# Automatic Speech Interruption Detection: Analysis, Corpus, and System

Martin Lebourdais, Marie Tahon, Antoine Laurent, Sylvain Meignier

# Automatic Speech Interruption Detection: Analysis, Corpus, and System

**Martin Lebourdais[1,2], Marie Tahon[1], Antoine Laurent[1] and Sylvain Meignier[1]**
[1]LIUM, Le Mans Université, France
[2]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
martin.lebourdais@irit.fr
{marie.tahon, antoine.laurent, sylvain.meignier}@univ-lemans.fr

**Abstract**

Interruption detection is a new yet challenging task in the field of speech processing. This article presents a comprehensive study on automatic speech interruption detection, from the definition of this task, the assembly of a specialized corpus, and the development of an initial baseline system. We provide three main contributions: Firstly, we define the task, taking into account the nuanced nature of interruptions within spontaneous conversations. Secondly, we introduce a new corpus of conversational data, annotated for interruptions, to facilitate research in this domain. This corpus serves as a valuable resource for evaluating and advancing interruption detection techniques. Lastly, we present a first baseline system, which use speech processing methods to automatically identify interruptions in speech with promising results. In this article, we derivate from theoretical notions of interruption to build a simplification of this notion based on overlapped speech detection. Our findings can not only serve as a foundation for further research in the field but also provide a benchmark for assessing future advancements in automatic speech interruption detection.

**Keywords:** Interruption, Overlap, Corpus

## 1. Introduction

The automatic detection of interruption is a rather new topic of interest. Thus, applications have not yet been used at a large scale. We can still think of some use cases where this knowledge, and the ability to process it automatically would be of interest. Foremost, we can imagine interruption detection being used as a tool to ensure that speaking turns are respected during official political debates. A similar use could be made by companies wishing to improve the flow of meetings, by inviting participants who interrupt too much to give the floor to other speakers. In such cases, interruptions are treated as a form of discourse disruption, without taking into account their causes. Finally, the interruption information can help researchers working on these topics in a more fine-grained manner to speed up the process of collecting area to analyze, for example on work about the link between interruption and dominance (Ferguson, 1977) or between interruption and gender (West and Zimmerman, 1975).

The first question that can be asked is ”What is an interruption?”. If we consider the Merriam-Webster definition, an interruption is ”something that causes a stoppage or break in the continuity of something”. A precise definition is much more subjective than it seems. Multiple works have been conducted on this topic and rely on specific definition or coding convention that simplify the problem to decide if an event is or is not an interruption. Our starting point is the work presented by (West and Zimmerman, 1975). The authors defined an interruption as a subclass of overlapped speech, a segment of speech with at least two concurrent speakers, with a speaker change that occurs before an arbitrary distance to the end of the complete proposition. This definition is a good approximation of an interruption as stated by (Guillot, 2005), but lacks the separation between intentional interruption and anticipation of the end of turn. A second simplification of our work is that by defining an interruption as a subclass of overlapped speech, it ignores the interruptions without overlapped speech, thus missing around 12% of interruptions as stated by (Ferguson, 1977).
In our work, we want to automatically detect interruptions from speech on the basis of two previous studies. The first is the simplification of (West and Zimmerman, 1975) presented earlier that creates a link between interruption and overlapped speech. The second study is the classification of overlapped speech presented by (Adda-Decker et al., 2008). This classification split overlapped speech into four different classes:

- Backchannel: A short interjection that marks the attention of a listener
- Anticipated turn taking: Anticipation of the end of a turn unit without any intention of interrupting the current speaker
- Complementary information: The second speaker adds short information to the main proposition without intention to take the floor

- Interruption: The second speaker overlaps the first before the end of his proposition with intention to take the floor.

As aforementioned, interruption detection is a new task, resources are therefore scarce. To be able to train an automatic detection model, we collect a new corpus annotated in overlapped speech classes. To allows further studies on the emotional impact of such interruptions, we also annotate discrete emotions classes for each of the segments presented. We then propose a deep learning model along with an evaluation protocol to be able to detect if a speech segment contains an interruption. In this work we present a new task, interruption detection, associated with an use-able corpus for training and testing purposes as well as a baseline architecture to provides a comparison.
This work is separated into three main parts. Section 3 presents the different task we'll be approaching, the overlapped speech detection and the interruption detection. Section 4 presents the corpus collection, with the choices made, the analysis of annotations, as well as annotator's ratings fusion solutions. Finally, section 5 presents the automatic detection system, the experiments conducted to validate it and the results obtained.

