Paper Instructions and Template for Interspeech 2025

Anonymous submission to Interspeech 2025

Abstract

Conflict (de-)escalation speech corpora are limited in the literature. Of the corpora that exists, the domain of the data is usually from legal or political debate settings. However, conflict escalation and de-escalation are a critical part of the human experience. Yet, there is a lack of data-driven analysis of the speech signals (e.g. prosody, shimmer, F0) that characterize (de-)escalation in everyday debates and collaborations. This paper is a starting point for addressing this gap by using data-driven approaches to analyzing the statistical features and speech signals common to starting and ending ones turn in collaborations and debates from conversations about furthering the development of a hypothetical town. More specifically, understanding the implications of our speech in tandem with what speech signals indicate the beginning and end of a turn could further characterize these situations from a more empirical perspective.Thus, two key research questions are addressed. First, what speech signals are most common for predicting when a turn would start vs end when debate is the goal? Secondly, what speech signals are most common for predicting when a turn would start vs end when collaboration is the goal? Lastly, what clusters of speech signals characterize collaboration vs a debate? Few-shot learning with an LLM will be utilized to classify dialogue acts to capture the action of the speech being produced. Moreover, feature importance will be performed to understand what linguistic cues and speech signals hold more weight with determining when turns are likely to begin or end. Lastly, k-means clustering will be utilized on the speech signals and dialogue acts to identify natural groupings of turn-taking behaviors in debates vs collaborations. Overall, this analysis will contribute to improving our understanding of (de)-escalation in a more casual setting, and to operationalize speech for the training of an AI mediator.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Escalation creates risk in interaction with the potential to result in harm, while de-escalation mitigates such risks. Having a data-driven approach to understanding de-escalation can standardize approaches to reducing the intensity of a conflict, and keep civilians safe. De-escalation is an understudied phenomenon from an empirical perspective in general Ecologically valid (de-)escalation corpora is limited in the literature. Researchers do not want to replicate real-life conflict entirely. Another challenge is that most data related to escalation and aggression in laboratory environments involve at least one simulated interlocutor or confederate. Debate is a common setting where escalation occurs due to the competitive telos of the task, and debate corpora exists in the public domain. However, the domain of these debates is in a formalized setting (describe the different kinds of debates). While escalation does occur in this context, this still does not replicate escalation that may occur from everyday conflicts. For example, in a certain kind of debate, logical fallacies such as Ad-Hominium attacks are not permissible. These boundaries do not exist in everyday conflicts. De-escalation is the reduction of escalation. Due to the lack of escalation data in informal, naturalistic settings, this contributes to the limited de-escalation corpora in these settings. As a result, there is a need to investigate escalation and de-escalation from a more data-driven perspective, namely through performing statistical analysis on speech signals is needed. Moreover, collaborative settings are where de-escalation become more feasible. Proper collaboration involves all parties working together effectively to come to consensus on a solution, based on a common goal. This idea applies in the work place, where collaborating on a solution with the goodwill of both parties in mind constitutes as a vehicle for de-escalation, and conflict management prevents workplace bullying [29, 42]. Collaborations can arise spontaneously, or through a formal process. One domain where collaboration is encouraged for conflict resolution is mediation. Mediation is where a third party, called a mediator, facilitates the resolution of a disagreement between individuals to better understand the concerns of each party and come to an agreement [9]. One significant area of conversation is the processes of turn taking. To address this gap, this research utilizes the Several Paired Interactions in Conflict (de-)Escalation (SPICE) corpus as a means of performing this analysis. The SPICE corpus is a multimodal corpus that comprises of debates and collaborations about furthering a hypothetical town in a gamified setting. The organic debates and collaborations are hypothesized to result in conflict escalation, as well as de-escalalation. Through the organiz debates and collaborations in this corpus, an improved analysis of the linguistic cues and speech signals could be performed to enhance and formalize our understanding of turn-taking in competitions vs collaborations in a more ecologically valid manner, key factors in de-escalation. This corpus is detailed in the methods section. This study pursues three research questions First, what speech signals are most common for predicting when a turn would start vs end when debate is the goal? Secondly, what speech signals are most common for predicting when a turn would start vs end when collaboration is the goal? 3a) Which combinations of features produce the most distinct clusters? 3b) Of the clusters discovered in 3a, which of these map onto collaborations vs debates? The corresponding hypotheses, as established by the related works section, make the following predictions H1: repeated backchannels signal from the (non-speaking?) interlocutor will determine when a turn is likely to end. H2: In competitive settings, there H3:

