the maximal possible value for this. $W = \{agent, customer\}$, $X \in W$, $Y \in W$ and $X \neq Y$.

  The value of F_{cond} is between 0 and 1, where 0 represents a strict turn-taking.

- Percentage of overlaps.

- Percentage of Cmp and Ncm on total overlap duration.

- Percentage of agent's and customer's speech
- Median duration of T_A, T_C, P_A, P_C, Cmp, Ncm, L_W and L_B.
- Probability of speaker X's turn after a Cmp: $P(X|Cmp)$ or Ncm: $P(X|Ncm)$.
- Probability of Cmp after speaker X's turn: $P(Cmp|X)$ or Ncm after speaker X's turn : $P(Ncm|X)$.
- Rates of each dialog act dimension with respect to speaker's speech duration.

8.3.1.2 Prosodic Features

We extracted prosodic features using openSMILE [178] with the frame size of 25 ms and a frame step of 10 ms. These low-level features such as pitch, loudness, and voice-probability together with their derivatives are then projected onto 24 statistical functionals such as mean and range among others. More details of these features are in [179].

We extract the prosodic features for agent and customer channels separately, then linearly merge them to design an equal sized vector for each conversation. Let $\boldsymbol{A_{s1} = \{A_1, A_2, ..., A_m\}}$ and $\boldsymbol{C_{s2} = \{C_1, C_2, ..., C_m\}}$ denote agent and customer channels' feature vectors respectively. The combined feature vector is $\boldsymbol{P_c = \{A_1, A_2, ..., A_m, C_1, C_2, ..., C_m\}}$ with $P_c \in R^{m+m}$.

8.3.1.3 Lexical Features

Lexical features are extracted from automatic transcriptions for the whole conversation from the ASR pipeline. The features are then transformed into a bag-of-words (vector space model) [180], to represent the words as numeric features. For this study, we extracted trigram features, to use the contextual benefit of n-grams. The frequencies in the feature vectors were then transformed into tf-idf values - the product of the logarithmic term frequency (tf) and inverse document frequency (idf).

8.3.1.4 Feature Combination

For this study, we also analyze the combined contribution of the feature sets. As shown in Figure 8.2, after extracting turn-taking, prosodic and lexical features we merge the feature vectors into a single vector and then use that for classification.

8.3.2 Classification and Evaluation

A Sequential Minimal Optimization (SMO), a support vector machine implementation of weka [191], is used to train the classifiers with feature values normalized within $[0, 1]$ interval. Due to the difference between the dimensionality of the feature vectors, we experiment with different kernels such as linear and RBF of SVM on the dev set. As for the evaluation, we report F-measure ($F1$) for individual classes, along with macro-averaged F-measure.

Table 8.2: Classification results for predicting user satisfaction state. Feat.Comb: Feature-level combination, D.Fuse: Decision level fusion, Oracle-D.Fuse: Oracle of D.Fuse. Reported value is F1 measure on the test set.

Experiments	**Pos**	**Neg**	**Neu**	**Overall**
Chance-Baseline	0.24	0.30	0.27	0.27
Lexical	0.44	0.58	0.35	0.48
Prosodic	0.33	0.32	0.52	0.40
Turn-Taking	0.61	0.57	0.62	0.61
Feat.Comb	0.49	0.57	0.55	0.54
D.Fuse	0.57	0.57	0.60	0.59
Oracle-D.Fuse	0.90	0.86	0.80	0.85

8.4 Results and Discussion

In Table 8.2 we present the results for predicting the state of user satisfaction in terms of *Pos*, *Neg* and *Neu*, using individual feature sets and their combination and decision level fusion. For comparison, a random baseline is calculated by randomly generating class labels based on prior class distribution.

It is observed that all the systems have higher performance than the baseline. Regarding overall system **F1**, the turn-taking features outperform all other systems. As for individual classes, turn-taking is noticed to be the best discriminator for Pos and Neu classes and has 1% F1 less in Neg class compared to the lexical feature set. This indicates the potential of lexical features to predict for Neg state of user satisfaction.

It is important to note that we have used the linear kernel of SVM for all the experiments except for turn-taking feature set, for which we used the RBF kernel, tuned on the dev set. The F1 of turn-taking features with linear kernel ($Tt - L$) and an optimized penalty parameter $C = 0.4$ are: Pos: 0.55, Neg: 0.52, Neu: 0.63 and Overall: 0.58. Even with linear kernel the turn-taking feature set exceeds the lexical and prosodic features by 10% and 18%, respectively.

Using feature combination (*Feat.Comb*), we have 6% and 14% improvement over lexical and prosodic feature sets but not over turn-taking feature set. One possible reason could be the fact that these feature sets vary in terms of dimensionality and their representations (dense *vs* sparse). The vector size for turn-taking feature is 34, which is very small compare to prosodic and lexical feature sets. The performance of the individual system is reflected in decision fusion result and the upper bound of decision fusion is shown by Oracle results in Table 8.2.

We use multilevel logistic regression [216], to understand the impact of turn-taking feature for predicting each state of user satisfaction. The result shows a significant positive effect on the presence of non-competitive overlaps and use of social turns by customers in Pos class, while the median duration of T_A has a negative effect. That is, the customer tends to be more satisfied when there is an increase of feedback and social turns flow rather than agent taking long turns. Similarly, the use of the time-management/feedback DA turns decrease the likelihood of the conversation to be Neg significantly, whereas the likelihood of Neg class increases when the percentage of competitive overlaps along with the use of DA-Other by agent increases. In [217], the authors reported that the automatic feature "BargeIns" were highly correlated with user satisfaction, which also supports our findings with Neg class.

8.5 Summary

In this study, we investigate the use of turn-taking in predicting user satisfaction in spoken conversations. We model user satisfaction as the final emotional manifestation of a conversation, which can be either positive, negative or neutral. We extract turn-taking features by designing a turn segmentation and labeling system. We compare turn-taking features with lexical, prosodic feature sets along with feature level combination and decision level fusion. We observe that turn-taking features outperform all other systems. The analysis of turn-taking features suggests that the use of non-competitive turns and social dialog acts increase the chance of a positive user experience, whereas competitive turns tend to decrease the chance of positive experience.

Chapter 9

Coordination between Interlocutors in Emotional Episodes

In this chapter, we aim to investigate the coordination of interlocutors behavior in different emotional segments. Conversational coordination between the interlocutors is the tendency of speakers to predict and adjust each other accordingly on an ongoing conversation. In order to find such a coordination, we investigated 1) lexical similarities between the speakers in each emotional segments, 2) correlation between the interlocutors using psycholinguistic features, such as linguistic styles, psychological process, personal concerns among others, and 3) relation of interlocutors turn-taking behaviors such as competitiveness. To study the degree of coordination in different emotional segments, we conducted our experiments using real dyadic conversations collected from call centers in which agent's emotional state include *empathy* and customer's emotional states include *anger* and *frustration*. Our findings suggest that the most coordination occurs between the interlocutors inside anger segments, where as, a little coordination was observed when the agent was empathic, even though an increase in the amount of non-competitive overlaps was observed. We found no significant difference between anger and frustration segment in terms of turn-taking behaviors. However, the length of pause significantly decreases in the preceding segment of anger where as it increases in the preceding segment of frustration.

9.1 Introduction

Behavioral and social signal processing are emerging interdisciplinary areas of research, which combine social science, psychology, and computer science. The aim of the research is to design computational models for processing human behavioral aspects, which can facilitate different domain experts while counseling, consulting and (or) providing services [2, 3, 218–220]. The idea is to analyze different overt and covert behavioral signals during social interactions and label them with some short and long term functional aspects (i.e., states and traits) in order to quantitatively measure them. The functional aspects include empathy, politeness, agreement, engagement, uncertainty, competitiveness and other typical, atypical, distressed and affective

Contents published in:
Firoj Alam, Shammur Absar Chowdhury, Morena Danieli, Giuseppe Riccardi, *How Interlocutors Coordinate with each other within Emotional Segments?*, COLING, Osaka, Japan, 2016.

social behaviors. Using these short and long term states and traits, one can design an informative behavioral profile of an individual from the daily-life interactions. The measured behavioral profile can help to predict the next behavioral outcome/consequence and/or actions of an individual. This kind of behavioral profile can help domain experts in different application scenarios such as call center, health-care and teacher-student interactions.

In the field of social and psychological science, researchers have been trying to understand these functional aspects for a very long time, however, very recently there are attempts to design automatic computational models for real-world applications. Designing such automatic systems for measuring these behavioral and social functional aspects is still infancy due to many different challenges.

One of the important challenges is to understand how different behavioral cues are associated with one another and how we express them in different interaction scenarios. In this study, *we investigated, the coordination of interlocutors behavior in different emotional segments and how conversational turn-taking dynamics are associated with emotional manifestations of the agent and customer*. For the study, the conversational coordination between the interlocutors is defined as the tendency of speakers to predict and adjust each other accordingly on an ongoing conversation. We explored the coordination in terms of psycholinguistic features, lexical and turn-taking features using correlation analysis, cosine similarity, and regression analysis, respectively. For this study, we analyzed dyadic human-human spoken conversations, collected from the call centers in the domain of after-sale customer care, which has been annotated with turn-taking dynamics and emotional expressions. The turn-taking dynamics include competitiveness of overlaps, pauses, and lapses among others. Emotional expressions has been annotated for agent and customer separately with agent's emotional state include *empathy*, and customer's emotions include *anger* and *frustration*.

It has been a few decades to the study of automatically recognizing emotion in affective computing, which has been done in the lab as well as in real settings. The study includes classifying Ekman's six basic categorical emotions [221] or dimensional levels of emotion such as valence and arousal [222]. Still, there are challenges to make emotion recognition research in its practical use, which includes lack of publicly available realistic databases, issues of fusing multi-modal information, automatic segmentation, robustness in terms of generalizability across the domain, cross-corpus [163, 223]. A detailed overview of emotion recognition research in terms of theories, computation models, and relevant applications is provided in [224].

The study of turn-taking dynamics such as speech overlap has also a long history. One of the first studies on speech overlap, as discussed in [4], suggested that turn changes with overlap is a very rare case and occurs as a result of self-selection, which projects turn endings. Where as a recent study of [5] suggests that overlap is, in fact, a frequent phenomenon and is much more than just a turn-taking signal, which has also been discussed in [177].