## 2. Related works

As stated in the introduction, the definition of interruption is too complex, bringing into account the "intention" of a speaker of interrupting, making it impossible to process automatically. To simplify this concept, we'll heavily rely on (Adda-Decker et al., 2008). The proposed classification allows us to consider the overlapped speech as a space that we can classify, and thus discriminate the interruption from the other classes.
The state of the art for automatic interruption detection is extremely limited. Firstly, (Caraty and Montacie, 2015) considers overlapping speech as an interruption, which is a significant oversimplification for our objective. Nevertheless, they used a SVM classifier based on acoustic descriptors to determine the presence of overlapping speech, and a second SVM that utilizes both acoustic descriptors and features derived from the overlapping speech detection system to estimate a conflict level.
The second article addressing interruption detection is more recent. It is an article by (Fu et al., 2022), which presents a system designed for video conference conversations to automatically detect unsuccessful attempts at speaking. To do so, the authors propose to use information provided by a self-supervised system to detect when a participant attempts to obtain the floor, but fails to interrupt. While different from our task, it is sufficiently similar that techniques used in this article could work in our case.
Finally, interruption detection is a subjective task, and will need multiple raters to take into account this subjectivity. As we'll have multiple annotations for the same sample, we want to fuse them into a single gold reference, which will be used to train our model. Multiple methods have been tried in fields relying on subjective annotations, especially in emotion recognition. Regarding discrete annotations, the gold reference can be obtained with a majority vote, or by keeping only consensual data (Tahon et al., 2011; Pappagari et al., 2020). Some further studies also showed that such consensual selection helps in bringing out the prototypical classes (Chou and Lee, 2019), thus allowing a good model of complex concepts. Finally, ensembling of models trained on a single annotator have also been used to take the subjectivity into account (Kim and Provost, 2015).

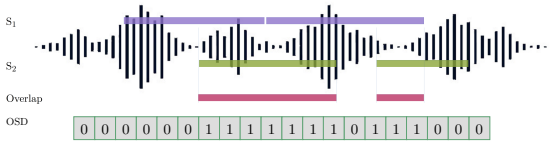



Figure 1: Segmentation by classification: Overlapped speech detection. $S_1$ and $S_2$ are the two speakers and overlap the overlapped speech segmentation. The output of the OSD is a binary vector with a value for each frame

## 3. General overview

Our approach works in two consecutive steps, a first one segment the audio speech to extract the overlapped areas and a second one classifies these areas into interruption or non interruption. The interruption thus lasts the entirety of the overlapped speech segment
The first part of our system is the detection of overlapped speech. This task is defined as a segmentation by classification. The model classifies each frame, sampled at 100Hz, into two classes, either overlapped speech or non overlapped speech. Figure 1 presents an overlapped speech detector (OSD) (Lebourdais et al., 2022). In our work we also use a variant defined as a 3 classes segmentation (Lebourdais et al., 2023) : either no speaker (0), 1 speaker (1) or more than 1 speaker (2) as depicted in Figure 2. This last approach allows getting the information of presence of speech as well as the information of presence of overlap with a single system instead of using a voice activity detector as well as an OSD.
The interruption detection task is thus defined as refinement of overlap segment and some time

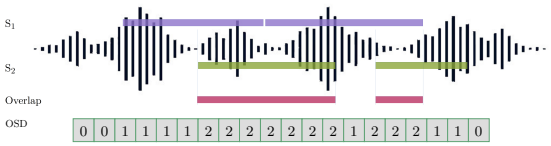



Figure 2: Segmentation by classification: 3 Classes overlapped speech detection. With same notations as Figure 1

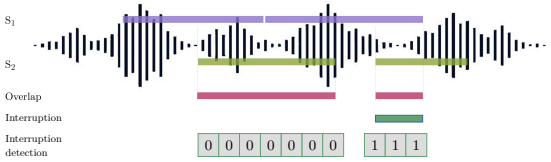



Figure 3: Interruption classification, with same notations as Figure 1

around the superposition to take into account the context into either interruption or non-interruption. Figure 3 shows the classification of overlapped areas.

