Multimodal classification of interruptions in humans' interaction

Liu Yang, Catherine Achard, Catherine Pelachaud

Multimodal Analysis of Interruptions

Liu YANG, Catherine ACHARD, and Catherine PELACHAUD

Institut des Systèmes Intelligents et de Robotique (CNRS-ISIR)
Sorbonne University, 75005, Paris, France
{yangl, catherine.achard, catherine.pelachaud }@isir.upmc.fr

Abstract. During an interaction, interactants exchange speaking turns. Exchanges can be done smoothly or through interruptions. Listeners can display backchannels, send signals to grab the speaking turn, wait for the speaker to yield the turn, or even interrupt and grab the speaking turn. Interruptions are very frequent in natural interactions. To create believable and engaging interaction between human interactants and embodied conversational agent ECA, it is important to endow virtual agent with the capability to manage interruptions, that is to have the ability to interrupt, but also to react to an interruption. As a first step, we focus on the later one where the agent is able to perceive and interpret the user's multimodal behaviors as either an attempt or not to take the turn. To this aim, we annotate, analyse and characterize interruptions in human-human conversations. In this paper, we describe our annotation schema that embeds different types of interruptions. We then provide an analysis of multimodal features, focusing of prosodic features (F0 and loudness) and body (head and hand) activity, to characterize interruptions.

Keywords: Interruption, Dyadic interaction, Multimodal signals, Turn taking

1 Introduction

Human-computer interfaces are becoming more and more frequent and appreciated in daily life, and the development of Embodied Conversational Agents (ECAs) is booming as they allow very natural interactions, without artifices. However, many difficulties arise since natural interactions are very complex and involve a multitude of research areas going from psychology to signal processing. A lot of work has already been done, both on verbal and non-verbal signals, and several embodied conversational agents have already been developed. However, one important faculty has not yet been sufficiently studied: the interruptions.

They are however very frequent in natural conversations [6] and appear when one interlocutor attempts to grab the turn while the other person is still holding it. Interruptions are an integral part of the turn-taking mechanism. In some early studies , interruptions were described as a symbol of dominance and power[22, 53, 37], since most of the conversations follow the rule of one-person-speaks-at-a-time. However, interruptions are essential in natural interactions, they help

to regulate the rhythm of the dialogue, to show an interest, to reinforce the engagement [57].

During natural interactions, speakers exchange turns quickly and naturally. Humans are able to predict the end of their partner's turn in order to smoothly take the floor [15], without any discontinuity in the fluidity of the exchange. In the same way, humans can easily recognize when their partners are displaying a backchannel as a sign of participation in the discussion. When an interruption occurs, the speaker can decide to give or not the speaking turn to the interruptee.

Our aim is to create Embodied Conversational Agent ECA able to engage their human interlocutor in natural interaction. We believe it is important to give ECA the ability to manage interruption [8], either by interruption their human interlocutor or by responding to an interruption. To this aim, the ECA should recognize when its human interlocutor produces multimodal signals if it is a backchannel or an interruption, that is an attempt or not to grab the speaking turn. The ECA should be able to recognize the different types of speaking turn exchanges.

To reach this objective, we study natural speaking turn exchanges in human-human interaction gathered in the dyadic corpus NoXi [9]. We propose a schema of annotate interruption. We also provide an analysis of multimodal features to study the non-verbal behaviors involved during each type of turn switches. Our goal is to define which features are used by humans to understand the situation and endow them to ECA. As multimodal features we consider prosodic features and body (head and hands) activity.

We first start by presenting studies on turn-taking and more particularly on interruption in human-human interaction in the following section. In Section 3, we follow by presenting a state of the art on existing works that focused on predicting turn-taking exchange and interruptions. Section 4 presents the NoXi corpus and Section 5 the annotation we have conducted. The multimodal features we have extracted automatically are presented in Section 6 and their analyses are described in Section 7.

2 Background

In this Section, we introduced major works on turn-taking and how they are marked multimodally.

2.1 Modeling turn taking

The study of interaction has interested many scholars since long. Emanuel A. Schegloff [47] defined sequencing rules that manage natural conversations. Ten years later, Harvey Sacks [46] proposed the idea of conversation analysis and described its most basic structure as turn-taking. Actually, interlocutors have to coordinate and exchange speaking floor based on rules to maintain the conversation with the hypothesis they cannot speak and listen at the same time. Using this basic structure, the turn taking, Kendon [38] and Duncan [18] introduced a model of conversation that uses three basic signals:

- Turn-yielding signals from the speaker: the listener may take the turn when a turn-yielding signal is displayed; the speaker yields the turn when the listener shows a willingness to take the floor.
- Attempt-suppressing signals from the speaker: the speaker uses attempt-suppressing signals to maintain the turn and prevent the listener to take the turn.
- Backchannel signals from the listener: the listener gives feedback information. It is not attempting to take the turn. It is not considered as a turn.

Sacks, Schegloff and Jefferson [46] proposed a conversation turn-taking model, often referred as the SSJ model, indicating the turn-taking mechanism. It is based on rules such as: (*i*) The current speaker may select the next speaker, the selected person must speak next. (*ii*) If the current speaker selects no one, then one of the participants may self-select to speak next. (*iii*) If no one is self-selected, the current speaker may continue to speak or terminate the conversation.

Sacks and colleagues made the hypothesis that interlocutors predict rather accurately the turn end timing, leading to 'no gap, no overlap' between speaking turns. However, Coates [12] analysed the distribution of timing interval during turn exchanges. He found a high number of overlaps occurrence at the end of a turn in different conversation settings, thus refuting the hypothesis 'no gap, no overlap'.