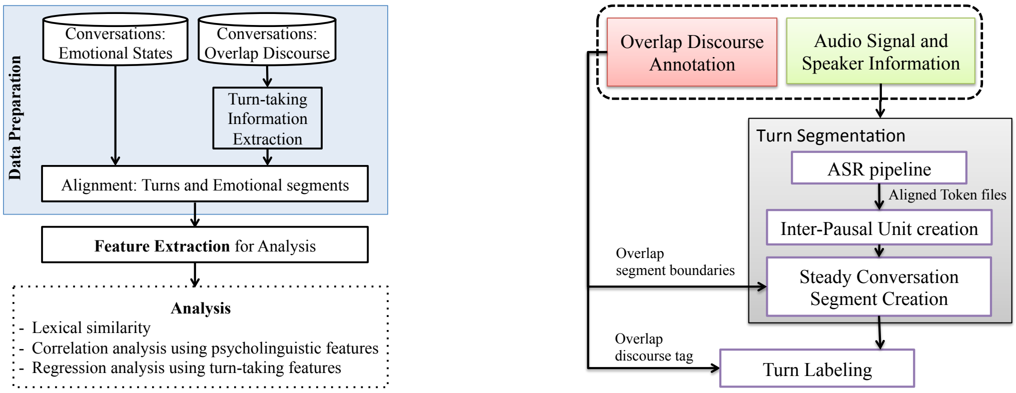
There has been a very few study, which explores finding how different turn-taking features are associated with emotional states. The association of turn-management labels, such as grab, accept, back-channel, and emotional states have been studied in [225]. The importance of turn-taking information for predicting user-satisfaction in terms of user manifested emotion have been studied in [226]. They discussed that turn-taking cues significantly helps in the automatic prediction of user-satisfaction. To the best of our knowledge, a very little study have been conducted to examine what actually happens within an emotional segment in terms of turn-taking. In our study, we present a call center conversation corpus (in Section 9.3) in which we have the manual annotation of emotional states and overlap discourse. Using which we explored the coordination of interlocutors behaviors as our preliminary study, presented in Section 9.2 and 9.6, which can shade a light in future for designing automated computational model.

9.2 Methodology

In Figure 9.1, we present the experimental system of our study. In the data preparation phase, we selected a subset of conversations in which we have annotations of emotional states and overlap discourse. The turn-taking information extraction system utilized an Automatic Speech Recognition (ASR) system [176] to create turn segments and extract turn information (see Section 9.4). Later, this information was aligned with the annotations of emotional segments to find the turn-taking information (more details can be found in Section 9.4.1). Using the aligned turn-taking information for an emotional segment, we extracted turn-taking features. We also used turn information to obtain lexical and psycholinguistic features per speaker from the segment. In the analysis phase of our experiment, we investigated lexical similarities and correlation of psycholinguistic features between speakers for different emotional segments. We also used multilevel logistic regression method to understand the association between turn-taking features and emotional segments, and how the association differs from one emotion to another.

9.3 Data Preparation

For the analytical study, we selected a set of 523 conversations with the manual annotation of emotional states and overlap discourse. This set includes 310 conversations with emotional segments. Among the emotional segments around 11.28% of emotion are annotated as anger, 26.11% as frustration and 62.61% of annotated emotion in agent channel has empathy. From the rest of the 213 conversations, containing no emotional annotations, we selected segments and labeled them with no-emotion (NoEmo).


Data Preparation
Conversations: Emotional States
Conversations: Overlap Discourse
Turn-taking Information Extraction
Alignment: Turns and Emotional segments
Feature Extraction for Analysis
Analysis
- Lexical similarity
- Correlation analysis using psycholinguistic features
- Regression analysis using turn-taking features
Overlap Discourse Annotation
Audio Signal and Speaker Information
Turn Segmentation
ASR pipeline
Aligned Token files
Inter-Pausal Unit creation
Overlap segment boundaries
Steady Conversation Segment Creation
Overlap discourse tag
Turn Labeling

(a) Experimental pipeline of this study. (b) Turn-taking information extraction system.

Figure 9.1: System diagrams.

During the data preparation, we faced two important problems in order to define and align the emotional segment in association with turn-taking discourse: 1) emotional segment are very short in length, which made the task very difficult to get sufficient turn information, 2) an speaker respond to other speaker's emotion with a latency. To overcome these problems, we re-defined the following boundary of manual emotion segment with an impact window of length $2 * d$, where d is the length of the manual annotation of the emotional segment. Hence, the length of our emotional segment is $d + 2 * d = 3 * d$. We also investigated preceding context of each customer's emotional segment and defined it as $Pre.Emo$ with a window of length $3 * d$. The $NoEmo$ segments have been selected from conversations where no emotion in both agent and customer side has been annotated. From the middle of each conversation, we selected and extracted two $NoEmo$ segments with a length of the average emotional segment, ($\approx$ 42 sec). We extracted the $NoEmo$ segments from both agent and customer channels. As mentioned earlier, empathy, Emp, has been annotated in the agent channel only. Thus the preceding context of agent's emotional segment is defined as $Pre.Emp$. Hence, the investigated emotional and non-emotional segments include Pre.Emp, Emp, Ang, Fru, Pre.Ang, Pre.Fru and NoEmo.

9.4 Turn-taking Information Extraction

The Turn-Taking Information Extraction System, described in Figure 9.1 (b), consists of a *turn segmentation and labeling system*. The system uses lexical and manual overlap discourse annotation information to segment and labels the turn types. The pipeline uses the time aligned ASR output as tokens to create Inter-Pausal Units (IPUs) for each input channel. IPUs are defined as the consecutive tokens with no less that 50 ms gaps in between. Using the start and end time information of inter-IPUs and intra-IPUs, we created a steady time line and binary representation (presence or absence of speech information) segments for both the channels. We then defined these segments as *steady conversation segments*. The labels of each segment were then defined by a set of rules. Labels of the segments are as follows:

- Turn (T): Maximal sequences of IPUs where one single speaker has the floor, and none of the IPUs from the interlocutor are present [201]. T_A and T_C represent agent and customer's turns respectively.
- Pause (P): Gaps between the turns of the same speaker with no less than 0.5 sec. P_A and P_C represent agent and customer's pauses respectively.
- Overlap Types Ov= $\{Cmp, Ncm\}$: Overlapping turns between the two interlocutors with competitive or non-competitive intention (see section 3.1.2 for details).
- Lapse between speakers (L_B): Floor switches between the speakers with a silence duration of 2 sec or more.
- Lapse within speaker (L_W): Gaps between a speakers' turns with a silence duration of 2 sec or more.
- Switch (S): Floor switches between the speakers with silence less than 2 secs or with overlapping frames, not more than 20 ms.

9.4.1 Alignment: Turns and Emotional Segments

For the turn level analysis, it is important to align the turn sequences with the boundary of emotional segments. It is evident from manual annotation that an emotional segment consist of different turn types and not all the turns start inside the boundary. There are some cases, as shown in Figure 9.2 where the start/end of emotional episode can be at the middle of a turn. We solved this problem using a rule-based approach. For example, if half of a mismatched turn fall inside an emotional segment we considered that as a part of emotional segment.

9.5 Feature Extraction

9.5.1 Lexical Features

We extracted lexical features from automatic transcriptions from an in-house developed Automatic Speech Recognition (ASR) System [176]. The word error rate of the system is 31.78% on the test set. To understand the utility of the automated transcriptions with such as error rate, in a different study we compared the performance between automatic and manual transcriptions for a automatic classification of emotions. The results show that performance differences are

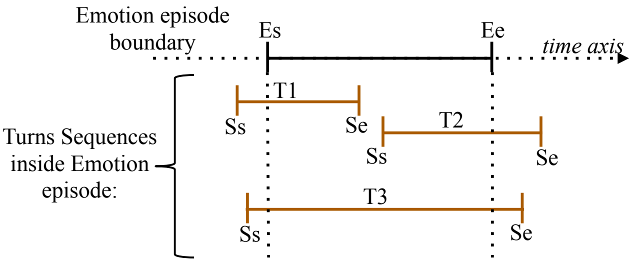


Figure 9.2: Type of mismatch between emotion segment boundaries and turns. E_s and E_e are the manual emotion segment boundaries. S_s and S_e are the turn boundaries. $T1$, $T2$, and $T3$ represents the type of turn boundary mismatch.

very low, only 1.2% drop with automated transcriptions [227]. Therefore, we found that the use of automatic transcriptions are reasonable for the experiment given that manual transcriptions are not available in call cases. For the experiments, the transcriptions of each segment were converted into bag-of-words vectors weighted with logarithmic term frequencies (tf) multiplied with inverse document frequencies (idf). We also reduced the size of the dictionary by removing stop-words and lower frequent words.

9.5.2 Psycholinguistic Features

Psycholinguistic features were extracted from the transcriptions, using Linguistic Inquiry Word Count (LIWC) [228]. It has been used to study personality, the role of speakers in overlaps [177, 229] among other social behaviors in order to understand the correlation between these attributes and word uses. The feature category includes linguistic (e.g., preposition, verb, word count), psychological (affect, positive, negative emotion, anxiety), personal concern (e.g., work, home, money), swear words, relativity among others. The LIWC is a knowledge-based system, which was designed using a set of dictionaries for different languages including Italian. In the dictionary, each word was labeled with feature categories mentioned above. During the feature extraction process the word in the transcriptions was matched with the dictionary. Then, the matched category was computed as frequency or relative frequency. The Italian version of the dictionary contains 85 word categories [230]. We also extracted 5 general and 12 punctuation categories constituting a total of 102 features. We then removed LIWC features that are not observed in our training dataset.

9.5.3 Turn-Taking Features

The turn-taking features were generated using the turn sequence output of the Turn-Taking Information Extraction System, described in Section 9.4. The sequences were first aligned with each corresponding emotional segment (see Section 9.4.1). To understand the impact of the choice of turn-taking behavior, we divided the feature sets, at both segment and individual speaker levels, into two groups. A brief description of extracted features, in the segment, are as follows:

- General information about emotional segment (G1):
  - Participation equality, shown in Equation 9.1

$$P_{eq} = 1 - (\frac{\sum_i^N (T_i - T)^2/T}{E}) \tag{9.1}$$

    where T is the average speech duration of the speakers. T_i is the total speech duration for each speaker. E represents the total speech duration. $N = 2$, represents two speakers as agent and customer inside the emotional segment.
  - Percentage of overlaps.
  - Percentage of Cmp and Ncm on total overlap duration.
- Length of different turn types (G2):
  - Median duration of T_A, T_C, P_A, P_C, Cmp, Ncm, L_W and L_B, inside emotional segment normalized by the median of speaker's respective turn in the whole conversation.