Our task definition allows to automatically detect interruption but suffer two main flaws, both coming from the overlap step. The interruption classification relies on the precision of the overlapped speech detector, thus any error in this first step will propagate further down the pipeline. The second flaw is that by using overlap as a proxy, we ignore the silent interruptions that we won't be able to address at all.

# 4. Corpus

## 4.1. Data selection

We focus on data coming from audiovisual data and are thus choosing to use a subpart of a meta corpus called ALLIES[1] (Shamsi et al., 2022). This meta corpus is composed of french media shows from the ESTER (Galliano et al., 2006), REPERE (Giraudel et al., 2012), and ETAPE (Gravier et al., 2012) corpora, harmonized and completed by a new set of data for over 500 h of data annotated for diarization. From this annotation we can extract the overlapped speech references and select a subset of 4639 segments of overlapping speech (mean duration 2.6s, standard deviation 2.2s).

We hypothesize that interruption is a subjective phenomenon and thus can rely on contextual information for the annotators to make a decision. Therefore, we add 4 seconds of context before and 4 seconds after the overlap to provide more information to the annotators. The addition of context might add some new overlap segment to the current one, and perturb the annotators. Therefore, we decided to remove the segments for which an overlap segment occurs in the 2/4sec before the current segment. An overlap after the current one is considered valid.

This selection process leads to a set of 4,639 segments with a length varying from 8 to 20 seconds.

## 4.2. Annotation protocol

### 4.2.1. Annotation tool

To annotate these data we recruited 4 annotators. The annotation have been done using a modified version of FlexEval (Fayet et al., 2020) This framework was initially dedicated for speech synthesis listening tests. All participants do not need to rate all samples, therefore samples are randomly distributed to each participant. In our case, we want all participants to rate all samples, however, we can not get rid of the random affection of the samples. Consequently, the use of this platform created a side-effect and made annotators get twice some segments, but we ensure all samples can be rated by all annotators.

### 4.2.2. Annotation of overlap class

The first part consists in assigning to each overlap a class. We initially selected the four classes presented in section 1 (backchannel, anticipated turn-taking, complementary information, interruption). But after a training phase, we decided to add three new classes on the basis of annotators' feedbacks :

- No overlap: The reference can be false and provide a segment without overlap to annotate. This class have been used as a "can't give a decision" class.
- Simultaneous start: The classes defined earlier are relevant in a dyadic situation. Our conversation may have more than two speakers, we can get in a situation where two speakers answer simultaneously a question asked by a third.
- Brouhaha: This last class contains the segments where multiple speakers are constantly overlapping.

### 4.2.3. Annotation of emotional state

In order to enrich the interruption annotation and better analyze this phenomenon, we have also proposed an emotion annotation task with two parts.

We first ask the annotator to assess the emotions of the main speaker before the overlap, during the overlap and after the overlap. The main speaker can change between these areas. We decided

[1] ALLIES will be included to ELRA catalogue in 2024.

| annotator | # uniq. seg | # dupl. seg. | consistency |
|---|---|---|---|
| A1 | 4331 | 308 | 0.594 |
| A2 | 4346 | 293 | 0.608 |
| A3 | 2444 | 171 | 0.860 |
| A4 | 4334 | 305 | 0.613 |

Table 1: Intra annotator consistency for the interruption detection annotation task, with the number of unique segments annotated, the number of duplicates and the consistency

not to use the Big six emotions (fear, anger, sadness, joy, disgust, surprise) (Ekman and Friesen, 1971) as they are mostly suitable for prototypical speech. As described in (Devillers et al., 2005) the expression of emotion highly rely on the context and the persons. Moreover, ambiguous emotions are frequently observed in real-life scenarios. Consequently, we selected a large set of emotions from Scherer circle (Scherer, 2005). This set of emotion is deliberately composed of vague concepts without narrow explanation to the annotators to capture the complex nature of human expressivity : frustrated, annoyed, impatient, calm, neutral, friendly, confident, attentive, feeling superior, enthusiastic, convinced, feel guilt, confused, indignant, hesitant, worried, contemptuous.