2.2 Taxonomy of speaking turn exchanges

Schegloff and Sacks [49] proposed to study some specific speaking turn exchanges corresponding to simultaneous speeches, that are classified as either, interruption, overlap or parenthetical comments such as backchannels. Backchannels are actually not tending to disturb the speech flow or to grab the floor, they are short messages to show the listener's attention, or if the listener agrees or not with the speech [1]. Overlap is when the listener takes the floor that the speaker is yielding but has not yet completed her speech; thus an overlap usually arises on the last word(s) or syllable(s) of the current speaker and the first word(s) of the listener (next speaker) [46]. On the contrary, interruption occurs when the listener grabs the floor against the speaker's wishes [49] without letting the speaker finish his/her utterance, and is described as a violation of the current speaker's turn which overlap is not [43].

Beattie [6] proposed another taxonomy of speaking turn exchanges using both simultaneous speech and willingness to yield the floor, as shown in Figure 1. The three considered classes are overlap, interruption and smooth switches.

Goldberg [24] described a taxonomy based on interruption meaning. He considered two main types: competitive and cooperative interruptions. Competitive interruptions are when the listener interrupts to take the control of the interaction, and disrupt the flow of dialogue between the partners, which can be seen as a conflict:

- Disagreement: The listener disagrees with the current speaker and expresses immediately his/her own opinion.

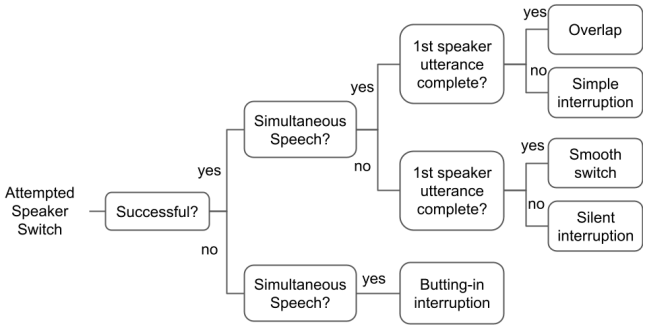


Fig. 1. Classification of interruption and smooth speaking turn exchange [6]

- Floor taking: The listener grabs the floor and expands on the current speaker's topic.
- Topic change: The listener grabs the turn and changes the current topic of conversation.
- Tangentialization: The listener grabs the turn and sums up the information received from the current speaker to prevent listening to more unwanted information.

On the opposite, a cooperative interruption helps to complete the conversation:

- Agreement: The listener shows understanding or support to the speaker.
- Assistance: The listener interrupts to provide the current speaker with a word, a phrase or an idea to help complete the utterance.
- Clarification: The listener expects the current speaker to clarify or explain the information about which the listener is not clear.

2.3 Characterization of speaking turn exchanges

Most of the studies indicate the importance of prosodic features such as fundamental-frequency (F0) or intensity during speaking turn exchanges [23, 39, 54]. Studies found that people raise their energy and voice when they attempt to interrupt the current speaker [50, 26]. Hammarberg [27] provided similar evidence regarding pitch and amplitude.

Features of the interrupters such as speech rate, cutoffs and repetitions are also analyzed by conversational analysts. For example, Schegloff [48] found that variations in prosodic profiles and repetitions are used by interrupters. He also mentioned that interrupting sentences usually have a faster speaking rate, thus providing evidence about the role of speech rate to deal with speakership conflicts.

Gravano and colleagues [26] analysed acoustic features of a telephonic conversation corpus and showed significant differences for interruptions in intensity

and pitch level, speaking rate and Inter-Pausal Unit IPU duration. An Inter-Pausal Unit (IPU) corresponds to a sequence of words surrounded by silences of 50ms or more.

2.4 Characterization of turn ending

Duncan [18] characterized speaker's multimodal signals at the end of the turn:

- phrase-final intonation other than a sustained, intermediate pitch level;
- a drawl on the final syllable of a terminal clause;
- the termination of any hand gesticulation;
- a stereotyped expression;
- a drop-in pitch and/or loudness in conjunction with a stereotyped expression;
- the completion of a grammatical clause.

Duncan also mentioned that the higher the joint frequency of these cues, the greater the probability that the listener take the floor. However, when the speaker is gesturing, the incidence of listener turn-taking attempts falls to zero. The speaker gesticulation is identified as an attempt suppression cue that cancel out the effects of turn-ending cues.

Another turn-taking attempt suppression cue has been proposed by Beattie [5] and Ball [2] that showed that filled pauses reduced the probability of a speaker-switch, at least for a short period after their occurrence.

To estimate the end of a speech turn, Riest *et al.* [45] and De Ruiter *et al.* [15] argued that semantic information is useful while contextual information is employed in [55] or [7].

Stivers et al. [52] mentioned that the intervals between speech turns last on average 200 ms. But psycholinguistic research has shown that it takes at least 600 ms to produce even a single word utterance [31]. Thus, certain cognitive processes by the next speaker must be involved to predict the end of the turn of the current speaker [16, 30]. Further results show that predictive processes work with other processing layers that allow simultaneous production planning and comprehension [15]. However the prediction of turn ending timing is sometime not that accurate and thus cause simultaneous speeches.

3 State of Art

In a conversation, the listener (to be the next potential speaker) has to project the exact timing when the current speaker will finish and what will be her words, so the listener can prepare to initiate the next turn at an appropriate timing and start planning what to say [30]. Several computational models are geared to predict turn-taking and to detect interruptions. They rely on models proposed by conversational analysts [49, 15].

3.1 Turn-taking prediction

A lot of studies on turn-taking prediction (turn ending or turn starting timing) have been done to investigate feature sets and prediction models. The majority of investigated feature sets include prosodic features such as fundamental frequency (F0) and energy [44, 25]. Linguistic features were also investigated such as syntactic structure, turn-ending markers, and language model [36, 42, 35]. Moreover, multimodal features, such as eye-gaze [14, 32], respiration [29, 33], and head-direction [51], were also considered.