9.6 Analysis and Results

For different feature sets, we investigated different experimental configurations. For the study of lexical similarities, our experimental conditions include: 1) lexical features from paired (i.e., agent and customer channel from same conversation) speakers' non-overlapping *vs* overlapping turns, 2) lexical features from non-paired (i.e., agent and customer channel extracted from unrelated conversation) speakers' non-overlapping *vs* overlapping turns. Where as for psycholinguistic features, we investigated features obtained from non-overlapping *vs* overlapping turns. For turn-taking features, we have not made any such distinctions. The non-overlapping turns include all the turns of the speakers excluding the overlaps. Where as the overlapping turns includes competitive (Cmp) and non-competitive (Ncm) overlaps.

9.6.1 Lexical Similarities

For the analysis, we computed cosine similarity of the agent and customer aligned segment representing different emotional states. For the lexical similarity we designed feature vector for agent $\overrightarrow{V_{S_A}}$ and customer $\overrightarrow{V_{S_C}}$ emotional segment using bag-of-word model and transformed

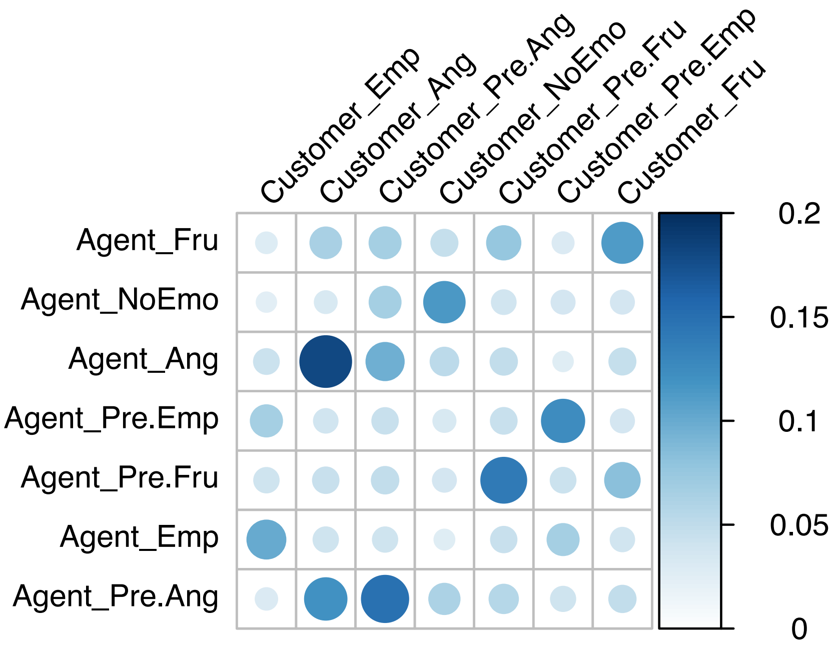


Figure 9.3: Lexical similarity between the emotional segment of the agent and the customer channel. Pre. represents preceding segments. Ang - anger, Fru - frustration, Emp - empathy, NoEmo - no-emotion

them into tf-idf. Then, we computed cosine similarity, as shown in Equation 9.2 between the feature vector of the agent and customer's segment. For a pair-wise comparison of emotional states, then, we computed mean and standard deviation with statistical significance using t-test.

$$sim(S_A, S_C) = \frac{\overrightarrow{V_{S_A}} \cdot \overrightarrow{V_{S_C}}}{\left|\overrightarrow{V_{S_A}}\right| \cdot \left|\overrightarrow{V_{S_C}}\right|} \tag{9.2}$$

As mentioned earlier, we have four different experimental configurations for the analysis of lexical similarities. As a baseline, we computed the similarities between non-paired speakers using the lexical features from non-overlapping turns for different emotional segments. The results are presented in a form of similarity map in Figure 9.3. From the results, we observed that the interlocutors entrain each other in non-overlapping turns when the customer is expressing anger, and the value of similarity ($sim = 0.181$) is significantly ($p < 0.05$) higher than the similarities in any other emotional segment.

In the experiment with competitive overlapping turns, we observed the highest similarity of 0.035 and 0.031 in preceding-anger and anger segments, respectively. In the case of non-competitive overlapping turns, a similarity of 0.034 was observed between the interlocutors in frustration segments. The results on overlapping turns are insignificant.

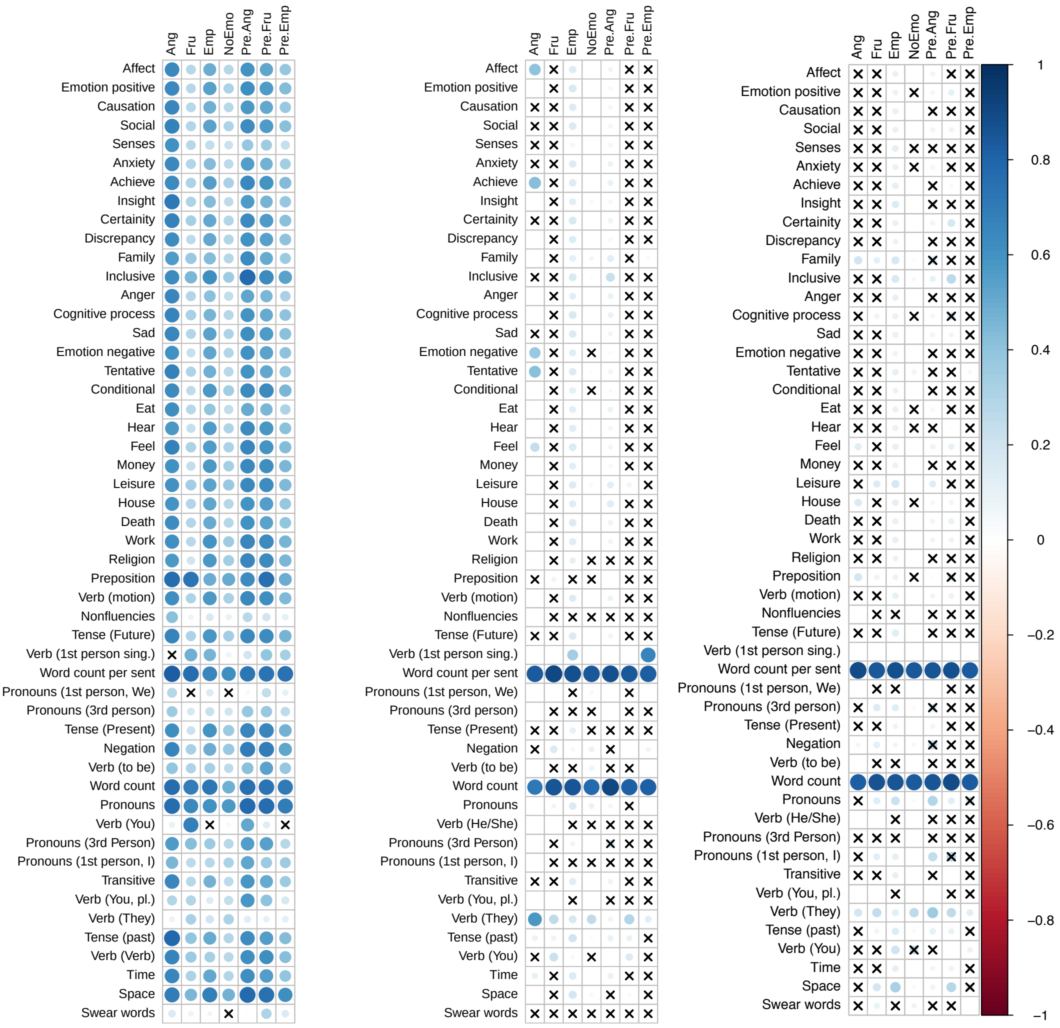
9.6.2 Psycholinguistic Features

We explored the degree of coordination using Pearson correlation coefficient (r) between the interlocutors' behaviors by correlating psycholinguistic features obtained from overlapping and non-overlapping turns, presented in Figure 9.4. For the sake of simplicity, the magnitude of r values are presented using colors where as '$\times$' symbol represent the corresponding r is not significant. These analyses are based on entire emotion segments from the agent and customer channels, irrespective of turns. The r is calculated for each psycholinguistic feature by correlating the agent and customer feature vectors of the conversations. We calculated the significance of the correlation coefficient r using t-test with a degree of freedom equal to $n - 2$, where n represent the total number of instances.

From the correlation plot, it is apparent that the non-overlapping turns of the interlocutors in anger (Ang) segments has high correlation values compared to other emotional segments non-overlapping turns and also compared to overlapping turns (Ncm and Cmp). Not surprisingly the magnitude of the correlation is significantly higher for psychological features like anxiety, affect, and sad between anger segments compared to frustration and empathy segments. Looking at the preceding-anger segments, we observed that the magnitude of r for personal concern along with psychological features are also stronger. It indicates that the cues of anger segment can be found in its preceding segments. The results also show that the uses of pronouns or negation words is directly proportional to the another speaker's usage. We also observed similar patterns in the uses of tenses. The magnitude of r is much higher for past-tense uses in anger compared to others emotional segment and preceding emotional context.

In the case of frustration, the strength of r decrease compared to the preceding segment of frustration. Unlike preceding-frustration segment, we observed that in frustration, there is less coordination between the interlocutors with an exception in preposition and word count features. Though a slight increase in r is observed in verb (they) feature. It is also observed that the interlocutors seem to be more coordinated in the use of swear words in preceding-frustration segments compared to all other segments.

In empathy segments, the coordination of the agent and customer improves compared to preceding-empathy, frustration, and no-emotion segments but the magnitude of coordination is not as impressive as anger segments.