### 4.3. Analysis of corpus

We formalize our corpus with the following conventions :

$M = 4$ nb of annotators,
$j$ indexed for annotator
$N = 2393$ nb of segments labeled by 4 annotators
$i$ index for a segment
$K = 7$ number of classes,
$k$ indexed for class

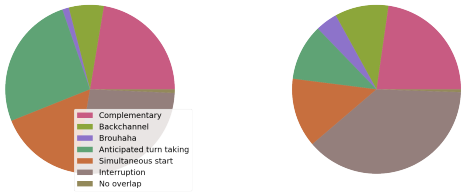
$A$ set of labels
$A_{ik}$ labels of segment $i$ and class $k$
$A_{ijk}$ labels of annotator $j$
for the segment $i$ with the class $k$

#### 4.3.1. Intra annotator consistency

The four annotators were required to process the entire set of 4639 segments. However, they did not annotate the same number of unique segments. , as indicated in Table 1. Nevertheless, annotators 1, 2, and 4 had the impression of annotating all segments. A side-effect of the implementation of our platform occasionally resulted in the re-presentation of segments already annotated. However, we can use this information to our advantage as it allows for the verification of the consistency of these annotations, *i.e.*, an annotator's ability to reproduce the same annotation on the same segment.

We define the consistency of a given annotator as the number of identical annotations for a segment $i$ divided by the total number of annotations by that annotator for each segment $i$. For example, a segment annotated three times with two similar classes obtain a consistency of 2/3. The obtained value corresponds to the probability that an annotator gives the same response for a sample presented twice. It should, therefore, be close to 1. The global consistency is the mean of consistency As presented in Table 1, the consistency for annotators 1, 2, and 4 is close to 0.6, meaning they have almost a fifty-fifty chance of changing their opinion on the same segment.

A3 did not complete the annotation but apparently paid attention to their quality. However, it is essential to consider that an annotator who consistently responds with the same class will have a consistency of 1 but may provide uninformative annotations. Consequently, it is also necessary to evaluate the distribution of the annotated classes.




(a) Annotator 1 (b) Annotator 2

(c) Annotator 3 (d) Annotator 4

Figure 4: Overlapped speech classes per annotator

Figure 4 summarizes the distribution of overlapped speech class annotated by the four annotators. At first glance, the distributions may appear similar, but several phenomena are worth noting. Firstly, the *interruption* class (brown) generally occupies the major portion of the annotations. However, the distributions highlights significant differences in the annotation strategy used by each annota-

tor. A1 uses the *interruption* class less frequently than others, instead favoring *anticipated turn taking* (green) which takes a substantial portion of the distribution. This situation is not surprising, as these two classes have also shown a high degree of confusion in the work of Adda-Decker et al. (Adda-Decker et al., 2008). One could argue that it is worthy to merge them, since both classes are perceptively hardly distinguishable. However in our work, we intend to discriminate anticipated turn taking from interruption that prevents the complete message to be transmitted.

A3 has an annotation distribution that is very different from the others. This annotator also took more time to process the files than the others, resulting in a smaller number of annotated segments. This study allows us to confirm the absence of an ill-defined class that would have hindered the smooth progression of subsequent analyses and predictions. Having analyzed the annotations of each annotator separately, the next logical step is to examine the agreement among these users.

#### 4.3.2. Inter annotator agreement

Agreement across annotators is usually measured using Fleiss' kappa. Considering only the segments annotated by the four annotators ($N = 2393$), we obtain a kappa value of $\kappa = 0.311$ on the 7 classes. While this kappa value is low, the large number of classes and their imbalance bias this metric against us.

The limited number of segments annotated by A3 poses an issue for the further analysis and utilization of these annotations. Therefore, we consider the option of excluding A3 from the annotation results. We calculate the kappa on segments annotated by the other three annotators, which are a lot more ($N = 4277$), and we obtain a $\kappa = 0.353$ on the 7 classes. Our result is still quite low but slightly better than the kappa with all four annotators. Moreover, the annotation distributions (Figure 4) exhibits greater homogeneity when considering only annotators 1,2 and 4. In (Tahon et al., 2011), the authors obtained a kappa of 0.4 for 5 classes of emotions in a real-life scenario and 2 annotators. Therefore, we conclude that the low agreement between raters highlights the difficulty of the task and the high subjectivity in the perception of these 7 classes. Furthermore, we exclude the annotator 3 from the evaluation for the rest of this article.