Hara et al. [28] took into account the concept of the Transition Relevance Place (TRP) in accordance with prosodic, speech and linguistic features to predict whether to take the turn at each instant by calculating the posterior probability of turn-switch. Each feature is modeled by an individual LSTM and the outputs of those LSTMs are concatenated and fed into a linear layer that outputs the posterior probability of the output label. More precisely, a first sub-model is used, using a single user, to detect TRP at the end of each IPU. Then, a second sub-model predicts whether to take the turn by calculating the posterior probability of turn-switch. An accuracy of 89.5% is obtained as the best result.

Ishii et al. [34, 35] proposed to predict the turn management willingness of speaker and listener to help predict the occurrence of turn switch using acoustic, linguistic and visual (gaze, head movement, respiration) cues from both speaker and listener. Results show that turn-management willingness and turn-exchange are predicted most precisely when all modalities from speaker and listener are used.

Coman et al. [13] proposed to build an automated system capable of estimating the dialog state and the appropriate turn-taking point token-by-token (word by word) in an incremental setting by exploiting lexical features. The authors developed two modules: an incremental Dialog State Tracker (iDST) that consists of an encoder-based classifier to track the dialog state change after each token (word); and an incremental Turn-Taking Decider (iTTD) that takes as input the output of iDST, and that decides if a turn switch should happen or not.

3.2 Interruption prediction

Several works have been dedicated to predicting when an interruption may occur.

Lee and Narayanan [41] use a hidden conditional random field (HCRF model) to predict occurrences of interruption in dyadic conversations. They found the following cues:

- Interrupter: mouth opening distance, eyebrow and head movement
- Interruptee: energy and pitch values of audio

The authors annotated the turn transitions into two classes: smooth transition and interruption. Their model predicts the upcoming turn exchange type with the behavior of the interrupter and interrupted one second before the relevant transition point.

Chýlek et al. [11] presented their study to predict the speaking turn switch timing. Three types of overlaps are defined by the voice activity: internal overlap (INT), overlap resulting in a switch of turns (OSW) and a clean switch of turns (CSW). INT corresponds to the case where speaker B starts speaking during speaker A's utterance, but speaker A continues his turn, OSW to the case where speaker A ends his utterance during the overlapping segment and CSW occurs when there is no overlap during the turn exchange. INT and OSW samples refer to overlaps (OVR), which are considered as interruptions in their work. Prediction of all three types (INT, OSW and OVR) is tested separately. The authors tested different ML models such as support vector machines, decision trees, and neural networks. The deep residual learning networks (ResNet-152) with only acoustic features gave the best performance.

Other works have been more interested in determining the interruption types and more particularly on classifying between cooperative and competitive interruptions. Yang [57] reported that competitive interruptions have higher pitch and intensity levels, while collaborative interruptions have a relatively lower pitch level. Lee and Narayanan [40] proposed a multimodal analysis method to classify the interruption type. They observed that the absence of hand motions signal the occurrence of cooperative interruptions with high probability. Moreover, the number of occurrences of disfluencies in the speech is significantly higher in the case of competitive interruptions. Their best classification results of interruption type combine hand motion with speech intensity. Khiet and colleagues [54] used SVM to classify overlaps with acoustic features, gaze behaviour and head movement annotations. With a delay of 0.6s after the start of overlap, the model begins to show a good accuracy only using overlapper's acoustic features. They also mentioned that slight improvement was obtained when gaze information during overlap was added, while adding acoustic information from the overlappee did not improve performance. Chowdhury et al. [10] classified competition and cooperative overlapping speech using a Sequential Minimal Optimization (SMO) model with prosody, voice quality, MFCC, energy and spectral features. Egorow et al. [19] considered two emotion dimensions (*control* and *valence*) combined with acoustic features to classify overlaps with SVM model.

4 Corpus

We choose the NoXi corpus [9] for our study. NoXi is composed of multimodal data (video and audio) that contains free dyadic interactions. For each dyad, both interactants have been recorded separately (video and audio) during a screen-mediated interaction, allowing to easily separate the audio sources, as shown in Figure 2. The video of each interactant shows almost their full body except the feet. Both interactants' audios and videos have been synchronized and transcribed.

In the NoXi database, participants take the role of either 'expert' or 'novice'. An 'expert' shares her knowledge on a subject (among over 45 given topics) with a 'novice' who is interested in that subject. Each interaction is about 20mn long.

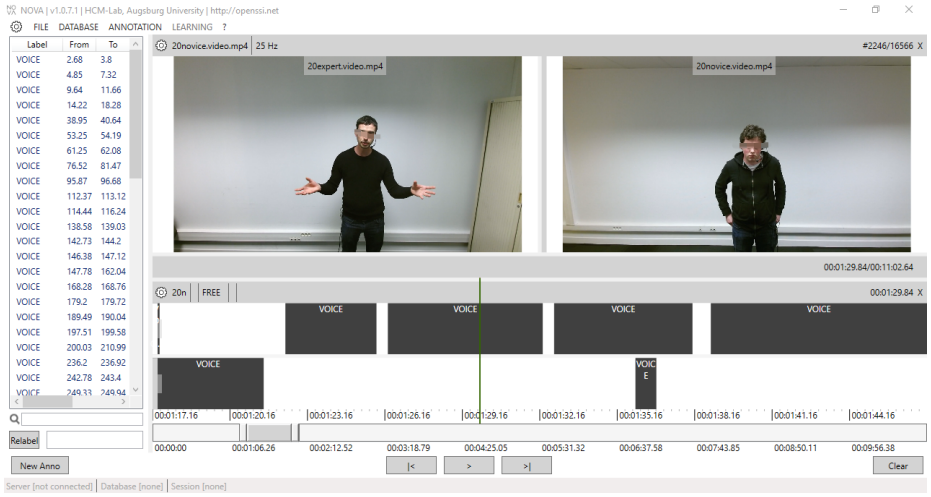


Fig. 2. Example of NoXi dyads

NoXi database has been acquired in seven languages. We chose the French part of NoXi corpus for our study, including 21 dyadic conversations, for about 7h in total (21*20mn).

