Multimodal classification of interruptions in humans' interaction

Liu Yang, Catherine Achard, Catherine Pelachaud

▶ **To cite this version:**

Liu Yang, Catherine Achard, Catherine Pelachaud. Multimodal classification of interruptions in humans' interaction. ACM International Conference on Multimodal Interaction ICMI, 2022, Bangalore, India. 10.1145/3536221.3556604 . hal-03845959v2

HAL Id: hal-03845959
https://hal.science/hal-03845959v2

Submitted on 4 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimodal classification of interruptions in humans' interaction

LIU YANG, Institut des Systèmes Intelligents et de Robotique, CNRS, Sorbonne University, France
CATHERINE ACHARD, Institut des Systèmes Intelligents et de Robotique, CNRS, Sorbonne University, France
CATHERINE PELACHAUD, CNRS, Institut des Systèmes Intelligents et de Robotique, Sorbonne University, France

During an interaction interruptions occur frequently. Interruptions may arise to fulfill different goals such as changing the topic of conversation abruptly, asking for clarification, completing the current speaker's turn. Interruptions may be cooperative or competitive depending on the interrupter's intention. Our main goal is to endow a Socially Interactive Agent with the capacity to handle user interruptions in dyadic interaction. It requires the agent to detect an interruption and recognize its type (cooperative/competitive), and then to plan its behaviours to respond appropriately. As a first step towards this goal, we developed a multimodal classification model using acoustic features, facial expression, head movement, and gaze direction from both, the interrupter and the interruptee. The classification model learns from the sequential information to automatically identify interruptions type. We also present studies we conducted to measure the shortest delay needed (0.6s) for our classification model to identify interruption types with a high classification accuracy (81%). On average, most interruption overlaps last longer than 0.6s, so a Socially Interactive Agent has time to detect and recognize an interruption type and can respond in a timely manner to its human interlocutor's interruption.

CCS Concepts: • **Human-centered computing** → **Human agent interaction (HAI)**.

Additional Key Words and Phrases: Nonverbal Behaviour, Interruption Classification, multimodality, Socially Interactive Agent

ACM Reference Format:
Liu YANG, Catherine ACHARD, and Catherine PELACHAUD. 2022. Multimodal classification of interruptions in humans' interaction. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3536221.3556604

1 INTRODUCTION

Speaking turn exchange is very well organized. People exchange speaking turns smoothly and cooperate to avoid overlaps since it's difficult to speak and listen at the same time. One of the early conversation analysis studies suggested that overlaps rarely occur during turn exchanges [27]. Interlocutors do not interrupt each other; the listener waits for the turn or sends signals to indicate the intention to take the turn. However, recent researches show that overlaps and interruptions are common phenomena in conversation. Helder *et al.* [16] stated that in dyadic conversations, over 40% of all turn shifts contain overlaps, and Shriberg*et al.* [31] observed that over 30% of turn exchanges contain overlaps in multi-party conversations. Before talking about interruption, we must distinguish the difference between interruption and overlap. During interruptions, the listener intends to take the floor of the speaker when the current utterance is not finished, against the speaker's will [29]. While overlap corresponds to the period when both interlocutors talk at the same time. Not all overlaps correspond to interruptions; they may correspond to the production of a slight overlay at the end of an utterance. Some of the overlaps can be identified as interruptions, which, play an important role during the conversation. For example, over 300 interruptions are identified during conversations of a total duration of about

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
Manuscript submitted to ACM

4 hours [21]. Interruptions can be interpreted as evidence of power and dominance as they disrupt the speech flow [11, 12, 34], and might be used to change the topic of conversation or to indicate the interlocutor knows already what the speaker is going to say. Such interruptions are referred as competitive interruptions [22]. However, depending on the conversation context, interruptions may also be cooperative, for example by providing support to the speaker to maintain the conversation[37], by completing the words the speaker does not recall, etc. Interruptions have therefore diverse functions and carry the underlying communicative intention, attitude and motivation of the interrupter [8, 35].

Therefore, to ensure simulating natural interaction between human and virtual agent, interruptions should be taken care of. Human user may interrupt the virtual agent during the interaction. The agent dialog system should be able to respond when human user interrupts to take the turn. To maintain the conversation, it should know whether to yield or keep the speaking turn, and how to do so. In order to plan an appropriate reaction, Reidsma et al. [26] argued that it is important to distinguish between competitive and cooperative interruptions, and thus to have a so-called "continuous conversation" during the human-agent interaction.

Conversely, interruptions can grab the interlocuter's attention and show a willingness to communicate, thus maintaining engagement in the interaction [23]. To have the human user engaged during the interaction, the dialog system should also be able to generate interruptions in a timely manner [14]. An agent that never interrupts may be perceived as less interactive, while interrupting too often may be perceived as much too aggressive [25].