With the view to having annotations robust enough to train an interruption detection model, we need a higher consensus between annotators. To do so, we have already discarded an outlier annotator. But another option would be to merge or discard classes for which the agreement is too low. One option is to compute agreement towards a specific

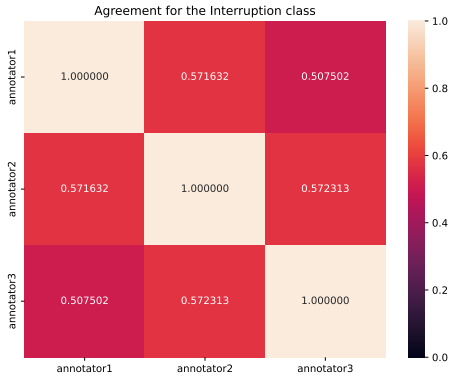



Figure 5: Agreement between Interruption and the 6 remaining classes for 3 annotators

class $k$, by considering the remaining classes as a unique $non(k)$ class. However, it is not straightforward to determine the detailed agreement by class. To obtain this information, we propose to define pairwise agreement between two annotators, 1 and 2, on a class $k$ as defined in Equation 1 where $A_{i1k}$ is the label for the segment $i$ and the class $k$ for the annotator $1$.

$$Agreement_{1,2,k} = \frac{Card(\{A_{i1k}, A_{i1k} = A_{i2k}\})}{Card(\{A_{i1k}\} \cup \{A_{i2k}\})} \quad (1)$$

For example, Figure 5 displays the agreement among the three annotators for the class we are interested into classifying, the *interruption* class. As we only consider unique segments and thus choosing the last annotated segment in the case of duplicate, the diagonal agreement represents the agreement of annotators with themselves, and this value cannot differ from 1, rendering it non-informative. The off-diagonal agreement, which increases as agreement between pairs of annotators improves, ranges between 0.51 and 0.57. Overall, this score is satisfying enough to build gold standard from these references.

### 4.4. Annotation fusion strategies

We aim to create an interruption detection system on overlapped speech segments with two classes: "interruption" or "non-interruption". We have three annotations for each segment and the goal is to derive a gold reference label for each segment from these annotations. To achieve this, we investigate different annotation fusion strategies.

For each of the three strategies described below, Table 2 summarized the total number of segments with a gold reference, the number of segments with a reference being Interruption, and test

**Majority Vote** This strategy assumes that a unique correct answer exists and is given by the

| Fusion | # Segments Train | Test | Interruption segments |
|---|---|---|---|
| Unanimous | 1448 | 449 | 571 |
| Majority decision | 4277 | 1309 | 1499 |
| Weighted by Agreement | 12831 | 1309 | 4679 |

Table 2: Number of available segments to train and test for each annotation fusion method.

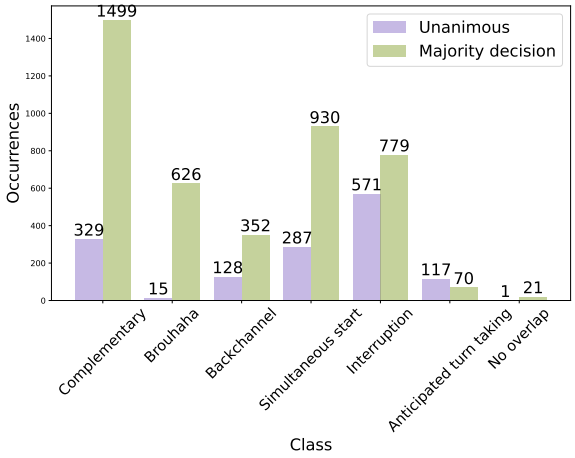



Figure 6: Distribution of overlapped speech classes for unanimous selection and majority decision

majority of annotators. An annotator whose response differs from the other two is not considered. To implement this, we consider the most frequently assigned class for a given segment as the gold reference. If none of the annotators agree, annotator A1 is designated as the reference, since they are deemed the most reliable. This solution has the advantage of keeping all overlap segments as shown in Table 2.

However, this solution does not fully satisfy us, as it removes the subjective nature of interruption perception. One person may perceive an interruption that another person does not, without either being necessarily wrong. Therefore, we will prioritize other annotation fusion solutions.