5 Annotation

We use this database to study the difference in multimodal signals occurring during different speaking turn exchange types. In a first step, we annotate the database based on voice activity detection VAD. For each of these changes in the vocal track, we annotate if it is a backchannel, a smooth speaking turn exchange or an interruption.

Three taxonomies of speaking turn exchanges have been presented in the background section. Altogether, they cover most of the situations we may encounter in daily conversations; but individually, they lack some details. Schegloff and Sacks [49] focused only on simultaneous speeches and thus, did not consider smooth speaking turn exchanges or interruptions that happened during a silence. Beattie's taxonomy [6] does not include backchannel and Goldberg's one [24] includes only different types of interruption. So, we propose a new annotation schema that merges and completes these three taxonomies.

Before presenting our taxonomy, we introduce some definitions.

5.1 Definition

The three speaking turn exchanges we consider in our study are defined below.

Smooth turn exchange: These speaking turn exchanges occur when one person ends speaking and another person takes the floor. Very often there is a

short silence between the two speeches. But sometimes, a short overlap exists as the listener has anticipated the end of the speaker's utterance [30]. The main characteristic of smooth turn exchange is that there is a willingness of the speaker to give up her speaking turn.

Backchannel: A backchannel is a multimodal signal, verbal and nonverbal, that the listener displays to indicate that she/he is listening, to show attitudinal (e.g. agreeing or not) and emotional (e.g. happiness) reactions to what the speaker is saying [1]. There is no desire to take the floor but just to show engagement and reactions in the conversation.

Interruption: During an interruption, the listener aims to take the floor while the speaker has not produced signs to yield the turn. It is common that an overlap is present during an interruption but this is not always the case; for example if the speaker is searching for a word, the listener may take this opportunity to bring up a new idea.

Interruptions can be further classified in **successful or failed interruptions**. A successful interruption corresponds when the listener grabs and maintains the speaking turn and the current speaker stops talking.

In a failed interruption, the current speaker does not give up her speaking turn and continues talking. We illustrate this distinction through examples taken from the NoXi database (and translated from French to English).

The following situations can be considered as a successful interruption:

- The interrupter grabs the turn successfully and the current speaker has to quit even she/he has not finished the current utterance.
  Example:
  Person A:... it's like sports, it's not physical, basically not phy[sical but I ...]
  Person B: [I agree with] you for example to train in football...
- The listener speaks over the speaker (e.g. by asking quickly a clarification question). The speaker doesn't stop and keeps her turn, but takes into account what the listener says (e.g. by answering the listener's question).
  Example:
  Person A: ...sometimes you can see the mushrooms, that's why you [have to be care]ful. yeah, especially the optics...
  Person B: [Mushrooms?]

Here are examples of failed interruption:

- The listener abandons the interruption before his utterance is completed and let the current speaker continues her turn; the speaker does not pay extra attention to the listener's attempt.
  Example:
  Person A: ...for competitions, maybe I'm wrong and I see your point of view, I unders[tand but finally] maybe it's easy ...
  Person B: [Ah no no you...]

– The listener begins to speak to get attention from the speaker. He does not respond after completion of his utterance. The speaker continues her speech.
Example:
Person A: ...I didn't pay even one euro for Hearthtone, and I uh I still [have my meta decks] up to now, I can...
Person B: [Ah me neither, it's useless no?]

5.2 Schema

In this section we introduce our annotation schema composed of three levels as presented in Figure 3.

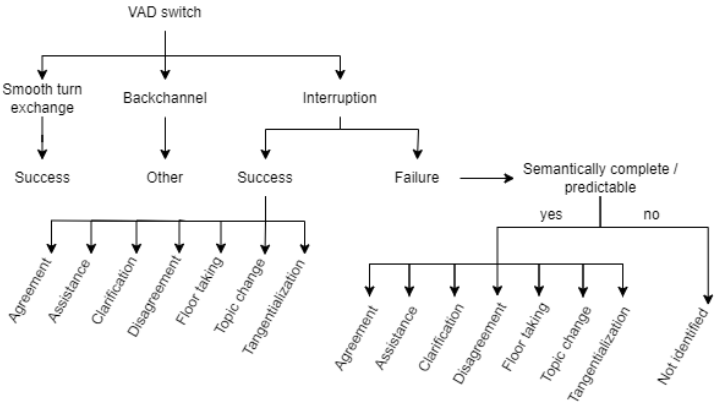


Fig. 3. Interruption annotation schema

At the first level, each VAD switch is classified into *interruption, backchannel* or *smooth turn-exchange* according to the definitions presented above.

The second level deals with the accomplishment of speaking turn exchange. Smooth turn exchanges are always annotated as successful (*success*) and backchannels, which are not aiming to grab the turn, are annotated as *Other*. Interruptions can be annotated as successful (*success*) or as failed (*failure*).

Finally, in the third level, the type of interruptions is annotated based on the speech content using the eight classes proposed by [24]: Agreement, Assistance, Clarification, Disagreement, Floor taking, Topic change and Tangentialization. Successful interruptions can easily be classified into their different classes while failed interruptions can be too short to perform this classification. In such a case they are annotated as *Not identified*. As in [24], we group the classes in "two super-classes" where Agreement, Assistance, Clarification belong to the cooperative interruption class while Disagreement, Floor taking, Topic change and Tangentialization belong to the competitive interruption class. The annotation

of interruption class at this level is based on the linguistic analysis of what is being said.