(a) Non-overlapping segments(b) Ncm overlapping turns (c) Cmp overlapping turns

Figure 9.4: Correlation analysis at the non-overlapping segment, and overlapping segments, where '✕' symbol represents that the corresponding r is not significant. Pre. represents preceding segments. Ang - anger, Fru - frustration, Emp - empathy, NoEmo - no-emotion

In competitive and non-competitive overlapping turns, a very few significant coordination has been observed. The experiment with non-competitive turns shows that the interlocutors coordinate in anger segment with the features such as affect, achieve, negative emotion, tentative, and verb (they). In the case of competitive overlaps, we observed weak positive correlations between the interlocutors in preceding-frustration segment with feature inclusive, preceding-anger segment with a verb (they), and in empathy segment with space feature.

9.6.3 Turn-Taking Features

For the experiment with turn-taking features, we applied a multilevel logistic regression to understand the association of turn-taking features with emotional expressions and how they differ from one emotional state to another. The association of turn-taking features with emotional segments are presented in Table 9.1, in terms of regression coefficients. In Table 9.1 (a), the coefficients are reported with respect to the preceding segment of each emotion, where as in Table 9.1 (b), the coefficients represents the association of each turn-taking feature with the preceding emotion segment *vs.* no-emotion segments.

The results indicates that compared to the preceding context of empathy (Pre.Emp) and no emotion (NoEmo) segments, $participationEquality$, $MedianTurnC$ and $MedianPauseC$ has a negative effect on empathy (Emp) segment, where as $\%Overlap$ and length of non-competitive overlap ($MedianNcm$) has a significant positive effect. Thus indicating the importance of non-competitive overlap in the empathic segment (Emp). The results also hypothesize that during this emotional episode, agents tends to talk more allowing less participation equality between the agent and the customer. The duration of customer's turn and pause tends to be small.

The features $\%Overlap$, the length of overlaps ($MedianNcm$ and $MedianCmp$) has a positive effect for anger segment compared to no-emotion segment. We also observed similar findings for $MedianCmp$ for preceding-anger w.r.t to no-emotion segment. It is observed that the length of non-competitive overlaps ($MedianNcm$) has a positive association where as the length of the lapse between the speakers ($MedianLb$) has a negative effect on anger with respect to preceding context (Pre.Ang). From the result of comparing preceding-anger w.r.t to no-emotion segments, we noticed that the positive association of the length of competitive overlap is present from the preceding context as an indication of anger.

The features $\%Overlap$, the length of overlaps ($MedianNcm$ and $MedianCmp$) has a positive effect for anger segment compared to the no-emotion segment. We also observed similar findings for $MedianCmp$ for preceding-anger w.r.t to the no-emotion segment. It is observed that the length of non-competitive overlaps ($MedianNcm$) has a positive association where as the length of the lapse between the speakers ($MedianLb$) has a negative effect on anger with respect to preceding context (Pre.Ang). From the result of comparing preceding-anger w.r.t to no-emotion segments, we noticed that the positive association of the length of competitive overlap is present from the preceding context as an indication of anger.

Apart from the results presented in Table 9.1, we also compared the association of turn-taking features with empathy, anger, and frustration with respect to each other. We found no significant difference between anger and frustration segments. However, the preceding context of anger and frustration shows that compared to the preceding-frustration, decrease of pause length is positively associated with preceding-anger segment, especially in agent's side. It is observed that an increase in the length of competitive overlap duration, $MedianCmp$, is positively associated with anger segments w.r.t empathy segments.

Table 9.1: Regression coefficient w.r.t preceding segment of each emotion and no-emotion segments.

Groups	Features	(a) Compared to preceding segment			(b) Compared to noemo segment					
		Emp	Ang	Fru	Emp	Ang	Fru	Pre.Emp	Pre.Ang	Pre.Fru
G1	participationEquality	**-0.948**	0.259	**-1.381**	-0.244	1.018	0.253	**0.823**	0.060	**1.669**
	% Overlap	**0.069**	0.059	**0.131**	**0.099**	**0.112**	**0.083**	0.035	0.046	-0.046
	% Cmp	0.002	0.008	-0.005	0.005	0.015	**0.011**	0.001	0.009	**0.015**
	% Ncm	0.007	0.000	-0.009	**0.010**	-0.002	0.000	0.002	0.000	**0.008**
G2	MedianTurnA	0.001	-0.002	-0.001	0.000	-0.001	-0.001	0.000	0.000	-0.001
	MedianTurnC	**-0.003**	0.001	**0.003**	-0.001	0.002	**0.003**	**0.002**	0.001	0.000
	MedianPauseA	0.001	0.001	-0.004	0.002	-0.005	-0.002	0.001	-0.004	**0.005**
	MedianPauseC	**-0.005**	-0.008	**-0.008**	**-0.003**	0.002	-0.002	0.002	0.003	**0.006**
	MedianCmp	0.001	0.002	0.001	**0.003**	**0.006**	**0.005**	0.002	**0.004**	**0.005**
	MedianNcm	**0.004**	**0.007**	0.002	**0.003**	**0.003**	0.002	0.001	0.001	0.000
	MedianLb	-0.002	**-0.011**	**-0.006**	**-0.005**	-0.004	-0.003	**-0.002**	0.000	0.000
	MedianLw	0.000	-0.002	-0.001	**-0.002**	-0.003	-0.001	-0.001	0.000	0.000

We also compared the duration of competitive and non-competitive overlap within different emotions and preceding emotional segments. In case of competitive, as shown in Figure ??, we observed that duration of mean competitive overlap in anger ($1.25s$) and frustration ($1.09s$) are significantly more compared to the empathy ($0.93s$), no-emotion ($0.80s$) while there is not significant difference between the duration of competitive in anger and frustration segment. In the case of preceding emotion segments, the duration of competitive overlap in frustration is significantly higher than that of preceding-frustration ($0.91s$), where as preceding-anger ($1.12s$) and preceding-frustration is significantly higher than no-emotion. It is also observed that competitive duration in empathy segment is also longer ($p < 0.05$) than no-emotion segments. As for non-competitive duration, shown in Figure ??, there is no significant difference between anger ($0.72s$), frustration ($0.68s$) and empathy ($0.69s$) segment. But it is observed that empathy has significantly longer non-competitive overlap compared to no-emotion ($0.53s$) and preceding-empathy ($0.61s$) segment. Even, the preceding context of empathy (Pre.Emp) has significantly longer non-competitive overlap duration than the non-competitive overlap where there is no emotion. While in anger and frustration, the non-competitive overlap length is significantly higher than the no-emotion segment.

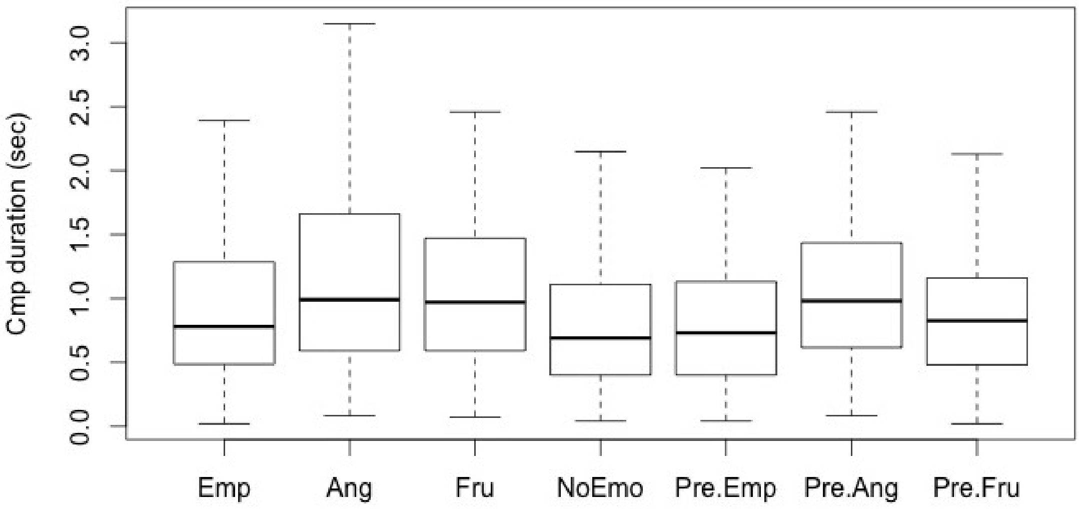


Figure 9.5: Duration distribution of competitive overlaps in different emotional segments.

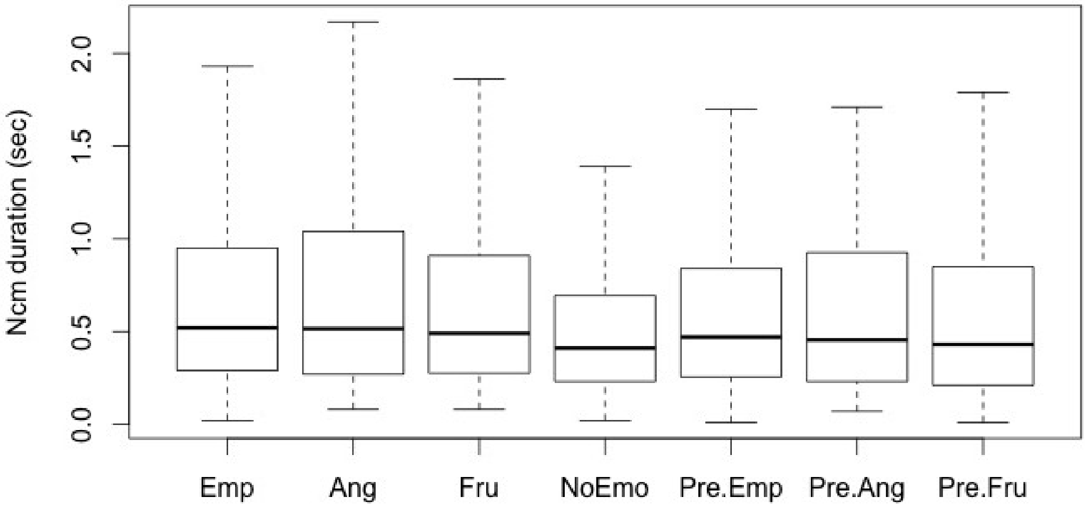


Figure 9.6: Duration distribution of non-competitive overlaps in different emotional segments.

Table 9.2: An automatic generated excerpt of a conversation between agent and customer, inside an anger segment. A: Agent, C: Customer. Each row represents a turn in the conversation. $[X]_{Cmp}$ presents competitive overlaps and $[X]_{Ncm}$ presents non-competitive overlaps.