In this paper, we focus on classifying interruption types. We consider only the interruption initiated by others and not self-interruption (which is initiated by the speaker-self). We present a new automatical classification approach based on multimodal features of both interrupter and interruptee. We also compare the performance of different classification models using different combinations of modalities. Finally, as we want to classify interruption types in real-time during human-agent interaction, we conduct classification performance tests by changing the time window length in which the classification happens after the beginning of an interruption. Only overlapped interruptions are considered in this work.

An overview of previous research on interruption is given in Section 2. Sections 3 and 4 describe respectively the corpus we used for this work and the feature we consider. We present different classification models in Section 5 and their results and comparison in Section 6. Finally, Section 7 concludes this work.

2 RELATED WORKS

Existing works on interruption have addressed the issue of detecting when an interruption occurs and recognizing its type.

Lee *et al.* [20] suggested that the presence of interruptions is not random, they follow certain rules and their occurrence can be predicted using contextual cues. This is also supported by the research of Shriberg *et al.* [30] that used prosodic features to predict overlap start point. Previously, Goldberg and colleagues [13] found that the location of an interruption may provide information to differentiate interruption type. Apart from the start point location, the duration of overlaps plays an important role in differentiating between cooperative and competitive overlaps or interruptions. The results in [17, 18] indicate that competitive overlaps last longer than cooperative ones.

Besides, more and more researches show evidence that cooperative and competitive interruptions also show differences in prosodic features. From the analysis results, Yang [37] argued that competitive interruptions have higher pitch and intensity than cooperative interruptions. Shriberg [30] and Hammarberg *et al.* [15] found that people raise their voice energy and pitch to interrupt the current speaker. Schegloff *et al.* [28] argued that the speaker uses prosodic variation and repetition to indicate strong turn competitiveness.

Several works have proposed models to automatically classify interruption types using multimodal features. Lee *et al.* [19] analysed hand motion activity, speech intensity and disfluency for competitive/cooperative interruption in spoken dyadic conversations. Hand motion activity and speech intensity are reported as reliable features differentiating interruption types. Compared to the multimodal classification model, of which the accuracy achieved 71.2%, using only one modality leads to significant lower classification performance.

Truong *et al.* [33] proposed to classify overlaps using a SVM model. As input, the authors used low-level signals such as acoustic features (F0, intensity and voice quality), combined with high-level annotations, namely focus of attention (gaze direction) and communicative function of head movements. With a sequence length of 0.6s after the overlap onset point, the SVM model achieved good performance with an EER of 32.1%. The authors also mentioned that gaze information slightly improves the accuracy, while adding acoustic information from the overlappee didn't. Chowdhury et al. [5] proposed a Sequential Minimal Optimization (SMO) model for competitive/cooperative overlaps classification. Prosody, voice quality, MFCC, energy and spectral features are used as inputs. With an optimal subset of selected features, the model achieved an F1-score of 0.69. Later on, the authors [6, 7] presented to classify cooperative/competitive overlaps with acoustic (prosodic, spectral, voice quality, MFCC and energy) and lexical features. Different models with feature combinations were tested during the experiment, the best performance (F1-score of 0.70) is achieved with a feed-forward neural network (FFNN), using both acoustic and lexical features. Combined with acoustic features, Egorow et al. [9] then took two emotion dimensions (*control* and *valence*) into account to classify overlaps on a telephone-based human conversation corpus. The SVM model's best performance was reported with an F1-score of 0.74.

From previous research, although hand activity, body motion, gaze and head gesture improve the classification accuracy, acoustic features have been found to be the key features in classifying interruptions and overlaps. However, previous studies do not take facial expressions into account. According to the experiment results of [24], conversational facial expression is associated with the dialog context and operates as nonverbal interjections, which should be able to provide additional information for the interruption classification model.

The importance of sequential information during interruptions and overlaps is also ignored in previous works and only statistical projections are provided as input to classification models. However, the analysis result of [33] shows the difference between two types in the temporal curve of acoustic features, that have been used to classify the overlap type. We thus propose a new method to classify the interruptions with acoustic profile, head activity, gaze behaviour, and facial expression. We also take the sequential information into account. As our goal is to implement an online model that classifies the interruption type as fast as possible during human-agent interaction, we also investigate how classification performance varies with the length of the time window used after the point of interruption.

3 CORPUS

For this work, we use two different corpora that are freely available: the AMI corpus [3] and the NoXi corpus [2]. Both corpora differ in their setting (group of four interlocutors vs dyads), conversation context (task focus vs more social focus), and language (English vs French). Using different corpora allows us to test the genericity of our approach.

3.1 The AMI Meeting Corpus

AMI [3] is a multimodal database that contains 100h free multi-party English meetings. In the AMI corpus, 4 participants are asked to discuss the given topic. Each interaction lasts about 25mn. The video of the participants shows their upper body. For our work, we focus on 9 meetings (for about 4h in total) of as they have been annotated for *(head function and focus of attention).*

Following the schema described in [36], we annotate the interruptions in these meetings into different types. The schema also presents the definition of successful interruption (if the interruptee's flow was disrupted) and failed interruption (if the interruption was rejected). We don't distinguish the interruption's accomplishment for this model, but only the types below:

Cooperative interruptions: include agreement, assistance and clarification, for both successful and failed interruptions.