5.3 Annotation process

We annotate NoXi database using our annotation schema. The annotation is done in three steps, one per level. For each dyad, we use the Nova tool [4] to display and synchronize the visual and audio channels of the video of both interactants. Annotation is done semi-automatically. When it requires semantic analysis, we rely on manual annotation. We now describe the steps we follow.

At first, we apply the automatic Voice Activity Detection (VAD). It gives us points where both interactants speak simultaneously or when there is a change of speaker. The next step consist to classify these points in either backchannel, smooth turn exchange or interruption (level 1). This classification requires analyzing the linguistic terms, and cannot rely solely on acoustic features. This step is done manually.

The distinction between failed and successful interruptions is done manually (level 2). For each occurrence of interruptions annotated in the first step, we listen to the occurring speech of the current speaker and of the other participant of the dyad.

The last step requires to analyse what is being said, to understand if an interruption is a cooperative or a competitive one, and which class among the eight possible it is (level 3).

In order to measure the annotation accuracy, all videos have been annotated three times by the same annotator, following the same process as just described. There was one-month interval between each round of annotation to ensure that the annotator has forgotten the video content and the annotations.

To compute the annotation accuracy, we consider the agreement for annotation on the three levels. After the first two rounds of annotation, we accepted all records with full agreement (for the three levels). For the other records, we applied a third round of annotation. After this third round, we accepted the records for which the three annotation levels (levels 1, 2 and 3) are identical to the first or second round annotations. We disregarded the other records (318 among 4301).

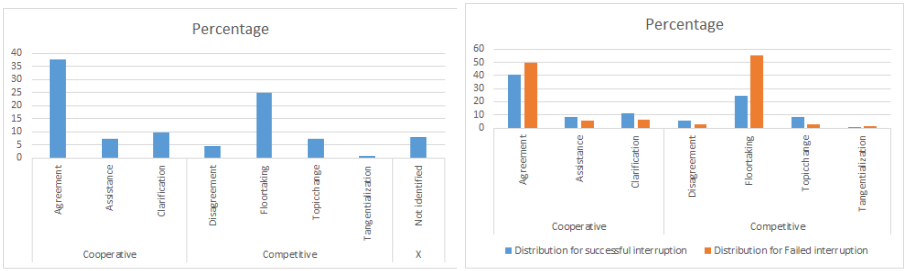
After three rounds of annotation, we have a global annotator self-acceptance of 92.6%. When comparing level 1 annotation value over the three annotation rounds, we have an agreement level of: 84.07% for interruption, 92% for smooth turn exchange and 98.8% for backchannel.

5.4 Annotation analysis

Following this annotation process, we obtain 3983 VAD switch points for the French part of the NoXi database. Among them, there are 1403 smooth turn exchanges, 1651 backchannels and 929 interruptions. When, removing backchannels that do not correspond to a speaking turn exchange, interruptions represent 33%

of turn-taking situations. This reinforces our intuition that they have a fundamental role in natural interactions.

Considering interruption, most of them (81.7%) are successfully performed and thus, the speaker succeeds in taking the floor. Moreover, among the successful interruptions, there are almost as many cooperative interruptions (54.36%) as competitive ones. The probability distribution according to the 8 types of interruption is given in Figure 4. *Agreement interruptions* are the most frequent ones over all the interruption types. They represent also the majority of the cooperative interruptions. For competitive interruptions, *Floor taking* ones are the most frequent. *Floor taking* interruptions do not involve a change of topic. The distribution of interruption classes may be specific to the NoXi database as it involves dyads chit-chatting on topics that one person wishes to share information about and another person wishes to learn about it. The interaction setting is rather a friendly one. This is congruent with the majority of interruption types that is found in the corpus.



(a) Distribution according to the 8 types of interruption.

(b) Distribution according to the 8 types of interruption for failed and successful interruptions.

Fig. 4. Distribution of interruption.

For 44.71% of failed interruptions, the type cannot be determined because they are too short to understand the meaning of the speech. When the type has been determined, the types distribution of successful and failed interruptions is represented in Figure 4. We find again that *Agreement* and *Floor taking* interruptions are the two largest types for both competitive and cooperative interruptions, whatever their accomplishment.

6 Features

For every conversation, we extracted separately the acoustic, facial and body features of each participant.

- The acoustic features are extracted using Opensmile [21], including: F0, loudness and 13 features of MFCC. We normalize each acoustic features by subtracting its mean value along the whole sequence.
- The facial expressions are extracted using Openface [3] and encoded with Action Units (coded with Facial Action Coding System (FACS) [20]).
- The facial positions are also estimated using Openface [3]. They include head movement (position & rotation in x-y-z axis) and gaze direction.
- The body features are extracted using Alphapose [56] and are composed of positions of 15 key joints except the feet (position in x-y axis). To standardize the position features, instead of using the absolute positions provided by Alphapose and Openface, we center the position by taking the middle of the two shoulders as the origin of the coordinate system (0,0). It allows us to avoid the bias caused by the interactant's initial position. This is used to calculate the *scaled joint position*. Moreover, for each video, we pick one frame when the interactant is facing the camera and note the distance between the two shoulders (*scale*). Then, the coordinate system is changed using a normalisation of x and y such as $scale = 1$.

Based on the extracted features and the voice activity detection records, we define several new variables that we used for our analysis.