**Unanimous Selection** The second solution attempts to address subjectivity by selecting examples that minimize it. By choosing segments on which all three annotators agree, we reduce the impact of subjectivity by excluding borderline cases. However, this solution drastically reduces the number of segments ($N = 1448$) .

Figure 6 presents the distribution of segment per class for a unanimous decision and a majority decision. As shown in Figure 6, unanimous fusion not only significantly reduces the number of segments but also modify the distribution of classes compared to the distributions of separate annotators represented in Figure 4. The interruption class stay relatively similar between unanimous and majority decision, meaning that if two annotators agree on putting a segment in the interruption class, the third will most of the time also answer interruption. We have chosen to keep only unanimous segments for all classes, although we group all non-interruption classes later. This decision is made because if annotators do not agree on a specific class, there is no guarantee that a fifth or sixth annotator would not hesitate to classify the excerpt as an interruption.

However, this fusion method lacks nuance. A system trained on this fusion may lose effectiveness in a real-world scenario, as it has not seen borderline cases and would only be suitable for detecting obvious interruptions.

**Agreement weighted data augmentation** The last strategy we investigate leverages machine learning properties for classification. The idea is to present the same segment multiple times to the system with different labels to disrupt the convergence of non-unanimous examples into a specific class. Consequently, this approach artificial augment the data by a weighted duplication based on agreement. In this approach, we consider all annotators' labels as gold references during the training phase. Because our system will return only one prediction, we can not apply the same strategy for the test segments, therefore we consider only a majority vote for evaluation.

This method has a variant involving weighting the segments to give more importance to unanimous excerpts. To achieve this, we duplicate the occurrences of an excerpt according to a power of 2. For example, an excerpt without any agreement (all the three labels are different) will have each of its annotations placed $1^2$ times, an excerpt with a majority class will have its majority class placed $2^2$ times in the corpus, and a unanimous excerpt will have its class placed $3^2$ times in the corpus. This approach artificially increases the number of examples in the corpus but requires more processing and introduces uncertainty about the success of the learning process.

In conclusion, each of the three strategies has its pros and cons. We decided to compare the performances of a system trained with unanimous labels and with the weighted by agreement data augmen-

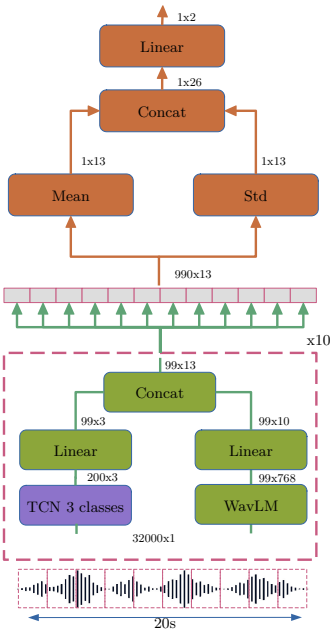



Figure 7: Interruption detection architecture, the bottom third of the figure is the feature extractor composed, with a frozen OSD using a TCN

tation approach.

## 5. Interruption detection from speech

### 5.1. Deep neural network description

Due to the subjectivity of the task, the amount of annotated data is limited. Consequently we decided to have a low amount of parameters in our model in order to prevent overfitting. The architecture of our deep neural network (cf Figure 7) is motivated by the fact that perception of an interruption relies on both the presence of overlapping speech, and additional high-level audio features summarized with mean and variance at the segment level. The first feature extraction branch (left) is a TCN-based overlapped speech detector (OSD) with 3 classes as described in (Lebourdais et al., 2023) trained on ALLIES corpus (475 hours) and frozen. The OSD takes input a WavLM (Chen et al., 2021) representation of a 2s speech signal and return 3 pseudo-probabilities for the absence of speech (0), the presence of a unique speaker (1) or the presence of multiple (speakers). The idea of using this system is to add information about overlapped speech presence as an input feature to the interruption detection model. The input segments vary in length accordingly with the length of the training overlap segments, ranging from 8 seconds to 20 seconds. They are padded with zeros to reach a length of 20 seconds. To be processed by the OSD, this 20-second segment is divided into 10 non-overlapping 2-second segments.

The second feature extraction branch (right) consists of a variant of WavLM proposed in the original paper (Chen et al., 2021) called ”Base plus”. This variant is a reduced version of the previously used model, trained on the same training corpus with 768 output dimensions.