Competitive interruptions: include disagreement, topic change, floor taking and tangentialization, for both successful and failed interruptions.

Not identified type includes short failed interruptions with not enough information to determine its type.

The annotation is performed twice by the same annotator according to the approach presented in [4]. There is a one-month interval between the two rounds of annotation. Cohen's Kappa measures were computed across the two annotations and were found to be mostly satisfactory ($k = 0.821$). Finally, 508 interruptions were annotated in the AMI corpus (230 cooperatives, 278 competitives).

3.2 The NoXi Corpus

NoXi is a multimodal corpus that contains free dyadic conversations. We also use the French part of the NoXi corpus to evaluate our model[2], including 21 conversations, for about 7h in total. Each conversation lasts around 20mn and both interactants are recorded separately. The video shows almost their full body except for the feet. All videos are synchronized and transcribed. We use the same annotation schema and double annotation method to annotate interruptions as for the AMI corpus. 929 interruptions are annotated (505 cooperative, 348 competitive) with $k = 0.84$.

4 FEATURE EXTRACTION

We propose a new approach to classify interruption using multimodal signal. To measure the shortest delay needed for classification, we first define a time window around the interruption onset point and then extract the multimodal features in this time window. We use acoustic features as well as the visual features of facial expression (eyebrow movement), head motion activity and gaze direction. Each modality consists of two types of features, *local* and *global*. The *local* features are the values extracted on every frame of the selected time window (15 frames per second). The *global* features are the statistical projection of each *local* feature over the defined time window. For the AMI corpus, we use two more provided features: the annotation of communicative functions of head movements and the occurrence of mutual gaze, which do not exist in NoXi.

4.1 Segmentation

To determine the interruption window, we analyze the interruption overlap duration for the annotated interruptions of both corpora. The statistical analysis result is shown in Figure 1. The average duration of cooperative interruption is 1.11s and of competitive interruption 1.49s.

For each interruption instance, as shown in Figure 2, most interruption overlaps last longer than 0.6s. We note its starting point as t_0, in order to classify the interruption before the end of overlap, we should choose a window size less than 0.6s. Truong and colleagues [33] suggest an interruption window length of 0.6s. We choose to use the same window size and then study the temporal segment between $t_0 - 0.6$s and $t_0 + 0.6$s that we called interruption windows in the following. This choice is also driven by our main aim which is to endow a virtual agent with the capacity to

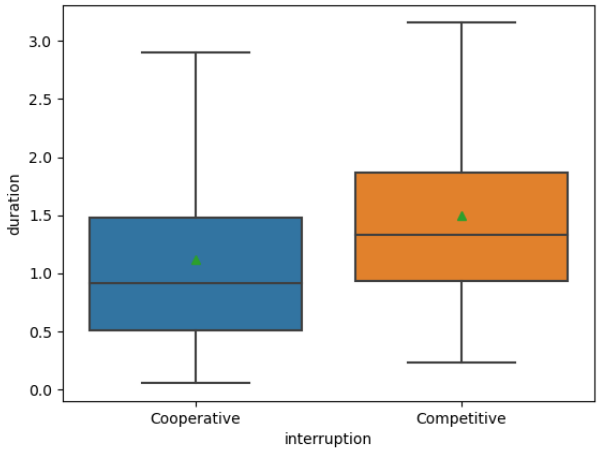

Fig. 1. Interruption overlap duration.

handle interruptions and respond to them as quickly as possible. It is important the agent classifies it before the overlap ends. During the first period, before the interruption occurs, from $t_0 - 0.6$s to t_0, only the overlappee is speaking while in the second period, both interlocutors are speaking, at least partly.

The classification of interruptions is based on the multimodal features that are extracted for both interrupter and interruptee over the interruption window. Depending on the method we employ for the classification model (Section 6.2), we extract *global* features over the interruption window, or, *local* features estimated at each time step of the window.

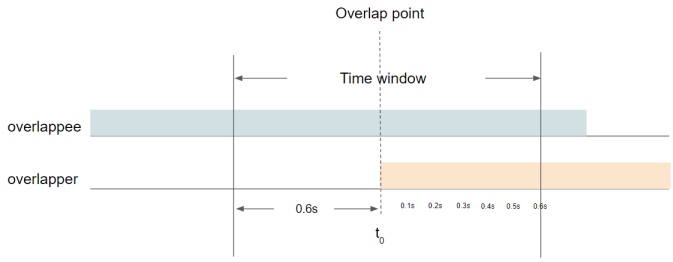

Fig. 2. Segmentation of features.