- **Acoustic features:** The acoustic features are averaged along the 600ms following each VAD switch points.
- **Hand activity:** After scaling the joint position, the left and right hand activity are computed on the 600ms following each VAD switch points. They can be interpreted as the amount of motion of each hand and are estimated using:

$$v_{Hand}(i) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{1}$$

and:

$$Hand\_act = \sum_{i}^{i+N} v_{Hand}(i)/N \tag{2}$$

where x_i and y_i are the coordinates of the hand (right or left) at time-step i, N is the number of frame corresponding to 600ms. i is the instant of a particular VAD switch points.
Hand activities are normalized using z-scores to be invariant to the quantity of behaviors of interactants:

$$(Hand)\_act = \frac{(Hand)\_act - \mu}{\sigma} \tag{3}$$

Where μ and σ are, respectively, the mean and standard deviation of $v_{Hand}(i)$ along the whole sequence. Doing this for both hands leads to the two features $left\_hand\_act$ and $right\_hand\_act$.
- **Head activity:** similar to hand activity, head activity is estimated over the 600ms following each VAD switch points using:

$$v_{Head}(i) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2} \tag{4}$$

and:

$$Head\_act = \sum_{i}^{i+N} v_{Head}(i)/N \tag{5}$$

Head activities are then normalized in the same way than hand activities.

– **IPU length:** After annotating all the conversations, we apply a script to split voice activity into Inter-Pausal Unit (IPU). An Inter-Pausal Units is defined as a speech unit of a single speaker without pauses longer than 200ms [17]. The feature *IPU length* corresponds to the length of the IPU following each annotated VAD switch point.

7 Analysis

7.1 IPU length analysis

We first analyzed the length of the IPU following each annotated voice activity change in the level 1 annotation as illustrated on the left side of Figure 5.

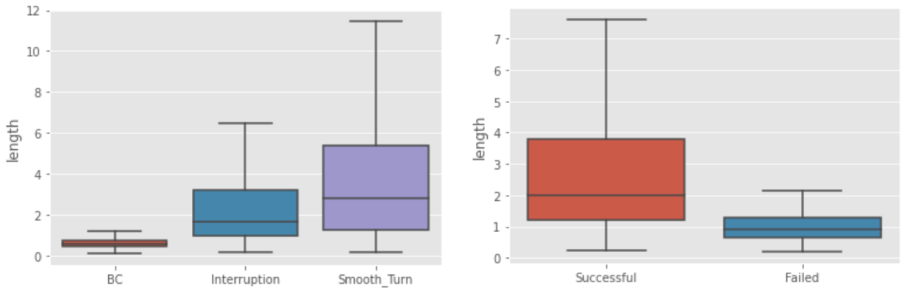


Fig. 5. Average value of IPUs or backchannels length for the level 1 (left) & the level 2 (right) annotation.

The IPU length following smooth turn exchanges are statistically longer than those following interruptions. Moreover, the length of backchannels is smaller than the IPU after an interruption or a smooth turn.

Considering the level 2 annotation, that is the accomplishment (success / failure) of interruptions, we can see on the right side of Figure 5 that IPUs that follow successful interruptions have longer duration (3.33 seconds on average) than those that follow failed interruptions (1.04 seconds on average).

For the level 3 annotation (left figure in Figure 6), IPUs that follow competitive interruptions are longer than than those that follow cooperative ones, but the difference is less significant between the different IPUs considered at the level 2.

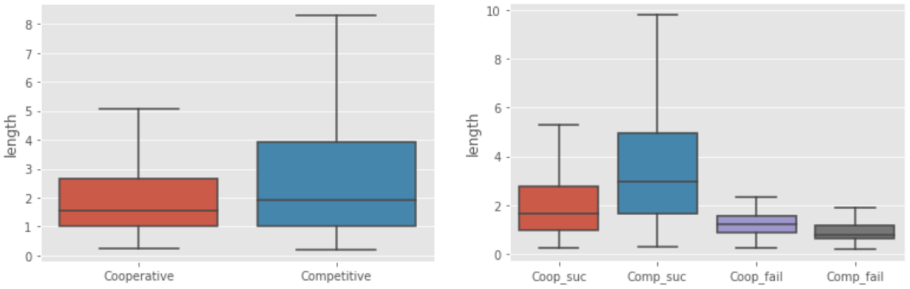


Fig. 6. Average value of IPUs length for the level 3 annotation (left) & combining the level 2 and 3 annotations (right).

When taking into account both accomplishment and interruption types (annotated respectively at level 2 and level 3), IPUs that follow successful competitive interruptions have longer duration than IPUs that follow successful cooperative ones. We do not find significant differences between IPUs that follow cooperative and competitive interruptions for the failed interruptions as illustrated on the right side of Figure 6.

7.2 Acoustic features analysis

We then analysed acoustic features (F0 and loudness) of the interrupter who initiates the interruption as illustrated in Figure 7

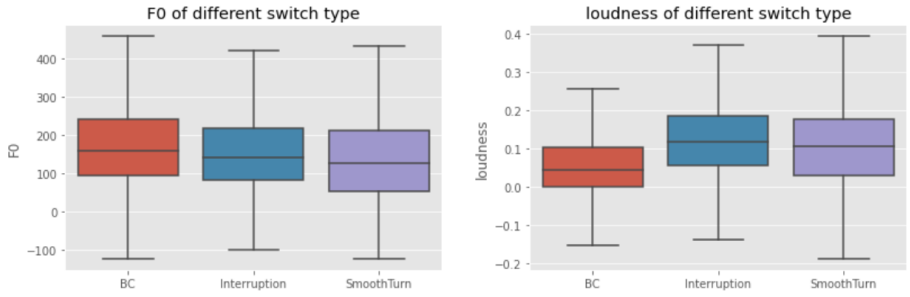


Fig. 7. Average value of F0 (left) & loudness (right) for the level 1 annotation.

No significant differences appear for F0 or loudness between interruptions and smooth turn exchanges. However, we note that backchannels have a lower loudness. We could not draw any conclusions by studying levels 2 and 3 of annotation.