A:	[negare quello che lei ci]$_{Cmp}$ *[deny what you us]*
C:	[no no no non so non]$_{Cmp}$ *[no no no I do not know not]*
C:	andate in diffida perché *gone on a warning because*
C:	[io ho chiamato ho mandato anche una raccomandata]$_{Cmp}$ [anche una raccomandata mi scusi mi faccia]$_{Cmp}$ *[I have called (I) have send also a registered (mail)] [also a registered (mail) excuse me let me]*
A:	[una signora io il sistema davanti quindi]$_{Cmp}$ *[one, Madame, I (am in front of) the system, then]*
C:	[parlare]$_{Ncm}$ *[speak]*
A:	[eh]$_{Ncm}$ *[huh]*
A:	eh però sul se scadesse al *huh however on if expires at*
A:	[sistema risultano in diffida]$_{Cmp}$ *[(the) system (they)appear (to be on) warning]*
C:	[sì un computer sono andate in]$_{Cmp}$ *[yes a computer (they) are gone in]*
C:	diffida quelle lì non io non le ho assolutamente pagate perché il la lettura del *warning those ones (I) do not have paid them absolutely because the reading of the*
C:	[del del]$_{Ncm}$ *[of the of the]*
A:	[eh]$_{Ncm}$ *[huh]*
C:	contatore *meter*
C:	[è stato fermo per sì]$_{Ncm}$ [*(it) was stopped for yes]*
A:	[certo lei]$_{Ncm}$ *[of course you]*
A:	non le ha pagate quindi sono andate in diffida quindi c è stato un interesse di mora *did not payed them so they had gone on warning so there has been a default interest ...*

9.7 Summary

In this study, we explored the coordination of interlocutors in different emotional segments using lexical, psycholinguistic and turn-taking features. We investigated such feature sets in terms of regression coefficients, cosine similarity and correlation analysis, respectively. We observed that the interlocutors match each other turns, in terms of lexical similarity and psycholinguistic features, significantly more in anger segment compared to other emotional segments. We also observed that in preceding segment of anger the speakers shows significant correlation with each other in terms of psycholinguistic features. In terms of turn-taking features, no significant differences between anger and frustration have been noticed, apart from the difference in length of pauses in the preceding segment of the emotion. It indicates that preceding context of anger has shorter pause with respect to frustration. Unlike anger, we found less coordination in the segment where the agent is empathic even though an increase in the percentage of non-competitive overlaps has been observed. This is our preliminary study towards utilizing these feature sets for the classification of emotional states and turn-taking discourse, which we will investigate in future.

Chapter 10

Conclusion and Future Works

10.1 Contributions and Discussion

The motivation of this dissertation is to design computational approaches for modeling turn-taking dynamics. To model turn-taking dynamics, this thesis focused on two aspects of conversational dynamics: 1) design *automated computational models for analyzing turn-taking behavior* in a dyadic conversation, and 2) predict the outcome of the conversations, i.e., *observed user satisfaction*, using turn-taking descriptors and understanding the coordination between the interlocutors inside emotional episodes.

Towards achieving this goal, the dissertation first studied what are the available research in turn-takings. It is observed that even though overlapping speech and long silences are common in every day conversation and carries behavioral information, however, a very little work has been done to model this behavior and their functions towards the conversational flow. Therefore, to model the turn-taking dynamics, we found that our primary priority is to push the boundary of the research on overlapping speech and silence.

Towards the goal of modeling discourse of overlapping speech, we needed an operational model to classify the competitiveness in overlapping speech. However, to model such discourses, we needed a novel overlap annotation guideline. Therefore, we designed an annotation guideline for segmenting and annotating the speech overlaps with the competitive and non-competitive discouses, which is one of the main contributions of this thesis.

A significant effort has been given on designing the discourse model of overlaps. The research first focused on the low-level acoustic features, such as spectral, mfcc, and prosodic feature, to evaluate the distinguishing capabilities of the features while categorizing competitiveness in overlaps. This investigation has been done by incorporating both the interlocutors' channel information. It is observed that spectral features along with prosodic features provide sufficient power to the model for classifying overlaps in absence of any other information. Later, the thesis also studied the linguistic features such as psycholinguistic and lexical features. In addition, this study presents how most relevant lexical ngrams act as a window for describing the overlap discourse.

To understand the role of speakers and the context, the study also focused on designing classifier using contextual information such as overlapper, overlappee, left, right, and their different combinations. Examining the results, it is observed that lexical choice of the overlapper is a good indicator of competitiveness in overlaps, where as for other information such as acoustic, the information regarding the discourse of overlaps can be found in both the interlocutors' segment along with the context.

Apart from investigating individual feature set, this thesis also focused on different combination techniques such as decision level or feature level combination. The purpose is to study is to develop an architecture to combine different information in classifying overlaps. It is observed, that the feature level combination of lexical information (Bag-of-ngrams or word embedding features) along with acoustic information outperforms any individual feature models and their decision level combination. In addition to explore different features and their combination, the thesis also focused on exploiting the power of linear (SVM) and non-linear (DNN) algorithms to classify the overlap discourse. We have observed that the unbalanced natural distribution presents a challenge for the performance of the discourse model especially for competitive class.

The dissertation also attempts to shed some lights on the functional aspects of long silence in between- and within- speaker turns. This is one of the most challenging tasks because as most of the literature on silence points out the silence is valueless, but the function of the silence is based on the context and the situation. But there is no operational categories or properties to define the functions. Therefore, to model the silence function, we first need to code silence in feature space. We designed the silence feature space using the preceding and succeeding turns of silences. To code the action in the surrounding turns, we used dimension and communicative functions of dialog acts. Following the design of the features, we categorized the functions of between- and within- speaker silences using a hierarchical concept learning technique, and defined general functional categories by selecting and merging the clusters (i.e., sub-trees), from a hierarchical tree, based on their functional similarity. We observed that there are different functions of these long silences, varying from response preparation to hesitation about some queries. Even though there can be other cognitive functions, however, they are out of scope of this dissertation. This study is our first attempt to understand long silence, and there is still more research needed to be done.

To model the turn-taking dynamics, we designed a framework that automatically segments the turns and turn-taking events, such as silence, and categorizes its discourse labels. Such a system can take the audio signal as an input to the system and uses state-of-the-art ASR pipeline with other discourse module researched in this dissertation (including dialog act segmenter and classifier, and overlap discourse model) to label the turn-taking behavior in the conversation.

To understand the role of the coded turn-taking behavior in predicting the outcome of the conversation, the dissertation also focused to design a computational model to automatically predict observed user satisfaction, as a measure of the conversational outcome. The dissertation defines the observed user satisfaction, as the final emotional manifestation of the conversation, which can be either positive, negative or neutral. Moreover, to predict user satisfaction, the dissertation also focused on engineering the turn-taking behavioral features. Our experimental finding suggests that turn-taking features are a powerful tool to predict the phenomena. A detailed analysis of turn-taking features suggests that the use of frequent non-competitive turns and social dialog can be a key to having a satisfied conversation.

To study the association of turn-taking dynamics with different emotional segments, the dissertation also focused on studying how the turn-taking behavior along with other features changes with different basic and complex emotional segments such as anger, frustration, empathy. The analysis of the study showed that in the preceding segments before anger, the pause length tends to be shorter with respect to frustration. The study also observed that interlocutors coordinate more with each other in anger segments compared to other emotional segments.

According to all the observations and experimental results found in this dissertation, turn-taking contains many behavioral information of the spoken conversation and can be used to predict human long and short time behavioral characteristics such as personality, dominance or even atypical conditions such as "flight of ideas". The studies presented in the dissertation opens many new avenues of research which can incorporate the turn-taking dynamics.

10.2 Possible Extensions

Throughout this dissertation, we designed, implemented, and evaluated computational models for overlap discourse classification, functions of silence, to segment and label turn-taking dynamics and to predict the outcome of the conversation. While this dissertation tried to address most of the possible queries regarding the models, but there are still scope for further extension of the research.

In regards to the study of overlapping speech discourse, one important study that can be explored in future is to investigate an adaptation or transfer learning approach to exploit unlabeled data. It is also necessary to understand how the computational model works across corpora in order to understand the generability of the model and the designed architecture. A small part of this thesis also focused on segmenting and classifying overlaps in a mono channel scenarios. To improve the performance, the mono-channel models needs more research attentions from the speech community.

The contribution in this thesis for modeling the functions of silence is the first step towards this research areas. Even though for decades many studies have tried to conceptualize the function of silence, however, there is a very little contribution from speech communities. To understand more about functions of silence, we need to combine the observational cognitive aspects along with the experimental techniques. This is a vast unexplored area which needs a lot of attention.

To improve the quality of artificial agents or enhance the research on behavioral studies of human interaction, we need to learn how to utilize the designed turn-taking dynamics models with expressive affective behavior models and design a combined system.

Bibliography

[1] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. R. Traum, “ISO 24617-2: A semantically-based standard for dialogue annotation.” in *LREC*, 2012, pp. 430–437.

[2] E. Stepanov, B. Favre, F. Alam, S. Chowdhury, K. Singla, J. Trione, F. Béchet, and G. Riccardi, “Automatic summarization of call-center conversations,” in *In Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[3] S. Narayanan and P. G. Georgiou, “Behavioral signal processing: Deriving human behavioral informatics from speech and language,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[4] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, pp. 696–735, 1974.

[5] M. Heldner and J. Edlund, “Pauses, gaps and overlaps in conversations,” *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[6] L. Ten Bosch, N. Oostdijk, and L. Boves, “On temporal aspects of turn taking in conversational dialogues,” *Speech Communication*, vol. 47, no. 1, pp. 80–86, 2005.

[7] C. West, “Against our will: Male interruptions of females in cross-sex conversation*,” *Annals of the New York Academy of Sciences*, vol. 327, no. 1, pp. 81–96, 1979.

[8] J. A. Goldberg, “Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts,” *Journal of Pragmatics*, vol. 14, no. 6, pp. 883–903, 1990.

[9] S. C. Levinson, “Pragmatics (cambridge textbooks in linguistics),” 1983.

[10] T. J. Bruneau, “Communicative silences: Forms and functions,” *Journal of Communication*, vol. 23, no. 1, pp. 17–46, 1973.

[11] E. Zerubavel, *The elephant in the room: Silence and denial in everyday life.* Oxford University Press, 2006.