Competitive settings will be distinguished with a cluster of experiments with higher average intensity intensity (figure out a way to formalize higher intensity)

Though utilizing a multimodal approach on organic debates and collaborations to formalize turn-taking strategies in everyday collaborations and competitions, we can better model and understand de-escalation. We can also apply this analysis and data to machine-mediated communication in these contexts [6]. For example, we can operationalize speech for the training of an AI mediator.

2. Related Work

2.1. Corpora that Involves Competitive and Collaborative Domains

- Tie competition back to escalation, collaboration back to de-escalation
- Two primary categories that corpora containing spoken dialogues can be categorized into are spontaneous and task-oriented.
- list cons of this - often times involves simulated interlocutors or confederates. Acted data is not ecologically valid, and leads to less authentic results.
- Moreover, outside of the laboratory, corpora can be obtained from more naturalistic settings.
- Two common examples are recordings of political or academic debates (discuss 1)
- list cons of this. These formal debates adhere to rules that do not apply to real-life competitive settings. For instance, utlilizing logical fallacies violate the rules of academic debates, and yet, logical fallacies are very common in escalated debates in our everyday lives.
- Collaborations have also been captured in naturalistic settings.
- For instance, [5] utilized transcripts from remote meetings of a global technology company to identify and predict competitive and collaborative turn overlaps. Despite the fact that collaborations were captured in the data successfully, the researchers note that competitive and collaborative overlaps in these settings are distorted because of the asymmetric nature of videoconferencing. Thus, the data-driven intepretations of these interactions fail to capture the true nature of natural conversation.
- State gap

Competitive debates often comprise of four key formats "The competitive debate corpus comprises four formats: British Parliamentary debate (BP), Lincoln-Douglas debate (LD), Karl Popper debate (KP), and Oxford debate (DOx).4"

2.2. Linguistic Turn-Detection

- A linguistic turn can be defined as...
- Turns can be categorized into holds or switches

2.3. Comparing Competitive and Collaborative Strategies in Corpora

Generally speaking, in competitive settings, the speakers vie for attention and the floor, resulting in reluctance to lose the floor. Backchannels in this setting are generally terse. In collaborative settings, back channels are more encouraging.

- Contextualizing turns and speaker intentions could benefit from utilizing pragmatic features
- discuss (4) here
-
- Thus, competitive and collaborative interactions have been captured in a variety of settings, in both laboratory and naturalistic settings. Yet, laboratory experiments eliciting collaborative and competitive dialogue often involve a simulated interlocutor or a confederate, leading to acted data. The data captured from academic debates and political debates involve the usage of rules that do not reflect the social norms of real, dynamic competitions. The videoconferencing domain suffers from asymmetry in interactions between interlocutors, which is also not representative of standard, naturalistic collaborations. This highlights the need to analyze a more ecologically valid corpus, without simulated interlocutors, confederates, and with organic debates and collaborations reflective of causal conversation. Utilizing the interactions in the SPICE corpus for this purpose will satisify this gap. More specifically, through utilizing automatic speech signal extraction, a more nuanced understanding of This will contribute to a more generalizable and rich analysis of two key features of escalation and de-escalation: competition and collaboration

3. Methods

3.1. The SPICE Corpus

The SPICE (Several Paired Interactions in Conflict (De-)Escalation) Corpus is a multimodal corpus comprising dialogues elicited to capture moderate escalation and de-escalation in three novel, gamified tasks involving a hypothetical town's development, hypothesized to spark escalation and de-escalation. Unlike prior studies, two interlocutors interact in competitive and collaborative scenarios, enhancing ecological validity. In addition to the dyads being recorded, they rate interactions for politeness, respect, frustration, and escalation/de-escalation after each of the three tasks.