7.3 Head and Hand activities analysis

For body motion, we only study the level 1 annotation (left side of Figure 8). No significant differences appear for head activity, except a small activity for backchannels.

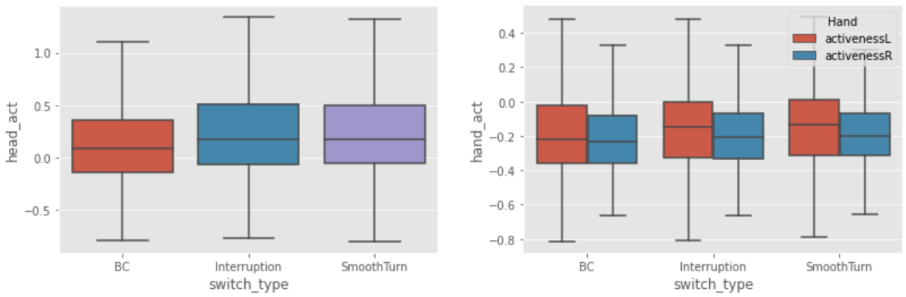


Fig. 8. Average value of head activity (left) & hand activity (right) for the first level of annotation.

When looking at the hands activity, we can see from the right side of Figure 8 that the left hand has larger spatial extent than the right hand. In contrast, there is no significant difference between backchannel, interruption and smooth turn.

8 Discussion

In our database, we detected 929 interruptions over 3983 VAD switches. Excluding backchannels, interruptions take almost 40% of the turn switches, which is a quite large number and shows the importance of interruptions in natural conversation.

In our corpus, the interruptions are successful most of the time. There are only 18.3% of the cases where the person does not succeed to take the turn.

Considering the acoustic and body features we analysed, some of them seem important to distinguish backchannel, interruption and smooth turn exchange such as the IPU length or loundness. We did not find significant differences for the other features. We could not replicate all the results from previous studies [40, 48, 26]. Such dissimilarities may come from the scenarios of the corpora used in the different studies. Other, which were supposed to be relevant, such as hand activity, did not show significant differences.

9 Conclusion and perspectives

Being interested by interruptions, we first propose a three levels schema annotation that allow characterizing each type of VAD switches. We used it to

annotated the NoXi corpus and found that interruptions are very frequent in natural interactions. Then, we extracted all the voice activity switches. We conducted analyses on acoustic and body movement features to computationally characterize these switches.

This work is a first step toward the modeling of interruptions for an embodied conversational agent. In the near future, we aim to endow ECA with the possibility to react to human's interruptions. It requires first to recognize if a speech overlap corresponds to a smooth turn exchange, a backchannel, or an interruption. Our next step is to introduce these features in a machine learning algorithm able to make this classification in real time. Once the agent knows if its human interlocutor interrupts, displays a backchannel or takes the turn slightly before the agent's speaking turn, the agent can plan how to respond to human's behavior.

10 Acknowledgements

This work was performed as part of ANR-JST-CREST TAPAS and ANR-JST-DFG PANORAMA project.

References

1. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of semantics **9**(1), 1–26 (1992)
2. Ball, P.: Listeners' responses to filled pauses in relation to floor apportionment. British Journal of Social & Clinical Psychology (1975)
3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018)
4. Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M.F., Schuller, B., André, E.: explainable cooperative machine learning with NOVA. KI - Künstliche Intelligenz (Jan 2020). https://doi.org/10.1007/s13218-020-00632-3, https://doi.org/10.1007/s13218-020-00632-3
5. Beattie, G.W.: Floor apportionment and gaze in conversational dyads. British journal of social and clinical psychology **17**(1), 7–15 (1978)
6. Beattie, G.W.: Interruption in conversational interaction, and its relation to the sex and status of the interactants. Walter de Gruyter, Berlin/New York Berlin, New York (1981)
7. Bögels, S., Torreira, F.: Turn-end estimation in conversational turn-taking: The roles of context and prosody. Discourse Processes **58**(10), 903–924 (2021)
8. Cafaro, A., Glas, N., Pelachaud, C.: The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions. In: Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems. pp. 911–920 (2016)
9. Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres Torres, M., Pelachaud, C., André, E., Valstar, M.: The noxi database: multimodal recordings of mediated novice-expert interactions. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 350–359 (2017)

10. Chowdhury, S.A., Danieli, M., Riccardi, G.: Annotating and categorizing competition in overlap speech. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5316–5320. IEEE (2015)
11. Chỳlek, A., Švec, J., Šmídl, L.: Learning to interrupt the user at the right time in incremental dialogue systems. In: International Conference on Text, Speech, and Dialogue. pp. 500–508. Springer (2018)
12. Coates, J.: 11 no gap, lots of overlap: Turn-taking patterns in. Researching language and literacy in social context: A reader p. 177 (1994)
13. Coman, A.C., Yoshino, K., Murase, Y., Nakamura, S., Riccardi, G.: An incremental turn-taking model for task-oriented dialog systems. arXiv preprint arXiv:1905.11806 (2019)
14. De Kok, I., Heylen, D.: Multimodal end-of-turn prediction in multi-party meetings. In: Proceedings of the 2009 international conference on Multimodal interfaces. pp. 91–98 (2009)
15. De Ruiter, J.P., Mitterer, H., Enfield, N.J.: Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. Language **82**(3), 515–535 (2006)
16. Dediu, D., Levinson, S.C.: On the antiquity of language: the reinterpretation of neandertal linguistic capacities and its consequences. Frontiers in psychology **4**, 397 (2013)
17. Demol, M., Verhelst, W., Verhoeve, P.: The duration of speech pauses in a multilingual environment. In: Eighth Annual Conference of the International Speech Communication Association (2007)
18. Duncan, S.: Some signals and rules for taking speaking turns in conversations. Journal of personality and social psychology **23**(2), 283 (1972)
19. Egorow, O., Wendemuth, A.: On emotions as features for speech overlaps classification. IEEE Transactions on Affective Computing (2019)
20. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978)
21. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1459–1462 (2010)
22. Ferguson, N.: Simultaneous speech, interruptions and dominance. British Journal of social and clinical Psychology **16**(4), 295–302 (1977)
23. French, P., Local, J.: Turn-competitive incomings. Journal of Pragmatics **7**(1), 17–38 (1983)
24. Goldberg, J.A.: Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts. Journal of Pragmatics **14**(6), 883–903 (1990)
25. Gravano, A., Brusco, P., Benus, S.: Who do you think will speak next? perception of turn-taking cues in slovak and argentine spanish. In: INTERSPEECH. pp. 1265–1269 (2016)
26. Gravano, A., Hirschberg, J.: A corpus-based study of interruptions in spoken dialogue. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
27. Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., Wedin, L.: Perceptual and acoustic correlates of abnormal voice qualities. Acta oto-laryngologica **90**(1-6), 441–451 (1980)
28. Hara, K., Inoue, K., Takanashi, K., Kawahara, T.: Turn-Taking Prediction Based on Detection of Transition Relevance Place. In: Proc. Interspeech 2019. pp. 4170–4174 (2019). https://doi.org/10.21437/Interspeech.2019-1537