To prevent the three OSD outputs from being overwhelmed by the 768 features from WavLM, a linear layer is trained to compress the WavLM information into 10 dimensions. Additionally, since the overlapped speech detector is not sampled at the same frequency as WavLM, we perform interpolation to downsample the overlapped speech detection to the WavLM frequency. The two types of features are concatenated first along the feature axis and then along the time dimension to obtain 20-second segments. This results in a segment with 990 samples and 13 features for a 20-second audio clip. Intuitively, 13 framewise features may appear insufficient to discriminate such a complex event as an interruption. Therefore, we chose to calculate the mean and standard deviation over time for this vector, and then concatenate the results to obtain 26 segment-level features that can be classified with a linear layer with two outputs. This technique is also commonly used in emotion recognition (Macary et al., 2020).

### 5.2. Unanimous model training protocol

As presented in Table 2, the number of training data drastically drops when only consensual annotations are kept. To address this issue, we employ a 5-fold cross-validation approach, with folds randomly selected and no overlap between them. The test set is kept aside from this cross-validation to enable comparison with other fusion methods.

Our model is trained on 4-folds using the Adam optimizer and a cross-entropy loss function and evaluated on the last folds (5 times). Given the imbalanced nature of the corpus, the evaluation metric employed is the F1-score, calculated separately for each fold. Lastly, since there is no prior state-of-the-art model available for assessing model performance, we compare our results with those obtained by replacing the model’s output with randomly uniform generated numbers of the same dimension before applying the softmax function. To select the best model over folds, we compare the relative improvement over the random baseline and choose the highest performing model. Each fold is trained for 10 epochs, and the model with the best validation performance is retained for that fold. The performances in terms of F1-score calculated on the interruption class are given in Table 4. While it is not the best absolute F1-score the fifth fold bring the best relative improvement over random and will be used as the model designated as *Unanimous*.

| System | Fusion strategy for test labels<br>Unanimous (449 segs) | Majority vote (1309 segs) |
|---|---|---|
| Unanimous | 76.57 | 59.79 |
| Weighted by Agreement | 77.04 | 61.48 |
| Random | 44.53±0.62 | 39.60±0.42 |

Table 3: Results of interruption classification on the two models with F1-score, recall, and precision expressed in percentage (%) and a system providing random results for comparison. The evaluation corpus contains only unanimous segments.

| Folds | F1-score | Random F1 | Gain |
|---|---|---|---|
| Fold 1 | 75.10 | 51.52 | 1.458 |
| Fold 2 | 72.65 | 45.38 | 1.601 |
| Fold 3 | 71.93 | 45.86 | 1.568 |
| Fold 4 | 73.08 | 45.91 | 1.592 |
| Fold 5 | 74.55 | 38.13 | **1.955** |

Table 4: Results of the interruption classification across 5 folds, presented in terms of F1-score in percentage (%) and a F1-score calculated on random results for comparison.

### 5.3. Results

In this experiment, we compare the unanimous and the weighted by agreement fusion strategies to train the network, while evaluating on unanimous reference and the complete data (majority vote). As we don't have a baseline in literature to compare with our models, we compare to 10 random models. All the results are given in terms of F1-score calculated on the interruption class.

We first evaluate the two models on data unanimously annotated by our annotators, resulting in 449 segments. Table 3 presents the results obtained by the *Unanimous* model and the *Weighted by Agreement* model on this test configuration. All results overpass random guess, which means that both models are able to learn to discriminate interruptions from the other classes of overlapping speech. We expected the Unanimous model to learn accurate and prototypical representation of interruption, however it performs slightly worse than the Weighted by Agreement model. However, the results of both models remain close, with F1-scores of 76.57% and 77.04%, respectively, for the *Unanimous* and *Weighted by Agreement* models.

As discussed in section 4, the unanimous fusion strategy does not fully capture the difficulty and the subjectivity of the interruption annotation task. Therefore, in a second phase, we evaluate the two aforementioned systems on reference obtained using majority vote. This method is expected to introduce more challenging cases compared to the previous evaluation.