4.2 Audio

4.2.1 Local acoustic features. The 33 acoustic features we considered are composed of pitch (Fundamental frequency F0, F0-envelope), loudness, voice- probability, jitter, shimmer, logarithmic harmonics-to-noise ratio (logHNR), Mel-frequency cepstral coefficients (MFCC 0-12), Logarithmic signal energy from pcm frames, Energy in spectral bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), roll-off points (25%, 50%, 70%, 90%), centroid, flux, max-position and min-position

as proposed in [7]. These features are extracted with openSMILE [10] with a frequency of 100 frames per second. The extracted audio features are then resampled into 15 frames per second to fit the frequency of visual features (eyebrow movement, head motion activity and gaze direction.).

All the extracted features are normalized by z-scores, using the mean and standard deviation of the interruption windows around all the interruptions. We obtain values for the acoustic features for each time window for both, the interrupter and the interruptee, and store them in a vector. We refer to this vector as the local acoustic features vector for a given time window around an interruption. The dimension of the local acoustic features vector A_l is then $(33 * 2) * 18$ (33 features for each interlocutor over 1.2s that lasts 18 frames).

When the interlocuter is not speaking, all her/his acoustic features are set to zero to avoid the impact of noises.

4.2.2 Global acoustic features. For each feature mentioned previously, we extract their statistical projection over the interruption windows as proposed by [7]. They are composed of values such as min and max position, range, linear and quadratic regression coefficients and approximation errors, variance, standard deviation, skewness, peaks, mean peak distance and mean peak.

These statistical projections are normalized by z-scores, using the mean and standard deviation of all windows. They composed the global acoustic features vector A_g of size 792 (33*2*12, the value of the 12 statistical projections for the 33 features for each interlocutor).

4.3 Eyebrow

4.3.1 Local eyebrow features. The facial expressions are extracted with Openface [1] and are encoded into Action Units from the Facial Action Coding System (FACS). We focus on the three action units AU01, AU02 and AU04 representing eyebrow movements to avoid the impact of mouth movement caused by the speech content. Eyebrow features are extracted with a frequency of 15 frames per second and are z-scores normalized on all the interruption windows to compose a feature vector E_l of size $(3 * 2) * 18$ (3 AUs for both interlocutors over 1.2 sec (18 frames)).

4.3.2 Global eyebrow features. Statistical projections of each eyebrow feature are estimated as done for the acoustic features. They are normalized by z-scores. They composed a vector E_g of size 48 (The value of the 8 statistical projections for the 3 features for each interlocutor)

4.4 Head activity

4.4.1 Local head activity features. For head movement features, we use the head position (in *x*-*y*-*z* axis) extracted by Openface at the frequency of 15 frames per second. To avoid the bias caused by the interactant's initial position, we do not use absolute position but the head motion activity using the following equation:

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

These features are normalized by z-scores. For the AMI corpus, we also use the head functions annotations that are provided: *concord*, *discord*, *deixis*, *emphasis*, *negative*, *turn*, and all *other* communicative head gestures. This annotation gives the communicative function of the head movements with timestamps of start and end timing. They are encoded using a one-hot encoding, binary values indicating the absence or presence of each event.

The final vector H_l is of size $(10 * 2) * 18$ (head activity + head function one-hot encoding for both interlocutors over 1.2 sec (18 frames)).

4.4.2 Global head movement features. The statistical projections of head activity are calculated and Z-score normalized on the interruption window. For the head function annotation mentioned above, we use a one-hot encoding by setting the value 1 if the event is present at least at one-time step of the interruption window.

The Global head movement vector is of size 34 (the value of the 8 statistical projections for 1 feature + the vectors one-hot encoding for each interlocutor).

4.5 Gaze

4.5.1 Local gaze features. The gaze direction (in *x-y* axes) is extracted by Openface with a frequency of 15 frames per second. Then they are z-scores normalized over all the interruption windows.

We also use the *focus of attention* annotations provided in the AMI corpus. A binary value is computed to indicate the presence of mutual attention (when the interrupter and interruptee are looking at each other) or its absence (when the interrupter or interruptee is looking somewhere else).

The final vector G_l is of size $(4 * 2) * 18$ (4 gaze features for both interlocutors over 1.2 sec (18frames)).

4.5.2 Global gaze features. Statistical projections of gaze direction are estimated as done for acoustic features. They are also normalized by z-scores. For the *focus of attention* annotation we use a binary value for mutual attention as presented in the local gaze features. The Global gaze vector is of size 34 (the value of the 8 statistical projections for 2 gaze direction features+ the vectors one-hot encoding for each interlocutor).

5 THE PROPOSED MODEL

Contrary to previous approaches [6, 7, 33], we do not extract manually global features but leave this task to a neural network which is more adapted to the problem. Moreover, using temporal series allows us to consider the real-time application as shown in the evaluation section.

Different from standard feed-forward neural networks, LSTM has connections that allow us to process not only single data points but also data sequences. An LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The three gates regulate the flow of information in and out of the cell. The long short-term memory (LSTM) architecture we used is depicted in Figure 3. The input at each time step x_t is composed by the concatenation of all local features. We only output the hidden state of the last time step, it is then passed to a dense layer that predicts the final classification. Although this model is very simple, it leads to the best results. Adding more complex architecture leads to an important over-fitting.