[12] M. Saville-Troike, “The place of silence in an integrated theory of communication,” *Perspectives on silence*, pp. 3–18, 1985.

[13] I. Nakane, *Silence in intercultural communication: Perceptions and performance*. John Benjamins Publishing, 2007, vol. 166.

[14] K. Agyekum, “The communicative role of silence in akan,” *Pragmatics*, vol. 12, no. 1, pp. 31–32, 2002.

[15] J. R. Hackman and N. Vidmar, “Effects of size and task type on group performance and member reactions,” *Sociometry*, pp. 37–54, 1970.

[16] E. Shriberg, E. Wade, and P. Price, “Human-machine problem solving using spoken language systems (sls): Factors affecting performance and user satisfaction,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 49–54.

[17] K. Forbes-Riley and D. J. Litman, “Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters,” in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 264–271.

[18] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, “Paradise: A framework for evaluating spoken dialogue agents,” in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 271–280.

[19] K.-P. Engelbrech, F. Gödde, F. Hartard, H. Ketabdar, and S. Möller, “Modeling user satisfaction with hidden markov model,” in *Proceedings of the SIGDIAL 2009 conference: the 10th annual meeting of the special interest group on discourse and dialogue*. Association for Computational Linguistics, 2009, pp. 170–177.

[20] F. Alam, S. A. Chowdhury, M. Danieli, and G. Riccardi, “How interlocutors coordinate with each other within emotional segments?” in *COLING: International Conference on Computational Linguistics*, 2016.

[21] M. A. Schmuckler, “What is ecological validity? a dimensional analysis,” *Infancy*, vol. 2, no. 4, pp. 419–436, 2001.

[22] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 2007.

[23] S. Al Moubayed, J. Beskow, B. Bollepalli, A. Hussen-Abdelaziz, M. Johansson, M. Koutsombogera, J. D. Lopes, J. Novikova, C. Oertel, G. Skantze *et al.*, "Tutoring robots," in *International Summer Workshop on Multimodal Interfaces*. Springer, 2013, pp. 80–113.

[24] G. Skantze, M. Johansson, and J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 67–74.

[25] S. Bernardini, K. Porayska-Pomsta, and T. J. Smith, "Echoes: An intelligent serious game for fostering social communication in children with autism," *Information Sciences*, vol. 264, pp. 41–60, 2014.

[26] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[27] D. Reidsma, I. de Kok, D. Neiberg, S. C. Pammi, B. van Straalen, K. Truong, and H. van Welbergen, "Continuous interaction with a virtual human," *Journal on Multimodal User Interfaces*, vol. 4, no. 2, pp. 97–118, 2011.

[28] W. Fisher, R. Groff, and H. Roane, "Applied behavior analysis: History, philosophy, principles, and basic methods," *Handbook of applied behavior analysis*, pp. 3–13, 2011.

[29] A. Norwine and O. Murphy, "Characteristic time intervals in telephonic conversation," *Bell System Technical Journal*, vol. 17, no. 2, pp. 281–291, 1938.

[30] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell System Technical Journal*, vol. 47, no. 1, pp. 73–91, 1968.

[31] J. Jaffe and S. Feldstein, *Rhythms of dialogue*. Academic Press, 1970, vol. 8.

[32] A. J. Sellen, "Remote conversations: The effects of mediating talk with technology," *Human-computer interaction*, vol. 10, no. 4, pp. 401–444, 1995.

[33] J.-P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[34] M. B. Walker and C. Trimboli, "Smooth transitions in conversational interactions," *The Journal of Social Psychology*, vol. 117, no. 2, pp. 305–306, 1982.

[35] S. Duncan, "Some signals and rules for taking speaking turns in conversations." *Journal of personality and social psychology*, vol. 23, no. 2, p. 283, 1972.

[36] E. Campione and J. Véronis, “A large-scale multilingual study of silent pause duration,” in *Speech prosody 2002, international conference*, 2002.

[37] L. F. Barrett, M. Lewis, and J. M. Haviland-Jones, *Handbook of emotions*. Guilford Publications, 2016.

[38] W. J. Levelt, *Speaking: From intention to articulation*. MIT press, 1993, vol. 1.

[39] C. E. Ford and S. A. Thompson, “Interaction and grammar, chapter interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns, pages 134–184,” 1996.

[40] H. Furo, *Turn-taking in English and Japanese: projectability in grammar, intonation and semantics*. Routledge, 2013.

[41] B. Oreström, *Turn-taking in English conversation*. Krieger Pub Co, 1983, vol. 66.

[42] W. L. Chafe, “Prosodic and functional units of,” *Talking data: Transcription and coding in discourse research*, vol. 33, 1993.

[43] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, “An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs,” *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[44] S. Nakajima and J. F. Allen, “A study on prosody and discourse structure in cooperative dialogues,” *Phonetica*, vol. 50, no. 3, pp. 197–210, 1993.

[45] E. A. Schegloff, “Reflections on studying prosody in talk-in-interaction,” *Language and speech*, vol. 41, no. 3-4, pp. 235–263, 1998.

[46] B. Wells and S. Macfarlane, “Prosody as an interactional resource: Turn-projection and overlap,” *Language and Speech*, vol. 41, no. 3-4, pp. 265–294, 1998.

[47] W. Wesseling, R. J. v. Son *et al.*, “Timing of experimentally elicited minimal responses as quantitative evidence for the use of intonation in projecting trps,” in *Interspeech*, no. 6, 2005, pp. 3389–3392.

[48] C. Goodwin, *Conversational organization: Interaction between speakers and hearers*. Academic Press, 1981.

[49] D. Gatica-Perez, “Automatic nonverbal analysis of social interaction in small groups: A review,” *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[50] E. A. Schegloff, “Discourse as an interactional achievement: Some uses of ‘uh huh’and other things that come between sentences,” *Analyzing discourse: Text and talk*, vol. 71, p. 93, 1982.

[51] V. H. Yngve, “On getting a word in edgewise,” in *Chicago Linguistics Society, 6th Meeting*, 1970, pp. 567–578.

[52] R. Gardner, *When listeners talk: Response tokens and listener stance*. John Benjamins Publishing, 2001, vol. 92.

[53] T. Stivers, “Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation,” *Research on language and social interaction*, vol. 41, no. 1, pp. 31–57, 2008.

[54] G. H. Lerner, “Collaborative turn sequences,” *Pragmatics and beyond new series*, vol. 125, pp. 225–256, 2004.

[55] ——, “Turn-sharing,” *The language of turn and sequence*, pp. 225–256, 2002.

[56] P. French and J. Local, “Turn-competitive incomings,” *Journal of Pragmatics*, vol. 7, no. 1, pp. 17–38, 1983.

[57] D. H. Zimmermann and C. West, “Sex roles, interruptions and silences in conversation,” *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pp. 211–236, 1996.

[58] C. D. West and D. H. Zimmerman, “Women’s place in everyday talk: Reflections on parent-child interaction,” *Social problems*, vol. 24, no. 5, pp. 521–529, 1977.

[59] C. West and D. H. Zimmerman, “Small insults: A study of interruptions in cross-sex conversations between unacquainted persons,” *Language, gender and society*, pp. 102–117, 1983.

[60] A. Esposito, “Sex differences in children’s conversation,” *Language and Speech*, vol. 22, no. 3, pp. 213–220, 1979.

[61] G. W. Beattie, “Interruption in conversational interaction, and its relation to the sex and status of the interactants,” *Linguistics*, vol. 19, no. 1-2, pp. 15–36, 1981.

[62] E. A. Schegloff, “Overlapping talk and the organization of turn-taking for conversation,” *Language in society*, vol. 29, no. 01, pp. 1–63, 2000.

[63] S. O. Murray, “Toward a model of members’ methods for recognizing interruptions,” *Language in Society*, vol. 14, no. 01, pp. 31–40, 1985.

[64] N. Ferguson, “Simultaneous speech, interruptions and dominance,” *British Journal of Clinical Psychology*, vol. 16, no. 4, pp. 295–302, 1977.

[65] D. James and C. Sandra, “Women, men, and interruptions: a critical review. gender and conversational interaction, ed. by deborah tannen, 231–280,” 1993.

[66] S. Feldstein and J. Welkowitz, “A chronography of conversation: In defense of an objective approach,” *Nonverbal behavior and communication*, pp. 329–378, 1978.

[67] G. Jefferson, “A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences,” *Semiotica*, vol. 9, no. 1, pp. 47–96, 1973.

[68] G. W. Beattie, “Turn-taking and interruption in political interviews: Margaret thatcher and jim callaghan compared and contrasted,” *Semiotica*, vol. 39, no. 1-2, pp. 93–114, 1982.

[69] D. B. Roger and A. Schumacher, “Effects of individual differences on dyadic conversational strategies.” *Journal of Personality and Social Psychology*, vol. 45, no. 3, p. 700, 1983.

[70] G. Jefferson, “Notes on some orderlinesses of overlap onset,” *Discourse analysis and natural rhetoric*, vol. 500, pp. 11–38, 1984.

[71] ——, “A sketch of some orderly aspects of overlap in natural conversation,” *PRAGMATICS AND BEYOND NEW SERIES*, vol. 125, pp. 43–62, 2004.

[72] E. Couper-Kuhlen, *English speech rhythm: Form and function in everyday verbal interaction*. John Benjamins Publishing, 1993, vol. 25.

[73] C.-C. Lee, S. Lee, and S. S. Narayanan, “An analysis of multimodal cues of interruption in dyadic spoken interactions.” in *Proc. of INTERSPEECH*, 2008, pp. 1678–1681.

[74] E. Kurtic, G. J. Brown, B. Wells, D. Barth-Weingarten, D. Dehé, and A. Wichmann, “Fundamental frequency height as a resource for the management of overlap in talk-in-interaction,” *Where Prosody Meets Pragmatics,. In: Studies in Pragmatics*, vol. 8, pp. 183–205, 2009.

[75] M. Danieli, C. Bazzanella, L. SpA, and I. Torino, “Linguistic markers in coming to understanding,” in *Proceedings of VIII Meeting of AIIA (Associazione Italiana Intelligenza artificiale), AIIA 2002*, 2002, pp. 10–13.

[76] C. Bazzanella, *Repetition in dialogue*. Walter de Gruyter, 1996, vol. 11.