29. Heldner, M., Edlund, J.: Pauses, gaps and overlaps in conversations. Journal of Phonetics **38**(4), 555–568 (2010)
30. Holler, J., Kendrick, K.H., Casillas, M., Levinson, S.C.: Turn-taking in human communicative interaction. Frontiers Media SA (2016)
31. Indefrey, P., Levelt, W.J.: The spatial and temporal signatures of word production components. Cognition **92**(1-2), 101–144 (2004)
32. Ishii, R., Otsuka, K., Kumano, S., Matsuda, M., Yamato, J.: Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In: Proceedings of the 15th ACM on International conference on multimodal interaction. pp. 79–86 (2013)
33. Ishii, R., Otsuka, K., Kumano, S., Yamato, J.: Using respiration to predict who will speak next and when in multiparty meetings. ACM Transactions on Interactive Intelligent Systems (TiiS) **6**(2), 1–20 (2016)
34. Ishii, R., Ren, X., Muszynski, M., Morency, L.P.: Can prediction of turn-management willingness improve turn-changing modeling? In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. pp. 1–8 (2020)
35. Ishii, R., Ren, X., Muszynski, M., Morency, L.P.: Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. pp. 131–138 (2021)
36. Ishimoto, Y., Teraoka, T., Enomoto, M.: End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous japanese speech. In: Interspeech. pp. 1681–1685 (2017)
37. Itakura, H.: Describing conversational dominance. Journal of Pragmatics **33**(12), 1859–1880 (2001)
38. Kendon, A.: Some functions of gaze-direction in social interaction. Acta psychologica **26**, 22–63 (1967)
39. Kurtić, E., Brown, G.J., Wells, B.: Resources for turn competition in overlapping talk. Speech Communication **55**(5), 721–743 (2013)
40. Lee, C.C., Lee, S., Narayanan, S.S.: An analysis of multimodal cues of interruption in dyadic spoken interactions. In: Ninth Annual Conference of the International Speech Communication Association (2008)
41. Lee, C.C., Narayanan, S.: Predicting interruptions in dyadic spoken interactions. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 5250–5253. IEEE (2010)
42. Maier, A., Hough, J., Schlangen, D., et al.: Towards deep end-of-turn prediction for situated spoken dialogue systems (2017)
43. Moerman, M., Sacks, H.: Appendix b. on "understanding" in the analysis of natural conversation. In: Talking Culture, pp. 180–186. University of Pennsylvania Press (2010)
44. Niebuhr, O., Görs, K., Graupe, E.: Speech reduction, intensity, and f0 shape are cues to turn-taking. In: Proceedings of the SIGDIAL 2013 Conference. pp. 261–269 (2013)
45. Riest, C., Jorschick, A.B., de Ruiter, J.P.: Anticipation in turn-taking: mechanisms and information sources. Frontiers in Psychology **6**, 89 (2015)
46. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn taking for conversation. In: Studies in the organization of conversational interaction, pp. 7–55. Elsevier (1978)
47. Schegloff, E.A.: Sequencing in conversational openings 1. American anthropologist **70**(6), 1075–1095 (1968)

48. Schegloff, E.A.: Overlapping talk and the organization of turn-taking for conversation. Language in society **29**(1), 1–63 (2000)
49. Schegloff, E.A., Sacks, H.: Opening up closings. Walter de Gruyter, Berlin/New York Berlin, New York (1973)
50. Shriberg, E., Stolcke, A., Baron, D.: Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In: Seventh European Conference on Speech Communication and Technology (2001)
51. Skantze, G., Johansson, M., Beskow, J.: Exploring turn-taking cues in multi-party human-robot discussions about objects. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 67–74 (2015)
52. Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.E., et al.: Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences **106**(26), 10587–10592 (2009)
53. Tannen, D., et al.: You just don't understand: Women and men in conversation. Virago London (1991)
54. Truong, K.P.: Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee. In: Interspeech. pp. 1404–1408 (2013)
55. Van Berkum, J.J., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P.: Anticipating upcoming words in discourse: evidence from erps and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition **31**(3), 443 (2005)
56. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose Flow: Efficient online pose tracking. In: BMVC (2018)
57. Yang, L.c.: Visualizing spoken discourse: Prosodic form and discourse functions of interruptions. In: Proceedings of the Second SIGdial Workshop on Discourse and Dialogue (2001)