6 RESULT & DISCUSSION

In this section, we present the experimental classification results we obtained using different modalities and compare them to the state of the art. Then, we introduce the study we conducted on the time window length to be considered.

6.1 Results

Training Procedure: The method has been implemented and tested with TensorFlow, we split the data into train (70%), validation (10%) and test (20%) datasets. The input of LSTM layer (with a dropout of 0.2) is composed of the concatenation of all local features mentioned above over the interruption window of 1.2s (18 frames), making a vector of size 100*18. The latent vector, of dimension 10, is passed to a dense layer of dimension 8, before the output layer of dimension 1 that has a sigmoid as an activation function.

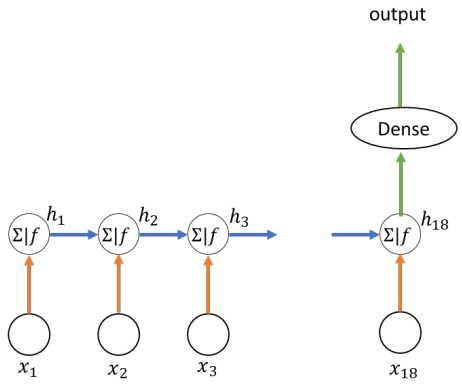

Fig. 3. The long short term memory (LSTM) architecture.

Table 1. Accuracy and F1 measure for FFNN, SVM and LSTM model with different combinations of modalities for the AMI corpus.

	Audio		Facial, head, gaze		All modalities	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN with modalities of [7]	0.74	0.73	-	-	-	-
FFNN with our modalities	0.74	0.73	0.69	0.67	0.79	0.78
SVM with modalities of [33]	0.69	0.66	0.65	0.61	0.72	0.72
SVM with our modalities	0.72	0.72	0.68	0.67	0.77	0.76
LSTM with our modalities	0.75	0.73	0.69	0.68	**0.81**	**0.80**

Table 2. Ablation study of FFNN, SVM and LSTM model for the AMI corpus with our modalities.

	All modalities		Audio, Facial, Head		Audio, Facial, Gaze		Audio, Head, Gaze	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN	0.79	0.78	0.75	0.75	0.77	0.77	0.77	0.76
SVM	0.77	0.76	0.74	0.73	0.75	0.74	0.75	0.75
LSTM	0.81	0.80	0.76	0.76	0.79	0.78	0.79	0.79

In all our experiments, the model is trained with mini-batches of 64 interruptions using an Adam optimizer and a learning rate fixed to 1e-5.

Results are presented in Table 1 for the AMI corpus. As in previous studies, acoustic features alone lead to good results. Adding Facial expressions, head movement and gaze improve the performances by reaching an accuracy of 81% and an F1-score of 0.80. Compared to facial expression, gaze and head activity, it is clear that acoustic features provide the main information for classification. Ablation studies are presented in Table 2. In these experiments, in order to study the contribution of each modality to the classification accuracy, we remove all features related to one modality at a time. We also perform ablation experiments for each feature, but the influence of a single feature on the results is too small. We do not present the results here. According to the results of the three models, after removing gaze-related features, the accuracy is significantly lower than that of all modalities, and the decrease is even larger than when the other two modalities are removed.

Table 3. Accuracy and F1 measure for FFNN, SVM and LSTM model with different combinations of modalities for the NoXi corpus.

	Audio		Facial, head, gaze		All modalities	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN with modalities of [7]	0.63	0.61	-	-	-	-
FFNN with our modalities	0.63	0.61	0.61	0.60	0.66	0.64
SVM with modalities of [33]	0.57	0.52	-	-	-	-
SVM with our modalities	0.61	0.57	0.59	0.57	0.62	0.60
LSTM with our modalities	0.65	0.62	0.62	0.60	**0.69**	**0.65**

Table 4. Ablation study of FFNN, SVM and LSTM model for the NoXi corpus with our modalities.

	All modalities		Audio, Facial, Head		Audio, Facial, Gaze		Audio, Head, Gaze	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN	0.66	0.64	0.65	0,65	0.65	0.64	0.64	0.63
SVM	0.62	0.60	0.61	0.60	0.60	0.58	0.61	0.61
LSTM	0.69	0.65	0.68	0,66	0.66	0.65	0.68	0.67

We compare our result with random accuracy, which is calculated with the equation [32]:

$$\begin{aligned} accuracy &= P(class = 0) * P(prediction = 0) + P(class = 1) * P(prediction = 1) \\ &= (348/(505 + 348))^2 + (505/(505 + 348))^2 \\ &= 0.5162 \approx 0.52 \end{aligned}$$

Where, in our case, class 0 represents competitive interruption and 1 represents cooperative, therefore $P(class = 0) = P(prediction = 0) = 348/(505 + 348) = 0.41$, and $P(class = 1) = P(prediction = 1) = 505/(505 + 348) = 0.59$.