[77] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, “Perceptual and acoustic correlates of abnormal voice qualities,” *Acta oto-laryngologica*, vol. 90, no. 1-6, pp. 441–451, 1980.

[78] E. Kurtic, G. J. Brown, and B. Wells, “Resources for turn competition in overlap in multiparty conversations: speech rate, pausing and duration.” in *Proc. of INTERSPEECH*, 2010, pp. 2550–2553.

[79] E. Kurtić, G. J. Brown, and B. Wells, “Resources for turn competition in overlapping talk,” *Speech Communication*, vol. 55, no. 5, pp. 721–743, 2013.

[80] C. Oertel, M. Wlodarczak, A. Tarasov, N. Campbell, and P. Wagner, “Context cues for classification of competitive and collaborative overlaps,” *Proceedings of Speech Prosody 2012*, 2012.

[81] K. P. Truong, “Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee,” in *Proc. of INTERSPEECH*, 2013, pp. 1404–1408.

[82] C.-C. Lee and S. Narayanan, “Predicting interruptions in dyadic spoken interactions,” in *Proc. of ICASSP*. IEEE, 2010, pp. 5250–5253.

[83] A. Gravano and J. Hirschberg, “A corpus-based study of interruptions in spoken dialogue.” in *Proc. of INTERSPEECH*, 2012.

[84] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, “The meeting project at icsi,” in *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics, 2001, pp. 1–7.

[85] E. Shriberg, A. Stolcke, and D. Baron, “Observations on overlap: findings and implications for automatic processing of multi-party conversation.” in *INTERSPEECH*, 2001, pp. 1359–1362.

[86] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “Switchboard: Telephone speech corpus for research and development,” in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[87] J. N. Cappella, “Talk and silence sequences in informal conversations ii,” *Human Communication Research*, vol. 6, no. 2, pp. 130–145, 1980.

[88] M. L. McLaughlin and M. J. Cody, “Awkward silences: Behavioral antecedents and consequences of the conversational lapse,” *Human communication research*, vol. 8, no. 4, pp. 299–316, 1982.

[89] T. J. Bruneau, “How americans use silence and silences to communicate.” *China Media Research*, vol. 4, no. 2, 2008.

[90] M. Kogure, “Nodding and smiling in silence during the loop sequence of backchannels in japanese conversation,” *Journal of Pragmatics*, vol. 39, no. 7, pp. 1275–1289, 2007.

[91] J. C. Damron, “Attitudes toward interpersonal silence within dyadic relationships.” Ph.D. dissertation, 2009.

[92] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[93] J. Oduro-Frimpong, “Semiotic silence: Its use as a conflictmanagement strategy in intimate relationships,” *Semiotica*, vol. 2007, no. 167, pp. 283–308, 2007.

[94] Z. Frankel, H. M. Levitt, D. M. Murray, L. S. Greenberg, and L. Angus, “Assessing silent processes in psychotherapy: An empirically derived categorization system and sampling strategy,” *Psychotherapy Research*, vol. 16, no. 5, pp. 627–638, 2006.

[95] J. Gale and B. Sanchez, “The meaning and function of silence in psychotherapy with particular reference to a therapeutic community treatment programme,” *Psychoanalytic Psychotherapy*, vol. 19, no. 3, pp. 205–220, 2005.

[96] N. Ladany, C. E. Hill, B. J. Thompson, and K. M. O’Brien, “Therapist perspectives on using silence in therapy: A qualitative study,” *Counselling and Psychotheraphy Research*, vol. 4, no. 1, pp. 80–89, 2004.

[97] E. Ronningstam, “Cultural function and psychological transformation in psychoanalysis and psychoanalytic psychotherapy,” *The International Journal of Psychoanalysis*, vol. 87, no. 5, pp. 1277–1295, 2006.

[98] D. Tannen and M. Saville-Troike, *Perspectives on Silence*. Ablex Publishing Corporation, 1985. [Online]. Available: https://books.google.com.qa/books?id=5GJjAAAAMAAJ

[99] A. Jaworski, *The power of silence: social and pragmatic perspectives*, ser. Language and language behaviors. Sage, 1993. [Online]. Available: https://books.google.mu/books?id=0NFoAAAAIAAJ

[100] V. L. DeFrancisco, "The sounds of silence: How men silence women in marital relations," *Discourse & Society*, vol. 2, no. 4, pp. 413–423, 1991.

[101] A. Jaworski and I. Sachdev, "Beliefs about silence in the classroom," *Language and Education*, vol. 12, no. 4, pp. 273–292, 1998.

[102] J. Liu, "Negotiating silence in american classrooms: Three chinese cases," *Language and intercultural communication*, vol. 2, no. 1, pp. 37–54, 2002.

[103] M. Sifianou, "Silence and politeness," *Silence: interdisciplinary perspectives*, pp. 63–84, 1997.

[104] J. T. Irvine, "Wolof" magical thinking" culture and conservation revisited," *Journal of Cross-cultural psychology*, vol. 9, no. 3, pp. 300–310, 1978.

[105] I. Weiner, W. Reynolds, and G. Miller, *Handbook of Psychology, Educational Psychology*, ser. Handbook of Psychology. Wiley, 2012. [Online]. Available: https://books.google.com.qa/books?id=VIhKqfn7YwQC

[106] S. Condon and N. Bonvillain, "Language, culture, and communication: The meaning of messages," 1994.

[107] M. Ephratt, "The functions of silence," *Journal of Pragmatics*, vol. 40, no. 11, pp. 1909 – 1938, 2008.

[108] J. V. Jensen, "Communicative functions of silence," *ETC: A Review of General Semantics*, pp. 249–257, 1973.

[109] R. L. Johannesen, "The functions of silence: A plea for communication research," *Western Journal of Communication (includes Communication Reports)*, vol. 38, no. 1, pp. 25–35, 1974.

[110] M. Ephratt, "The functions of silence," *Journal of pragmatics*, vol. 40, no. 11, pp. 1909–1938, 2008.

[111] D. Kurzon, *Discourse of Silence*, ser. New series]. J. Benjamins, 1998. [Online]. Available: https://books.google.com.qa/books?id=jt14WXcWNgsC

[112] I. Nakane, "Negotiating silence and speech in the classroom," *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, vol. 24, no. 1-2, pp. 75–100, 2005.

[113] M. Halliday, C. Matthiessen, and C. Matthiessen, *An Introduction to Functional Grammar*. Taylor & Francis, 2014. [Online]. Available: https://books.google.com.qa/books?id=JM3KAgAAQBAJ

[114] C. Bazzanella, "Le voci del silenzio," *Sul dialogo. Contesti e forme di interazione verbale*, pp. 35–44, 2002.

[115] G. E. Baker, *Rural versus urban political power: the nature and consequences of unbalanced representation*. Doubleday, 1955, vol. 20.

[116] D. Sperber and D. Wilson, "On defining relevance," *Philosophical Grounds of Rationality: intentions, categories, ends*, pp. 143–158, 1986.

[117] B. Brummett, "Towards a theory of silence as a political strategy," *Quarterly Journal of Speech*, vol. 66, no. 3, pp. 289–303, 1980.

[118] N. A. S. Sahin, "Silence as a multi-purpose speech act in turkish political discourse," *Procedia - Social and Behavioral Sciences*, vol. 15, 2011. [Online]. Available: http://gen.lib.rus.ec/scimag/index.php?s=10.1016/j.sbspro.2011.04.233

[119] P. Gilmore, "Silence and sulking: Emotional displays in the classroom," *Perspectives on silence*, pp. 139–162, 1985.

[120] G. R. Saunders, "Silence and noise as emotion management styles: An italian case," *Perspectives on silence*, pp. 165–83, 1985.

[121] D. Tannen, "Silence as conflict management in fiction and drama: Pinter's betrayal and a short story, great wits," *Conflict talk: Sociolinguistic investigations of arguments and conversations, ed. AD Grimshaw*, pp. 260–279, 1990.

[122] Y. Kamide, G. T. Altmann, and S. L. Haywood, "The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements," *Journal of Memory and language*, vol. 49, no. 1, pp. 133–156, 2003.

[123] J. Nivre, "Algorithms for deterministic incremental dependency parsing," *Computational Linguistics*, vol. 34, no. 4, pp. 513–553, 2008.

[124] Y. Kato, S. Matsubara, K. Toyama, and Y. Inagaki, "Incremental dependency parsing based on headed context-free grammar," *Systems and Computers in Japan*, vol. 36, no. 2, pp. 63–77, 2005.

[125] K. R. Thórisson, "Natural turn-taking needs no manual: Computational theory and model, from perception to action," in *Multimodality in language and speech systems*. Springer, 2002, pp. 173–207.

[126] ——, "Mind model for multimodal communicative creatures and humanoids," *Applied Artificial Intelligence*, vol. 13, no. 4-5, pp. 449–486, 1999.

[127] G. Ferguson, J. F. Allen *et al.*, "Trips: An integrated intelligent problem-solving assistant," in *AAAI/IAAI*, 1998, pp. 567–572.

[128] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "Towards a generic dialogue shell," *Natural Language Engineering*, vol. 6, no. 3, pp. 1–16, 2000.

[129] J. Allen, G. Ferguson, and A. Stent, "An architecture for more realistic conversational systems," in *Proceedings of the 6th international conference on Intelligent user interfaces*. ACM, 2001, pp. 1–8.

[130] G. Aist, "Expanding a time-sensitive conversational architecture for turn-taking to handle content-driven interruption," in *Fifth International Conference on Spoken Language Processing*, 1998.

[131] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa, "Learning decision trees to determine turn-taking by spoken dialogue systems." in *INTERSPEECH*, 2002.

[132] A. I. Rudnicky, E. H. Thayer, P. C. Constantinides, C. Tchou, R. Shern, K. A. Lenzo, W. Xu, and A. Oh, "Creating natural dialogs in the carnegie mellon communicator system." in *Eurospeech*, 1999.

[133] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may i help you?" *Speech communication*, vol. 23, no. 1, pp. 113–127, 1997.

[134] A. V. Ivanov and G. Riccardi, "Automatic turn segmentation in spoken conversations." in *INTERSPEECH*, 2010, pp. 3130–3133.

[135] K. A. Ericsson and H. A. Simon, *Protocol analysis*. MIT-press, 1984.

[136] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[137] T. Schmidt and K. Wörner, "Extensible markup language for discourse annotation (exmar-alda)," 2004.