This corpus was created through a within-subjects data collection experiment uses a 3x1 factorial design, comprising of three tasks. A diagram of the experiment flow is shown in figure 1.

In each, dyads engage in dialogue followed by self- and partner-ratings for escalation, respect, politeness, and frustration, after discussing a sequence of questions under either baseline, competitive or collaborative conditions. These ratings are averaged to represent one pair of participants. See Figure **Y** for visuals of the competitive and collaborative tasks.

As for rating politeness, respect, frustration, and escalation, 5 point Likert scales are used. The participants complete the data collection experiment in pairs. Each pair of participants complete all three tasks in one session. After each task, participants complete short questionnaires, rating their and their task partner's politeness, respect, and frustration levels, and the level of escalation. They also answer some distraction/fact questions to avoid revealing the factors being studied. Finally, they complete a post-experiment questionnaire, which asks if they have had experience with de-escalation training. The features of the SPICE Corpus are shown in Figure 3.

3.2. Preprocessing

Voice recordings will be transcribed with the Whisper API, and we will include timestamps of the linguistic units in milliseconds. Linguistic disfluencies are important for contextualizing

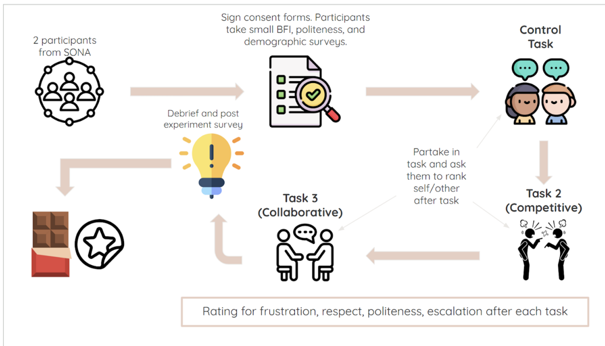

Figure 1: *Fig. 1. Flowchart describing the process of the experiment. Participants fill out The Big Five Short Inventory survey [62], a survey inspired by the Lernman Politeness Survey [47], and a Demographic survey using Qualtrics prior to arriving at the laboratory. Once they are in the laboratory for data collection, the participants complete the three tasks in order (baseline, competitive, collaborative) and for each they rate theirperceived and their partner's perceived escalation, politeness, respect, and frustration each time. Finally, the participants complete a post-experiment survey, are debriefed, and then receive chocolate, stickers, and SONA credits.*

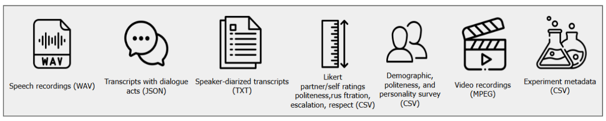

Figure 2: *Fig. 2. A visual of the features in the multimodal SPICE Corpus. The data is multilevel, and can be analyzed from the levels of pair ID, participant ID, and utterance.*

escalating and de-escalating interactions, so we will preserve the meaningful disfluencies within the corpus. The captured dialogues were computationally annotated for dialogue acts by research assistants following the **ISO** framework. This framework is particularly useful because it is the state of the art for dialog act classification. Moreover, it is a framework that is task and domain flexible, so the dialogue acts annotated could be modified with respect to the task or domain. A few-shot learning pipeline utilized these examples to perform an automatic dialogue act classification task using a local Llamma 2 model. they will be processed using automatic speech recognition with a reasonably low word error rate, using post-correction as needed, for feature extraction. Automatic speech signal extraction will be performed using Pratt. While the unit of the corpus will at the utterance level, following the formalized definition of an utternace as noted above, linguistic turn detection will be performed

SHAP will be utilized for determining feature importance of each of the features to address RQ . Moreover, dist

4. References