Results of our models are presented in Table 3 for the NoXi Corpus. The best results are also obtained with the multimodal features, but are only slightly above random (0.52). Compared to facial expression, gaze and head activity, acoustic features provide the main information for classification. Ablation studies are presented in Table 4. Same as for the AMI corpus, we remove all features related to one modality at a time. From the results, we do not observe significant differences for all three modalities. Except for language differences (French for NoXi and English for AMI), NoXi is composed of dyadic screen-mediated conversations while AMI is composed of four-parties face-to-face conversations. Even if multi-party interruptions are not considered, dyadic interruptions occur more frequently in the AMI than in the NoXi databases. The AMI scenario involves 4 interlocutors brainstorming to design a product. The design task may imply trying to impose one's ideas on others giving rise to more competitive interruptions. Being more participants could also explain the increase in the number and frequency of interruptions. In the NoXi database, interlocutors are asked to play either the role of an expert that shares information and knowledge on a given topic, or of a novice that is interested in learning about a given topic. The expert tends to talk much more than the novice during the interaction, most interruptions are initiated by the novice to express their opinion or to ask for information, which is not so competitive. These two scenarios are different and lead to more or less dynamic interactions, involving interruptions with different degrees of competitiveness. The differences in settings and scenarios could explain the differences in the results we obtain for both corpora. Missing several high-level annotation features for the NoXi corpus could also be another reason.

6.2 Comparative study

We compare our method with previous ones that have been implemented to classify interruptions. Truong's SVM model is trained using the same features as presented in [33] and using our features. Chowdhury's FFNN model is trained with the acoustic features mentioned in [7] and using our global multimodal features that have been automatically extracted from the local ones as presented in Section 4 for a fair comparison. Both methods are based on a global features vector that is manually designed, contrary to our method where a neural network learns the more adapted features.

Results presented in Table 1 for the AMI database and in Table 3 for the NoXi database, show that, for all models, acoustic features provide more information for the classification task compared to facial, head and gaze features. Actually, the accuracy obtained with all models is not so good when classifying only with facial expression, head activity and gaze direction. The proposed multimodal features could improve the results for all models (5% for FFNN model, 6% for LSTM model and 3% for SVM model). We can thus conclude that important information is also carried by facial expressions and head activities to classify interruptions. Our multimodal LSTM classification model leads to the best results, which proves the importance of sequential information from the time series and the model is able to extract more relevant features and thus to recognize interruption in real-time as presented in section 6.4, since no ending information is requested.

6.3 Interruption Window length

In order to study the influence of the interruption window length on the classification accuracy, we fix the window length before the onset point of interruption (0.6s) and vary the window length after the onset point from 0s to 1s for every 0.2s. We conduct such a study to endow a virtual agent with the capacity to react as quickly as possible and in a more appropriate manner. It requires the agent to know the interruption's type in the shortest delay when it occurs. Results presented in Figure 4 show that the accuracy increases with the length of the interruption window using our LSTM model. However, some cooperative interruptions present a small overlap duration as illustrated in Figure 1, and longer interruption window length implies the virtual agent's reaction will come later. To compromise between accuracy and reaction time, the *a priory choice* of 0.6s after the onset point of interruption seems to be a good choice. Another idea, possibly using temporal series as proposed with the LSTM model, is to adapt this temporal length during real-time application as proposed just below.

6.4 Adaptation of the temporal length during real-time application

The ultimate goal of our model is to classify interruptions in real-time in human-agent interaction, the agent responding to the interruptions of human users. Thus, as soon as an interruption is detected (beginning of the overlap), we start the inference using the multimodal signals in the temporal window $\{t_0 - 0.6\text{s}, t_0\text{s}\}$. Running our LSTM model on this temporal window allows us to initialize the latent vector introduced in the architecture as presented in Figure 3. An advantage of LSTM is that we can update this latent vector frame by frame, starting from t_0, and predict at each time the cooperative or competitive interruption probabilities (the result of the sigmoid layer is considered as probabilities in this part, even if it is not theoretically founded). A threshold can then be settled to perform or not the classification at each frame. As some samples of interruptions are easier to classify than others and as the overlap length can last more or less time, using LSTM allows us to make the classification more or less quickly, according to the studied sample and thus, to adapt the reaction time. To study the feasibility of this idea, for each interruption, we study the classification results starting from t_0 and adding new incoming data, time step by time step until 1.2s (18 frames) after t_0. Let us note

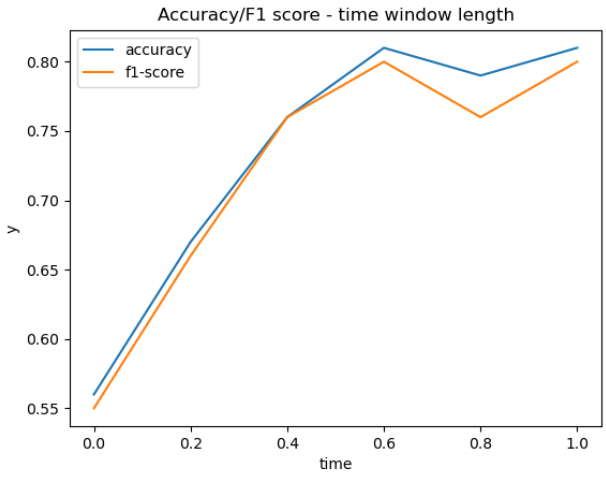

Fig. 4. Accuracy & Macro F1-score with different interruption window lengths.