[138] C. J., "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[139] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[140] J. L. Fleiss, "Measuring agreement between two judges on the presence or absence of a trait," *Biometrics*, pp. 651–659, 1975.

[141] G. Hripcsak and A. S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, 2005.

[142] J. J. Gross, "The emerging field of emotion regulation: An integrative review," *Review of General Psychology*, vol. 2, no. 3, p. 271, 1998.

[143] K. R. Scherer, "Psychological models of emotion," *The neuropsychology of emotion*, vol. 137, no. 3, pp. 137–162, 2000.

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

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

[146] M. L. Hoffman, "Empathy and prosocial behavior," *Handbook of Emotions*, vol. 3, pp. 440–455, 2008.

[147] M. Danieli, G. Riccardi, and F. Alam, "Annotation of complex emotion in real-life dialogues," in *Proc. of 1st Italian Conf. on Computational Linguistics (CLiC-it) 2014*, R. Basili, A. Lenci, and B. Magnini, Eds., vol. 1, no. 122–127, 2014.

[148] ——, "Emotion unfolding and affective scenes: A case study in spoken conversations," in *Proc. of Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015,*. ICMI, 2015.

[149] G. R. Firoj Alam, Morena Danieli, "Annotating and modeling empathy in spoken conversations," *arXiv*, 2017.

[150] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the international conference on new methods in language processing*, vol. 12. Citeseer, 1994, pp. 44–49.

[151] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Machine Learning: ECML-94*. Springer, 1994, pp. 171–182.

[152] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, "Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics," in *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*, Athens, Greece, 2009.

[153] S. Quarteroni, G. Riccardi, S. Varges, and A. Bisazza, "An open-domain dialog act taxonomy," University of Trento, Tech. Rep. DISI-08-032, August 2008.

[154] M. G. Core and J. F. Allen, "Coding dialogs with the damsl annotation scheme," in *Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.

[155] D. Traum, "Conversational agency: The trains-93 dialogue manager," in *Proceedings of Twente Workshop on Language Technology, TWLT-II*, 1996.

[156] H. Bunt, "A framework for dialogue act specification," in *In Proceedings of SIGSEM WG on Representation of Multimodal Semantic Information*, 2005.

[157] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary *et al.*, "Towards an ISO standard for dialogue act annotation," in *Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[158] A. C. Fang, J. Cao, H. Bunt, and X. Liu, "The annotation of the Switchboard Corpus with the new ISO standard for dialogue act analysis," in *Workshop on Interoperable Semantic Annotation*, 2012.

[159] H. Bunt, A. C. Fang, X. Liu, J. Cao, and V. Petukhova, "Issues in the addition of ISO standard annotations to the switchboard corpus," in *Workshop on Interoperable Semantic Annotation*, 2013.

[160] V. Petukhova, A. Malchanau, and H. Bunt, "Interoperability of dialogue corpora through ISO 24617-2-based querying," in *LREC*, 2014.

[161] S. A. Chowdhury, M. Calvo, A. Ghosh, E. A. Stepanov, A. O. Bayer, G. Riccardi, F. García, and E. Sanchis, "Selection and aggregation techniques for crowdsourced semantic annotation task," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[162] B. Schuller, "Voice and speech analysis in search of states and traits," in *Computer Analysis of Human Behavior*. Springer, 2011, pp. 227–253.

[163] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[164] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[165] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[166] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[167] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

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

[169] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[170] O. A. Abbas *et al.*, "Comparisons between data clustering algorithms." *Int. Arab J. Inf. Technol.*, vol. 5, no. 3, pp. 320–325, 2008.

[171] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[172] A. Donner and N. Klar, "The statistical analysis of kappa statistics in multiple samples," *Journal of clinical epidemiology*, vol. 49, no. 9, pp. 1053–1058, 1996.

[173] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering." in *NIPS*, vol. 17, no. 1601-1608, 2004, p. 16.

[174] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[175] F. Alam and G. Riccardi, "Comparative study of speaker personality traits recognition in conversational and broadcast news speech." in *INTERSPEECH*, 2013, pp. 2851–2855.

[176] S. A. Chowdhury, G. Riccardi, and F. Alam, "Unsupervised recognition and clustering of speech overlaps in spoken conversations," in *Proc. of Workshop on Speech, Language and Audio in Multimedia*, 2014.

[177] S. A. Chowdhury, M. Danieli, and G. Riccardi, "The role of speakers and context in classifying competition in overlapping speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[179] S. A. Chowdhury, M. Danieli, and G. Riccardi, "Annotating and categorizing competition in overlap speech," in *Proc. of ICASSP*. IEEE, 2015.

[180] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. Nédellec and C. Rouveirol, Eds. Springer Berlin Heidelberg, 1998, vol. 1398, pp. 137–142. [Online]. Available: http://dx.doi.org/10.1007/BFb0026683

[181] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proc. of ACL*, vol. 1, 2014, pp. 238–247.

[182] J. Bian, B. Gao, and T.-Y. Liu, "Knowledge-powered deep learning for word embedding," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 132–148.

[183] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[184] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the International Conference on Learning Representations*, 2013, available as arXiv preprint arXiv:1301.3781.

[185] T. Mikolov and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013.

[186] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms*. Cambridge University Press, 2011.

[187] J. Nickolls and W. J. Dally, "The gpu computing era," *Micro, IEEE*, vol. 30, no. 2, pp. 56–69, 2010.

[188] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of ICASSP*. IEEE, 2013, pp. 6645–6649.

[189] S. Liu, N. Yang, M. Li, and M. Zhou, "A recursive recurrent neural network for statistical machine translation." in *Proc. of ACL*, 2014, pp. 1491–1500.

[190] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *JMLR*, vol. 11, no. Feb, pp. 625–660, 2010.

[191] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[192] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[193] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[194] S. A. Chowdhury and G. Riccardi, "A deep learning approach to modeling competitiveness in spoken conversation," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[195] NIST, *The 2009 RT-09 RIch transcription meeting recognition evaluation plan*, NIST, 2009.

[196] D. Castán, A. Ortega, A. Miguel, and E. Lleida, “Audio segmentation-by-classification approach based on factor analysis in broadcast news domain,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[197] V. P. Richmond, J. C. McCroskey, and S. K. Payne, *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ, 1991.

[198] D. H. Fisher, “Knowledge acquisition via incremental conceptual clustering,” *Machine Learning*, vol. 2, pp. 139–172, 1987.

[199] M. A. Gluck and J. E. Corter, *Information, uncertainty, and the utility of categories*. Proceedings Seventh Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, 1985.

[200] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldi speech recognition toolkit,” in *Proc. of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011, pp. 1–4.

[201] Š. Beňuš, A. Gravano, and J. Hirschberg, “Pragmatic aspects of temporal accommodation in turn-taking,” *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[202] S. Joty, G. Carenini, and C.-Y. Lin, “Unsupervised modeling of dialog acts in asynchronous conversations,” in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011.

[203] M. Tavafi, Y. Mehdad, S. Joty, G. Carenini, and R. Ng, “Dialogue act recognition in synchronous and asynchronous conversations,” in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 117–121.

[204] T. Kudo, “CRF++,” http://taku910.github.io/crfpp/, 2013.

[205] R. L. Oliver, *Satisfaction: A behavioral perspective on the consumer*. Routledge, 2014.

[206] Š. Beňuš, A. Gravano, and J. Hirschberg, “Pragmatic aspects of temporal accommodation in turn-taking,” *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[207] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, “Speech and crosstalk detection in multichannel audio,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[208] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, “Overlapped speech detection for improved speaker diarization in multiparty meetings,” in *Proc. of ICASSP*. IEEE, 2008, pp. 4353–4356.

[209] E. Shriberg, A. Stolcke, and D. Baron, “Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech,” in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[210] G. , *Two explorations of the organization of overlapping talk in conversation.* Tilburg University, Department of Language and Literature, 1982.

[211] C. Lai, J. Carletta, and S. Renals, “Modelling participant affect in meetings with turn-taking features,” in *Proc. Workshop of Affective Social Speech Signals*, 2013.

[212] B. Wrede and E. Shriberg, “Spotting" hot spots" in meetings: human judgments and prosodic cues.” in *INTERSPEECH*, 2003.

[213] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, “Being bored? recognising natural interest by extensive audiovisual integration for real-life application,” *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[214] S. A. Chowdhury, E. A. Stepanov, and G. Riccardi, “Transfer of corpus-specific dialogue act annotation to iso standard: Is it worth it ?” in *Proc. of LREC*, 2016.

[215] J. Carletta, S. Garrod, and H. Fraser-Krauss, “Placement of authority and communication patterns in workplace groups the consequences for innovation,” *Small Group Research*, vol. 29, no. 5, pp. 531–559, 1998.

[216] N. Landwehr, M. Hall, and E. Frank, “Logistic model trees,” vol. 95, no. 1-2, pp. 161–205, 2005.

[217] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, “Evaluating spoken dialogue agents with paradise: Two case studies,” *Computer Speech & Language*, vol. 12, no. 4, pp. 317–347, 1998.

[218] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing: Survey of an emerging domain,” *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[219] M. Pantic, R. Cowie, F. D'Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli, "Social signal processing: the research agenda," in *Visual analysis of humans*. Springer, 2011, pp. 511–538.

[220] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[221] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, pp. 45–60, 1999.

[222] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[223] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[224] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[225] M. Koutsombogera, D. Galanis, M. T. Riviello, N. Tseres, S. Karabetsos, A. Esposito, and H. Papageorgiou, "Conflict cues in call center interactions," in *Conflict and Multimodal Communication*. Springer, 2015, pp. 431–447.

[226] S. A. Chowdhury, E. Stepanov, and G. Riccardi, "Predicting user satisfaction from turn-taking in spoken conversations," in *Proc. of INTERSPEECH*, 2016.

[227] A. Firoj, D. Morena, and R. Giuseppe, "Can we detect speakers' empathy?: A real-life case study," in *7th IEEE International Conference on Cognitive InfoCommunications*, 2016.

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

[229] F. Alam and G. Riccardi, "Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition," in *Proc. of ICASSP2014 - SLTC*, May 2014.

[230] F. Alparone, S. Caso, A. Agosti, and A. Rellini, “The italian liwc2001 dictionary.” LIWC.net, Austin, TX, Tech. Rep., 2004.