Table 5. Mean accuracy & mean reaction time with different thresholds

Threshold	Mean accuracy	Mean reaction time length (frame and second)
0.5	0.66	0 frame (0s)
0.6	0.72	1 frame (0.07s)
0.7	0.76	4 frames (0.27s)
0.8	0.83	8 frames (0.53s)
0.9	0.87	12 frames (0.8s)

y, the output of the sigmoid at the studied frame. The classification is performed at this frame if $max(y, 1-y)$ reaches the threshold. If not, the following frame is considered. Naturally, the classification is also done, independently of the threshold, if the overlap stops or if the maximal interruption window length of 1.2s after t_0 is reached. Using a smaller threshold should thus lead to a smaller reaction time but perhaps to lower accuracy.

We plot in Figure 5 the percentage of interruptions classified according to the number of frames after t_0 for different thresholds. As expected, using a threshold of 0.5 leads to classifying all interruptions at $t0$. Increasing the threshold makes the classification time, and therefore the reaction time, longer. Fewer interruptions can be classified if the threshold is too high since it is hard to reach such a classification score: only about 50% of interruptions are classified after t_0+14 frames (around 1s) when the threshold is settled to 0.9. On the opposite, over 95% of interruptions are classified after 0.4s (6 frames) when the threshold id settled to 0.6 while the 5% of remaining ones will take a longer time. Independently of the threshold, LSTM allows us to adapt the reaction time according to the classification difficulty. But proceeding quickly with poor classification results is not acceptable as well. We thus present in Table 5 the mean accuracy and the mean reaction time according to the threshold values. As expected, the higher the threshold, the better the accuracy and the longer the response time. As our objective is to classify interruptions in real-time interaction to generate an appropriate response, the model is requested to classify the interruption as soon as possible with acceptable accuracy. We find from Figure 5 and Table5 that a threshold of 0.8 seems to be a good choice.

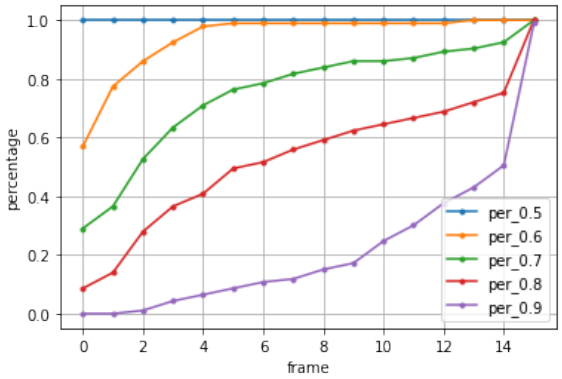

Fig. 5. Percentage of classified interruptions according to the number of frames after the beginning of the overlap, for different thresholds.

7 CONCLUSION

In this study, we proposed to classify cooperative and competitive interruptions in conversation with an LSTM model and evaluated different existing models. We experimented with different combinations of modalities and achieved our best performance using the acoustic profiles, facial expression, head movement and gaze features from both interrupter and interruptee. The experiments indicate that our LSTM model is able to learn accurate information from sequential series and improves classification performance. Tested with different interruption window lengths, the designed LSTM model is able to classify the interruption 0.6s after its start point with an accuracy of 81%.

Another advantage of LSTM is to make the classification more or less faster, according to the difficulty of the interruption. Using a classification threshold fixed to 0.8 allows us to classify interruption within 0.53s on average, with an accuracy of 83%. 40% of interruptions will be classified in less than 0.26s while the 30% of the most difficult examples will take more than 0.8s to be classified. In future work, we are going to explore more non-verbal behaviours to improve our LSTM model and implement our classification model in a virtual agent dialog system to classify interruptions in real-time with video and audio pre-processing. Then we will develop a non-verbal agent behaviour generation model to respond to the human user's interruption.

ACKNOWLEDGMENTS

This work was performed as a part of IA ANR-DFG-JST Panorama and ANR-JST-CREST TAPAS (19-JSTS-0001-01) project.

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. (2018), 59–66.
[2] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. (2017), 350–359.
[3] Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation* 41, 2 (2007), 181–190.
[4] Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud. 2019. A Methodology for the Automatic Extraction and Generation of Non-Verbal Signals Sequences Conveying Interpersonal Attitudes. *IEEE Trans. Affect. Comput.* 10, 4 (2019), 585–598.
[5] Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi. 2015. Annotating and categorizing competition in overlap speech. (2015), 5316–5320.

[6] Shammur Absar Chowdhury and Giuseppe Riccardi. 2017. A Deep Learning approach to modeling competitiveness in spoken conversations. (2017), 5680–5684.

[7] Shammur Absar Chowdhury, Evgeny A Stepanov, Morena Danieli, and Giuseppe Riccardi. 2019. Automatic classification of speech overlaps: feature representation and algorithms. *Computer Speech & Language* 55 (2019), 145–167.

[8] Shammur Absar Chowdhury, Evgeny A Stepanov, Giuseppe Riccardi, et al. 2016. Predicting User Satisfaction from Turn-Taking in Spoken Conversations. (2016), 2910–2914.

[9] Olga Egorow and Andreas Wendemuth. 2019. On Emotions as Features for Speech Overlaps Classification. *IEEE Transactions on Affective Computing* (2019).

[10] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. (2013), 835–838.

[11] Nicola Ferguson. 1977. Simultaneous speech, interruptions and dominance. *British Journal of social and clinical Psychology* 16, 4 (1977), 295–302.

[12] Peter French and John Local. 1983. Turn-competitive incomings. *Journal of Pragmatics* 7, 1 (1983), 17–38.

[13] Julia A Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts. *Journal of Pragmatics* 14, 6 (1990), 883–903.

[14] Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25, 3 (2011), 601–634.

[15] Britta Hammarberg, Bernard Fritzell, J Gaufin, Johan Sundberg, and Lage Wedin. 1980. Perceptual and acoustic correlates of abnormal voice qualities. *Acta oto-laryngologica* 90, 1-6 (1980), 441–451.

[16] Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38, 4 (2010).

[17] Gail Jefferson. 2004. A sketch of some orderly aspects of overlap in natural conversation. *Pragmatics and beyond new series* 125 (2004), 43–62.

[18] Emina Kurtic, Guy J Brown, and Bill Wells. 2010. Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration. (2010), 2550–2553.

[19] Chi-Chun Lee, Sungbok Lee, and Shrikanth S Narayanan. 2008. An analysis of multimodal cues of interruption in dyadic spoken interactions. (2008).

[20] Chi-Chun Lee and Shrikanth Narayanan. 2010. Predicting interruptions in dyadic spoken interactions. (2010), 5250–5253.

[21] Han Z. Li. 2001. Cooperative and Intrusive Interruptions in Inter- and Intracultural Dyadic Discourse. *Journal of Language and Social Psychology* 20, 3 (2001), 259–284.

[22] Han Z Li. 2001. Cooperative and intrusive interruptions in inter-and intracultural dyadic discourse. *Journal of language and social psychology* 20, 3 (2001), 259–284.

[23] William G Lycan. 1977. Conversation, politeness, and interruption. *Paper in Linguistics* 10, 1-2 (1977), 23–53.

[24] Michael T Motley. 1993. Facial affect and verbal context in conversation: Facial expression as interjection. *Human Communication Research* 20, 1 (1993), 3–40.

[25] Raja Parasuraman and Christopher A Miller. 2004. Trust and etiquette in high-criticality automated systems. *Commun. ACM* 47, 4 (2004), 51–55.

[26] Dennis Reidsma, Iwan de Kok, Daniel Neiberg, Sathish Chandra Pammi, Bart van Straalen, Khiet Truong, and Herwin van Welbergen. 2011. Continuous interaction with a virtual human. *Journal on Multimodal User Interfaces* 4, 2 (2011), 97–118.

[27] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. (1978).

[28] Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society* 29, 1 (2000), 1–63.

[29] Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. (1973).

[30] Elizabeth Shriberg, Andreas Stolcke, and Don Baron. 2001. Can Prosody Aid the Automatic Processing of Multi-Party Meetings? Evidence from Predicting Punctuation, Dis uencies, and Overlapping Speech. (2001).

[31] Elizabeth Shriberg, Andreas Stolcke, and Don Baron. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. (2001).

[32] Alaa Tharwat. 2020. Classification assessment methods. *Applied Computing and Informatics* (2020).

[33] Khiet P Truong. 2013. Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee. (2013), 1404–1408.

[34] Bill Wells and Sarah Macfarlane. 1998. Prosody as an interactional resource: Turn-projection and overlap. *Language and speech* 41, 3-4 (1998), 265–294.

[35] Candace West. 1979. Against our will: Male interruptions of females in cross-sex conversation. *Annals of the New York Academy of Sciences* (1979).

[36] Liu YANG, Catherine ACHARD, and Catherine PELACHAUD. 2022. Annotating Interruption in Dyadic Human Interaction. *LREC* (2022).

[37] Li-chiung Yang. 2001. Visualizing spoken discourse: Prosodic form and discourse functions of interruptions. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.