From Table 3 we can see that the obtained results are lower than those previously achieved, but the task involves more subjectivity, and a greater number of segments are considered (1309 compared to 449). As anticipated, the model trained on data weighted by agreement outperforms its counterpart trained on unanimous data, with F1-scores of 61.48% and 59.79%, respectively. Even if these performances are quite low for a binary classification task, we demonstrate that our model is able to capture some representation of interruption despite the fact that the annotator agreement was weak. We also show that our model performs quite well on non-ambiguous consensual data.

## 6. Conclusion

In this article, we introduce the field of interruption detection. This field lacks established frameworks and resources, necessitating the collection of relevant data. To accomplish this, we propose a corpus extracted from overlapping speech segments in debate TV/radio shows sourced from ALLIES corpus. The overlapping speech samples have been annotated in 7 classes, including interruptions. While this corpus includes annotations suitable for interruption detection, there are several aspects that require improvement. The extensive number of classes likely mitigates bias towards the interruption class by redistributing uncertain segments among other classes. However, it also substantially reduces inter-annotator agreement, making annotations challenging to merge. The analysis of annotations demonstrates 1) a real difficulty to perceptively discriminate between these classes, and 2) the high subjectivity in the perception of interruptions in such data. The annotations produced are available at https://lium.univ-lemans.fr/en/corpus-allies/.

We also presented an interruption classifier which includes an overlapping speech detector and predicts if an interruption occurs on the given segment or not. In order to take into account the subjectivity, different label fusion strategy to train the models and to evaluate them, were investigated. All the proposed models demonstrate classification performance largely exceeding random chance. This result illustrates that predicting a distinctly subjective dimension is possible with neural networks. We conclude from this extensive work

that the definition of interruption in our study corresponds to our annotators' notion of interruption, which is clearly not universal. Therefore, a study with a larger dataset, a greater number of annotators, and more diverse audio samples, should probably lead to better results in this task.

## 7. Acknowledgements

This work has been partially funded by the French National Research Agency (project Gender Equality Monitor - ANR-19-CE38-0012). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101007666. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011012565).

## 8. Bibliographical References

Marie-José Caraty and Claude Montacie. 2015. Detecting Speech Interruptions for Automatic Conflict Detection. *Conflict and Multimodal Communication: Social Research and Machine Intelligence*, pages 377–401.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv*.

Huang-Cheng Chou and Chi-Chun Lee. 2019. Every Rating Matters: Joint Learning of Subjective Labels and Individual Annotators for Speech Emotion Classification. pages 5886–5890.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422. Emotion and Brain.

Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129. Place: US Publisher: American Psychological Association.

Cédric Fayet, Alexis Blond, Grégoire Coulombel, Claude Simon, Damien Lolive, Gwénolé Lecorvé, Jonathan Chevelu, and Sébastien Le Maguer. 2020. FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, pages 22–25, Nancy, France. ATALA.

Szu-Wei Fu, Yaran Fan, Yasaman Hosseinkashi, Jayant Gupchup, and Ross Cutler. 2022. Improving Meeting Inclusiveness using Speech Interruption Analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pages 887–895, New York, NY, USA. Association for Computing Machinery.

Yelin Kim and Emily Mower Provost. 2015. Leveraging inter-rater agreement for audio-visual emotion recognition. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 553–559, Xi'an, China. IEEE.

Martin Lebourdais, Théo Mariotte, Marie Tahon, Anthony Larcher, Antoine Laurent, Silvio Montresor, Sylvain Meignier, and Jean-Hugh Thomas. 2023. Joint speech and overlap detection: a benchmark over multiple audio setup and speech domains. ArXiv:2307.13012 [cs, eess].

Martin Lebourdais, Marie Tahon, Antoine Laurent, and Sylvain Meignier. 2022. Overlapped speech and gender detection with WavLM pre-trained features. In *Interspeech 2022*, pages 5010–5014. ISCA.

Manon Macary, Martin Lebourdais, Marie Tahon, Yannick Estève, and Anthony Rousseau. 2020. Multi-corpus Experiment on Continuous Speech Emotion Recognition: Convolution or Recurrence? *SPECOM*, pages 304–314. MAG ID: 3088876910.

R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak. 2020. X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7169–7173.

Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729.

Marie Tahon, Agnes Delaborde, and Laurence Devillers. 2011. Real-life emotion detection from speech in human-robot interaction: experiments across diverse corpora with child and adult voices. In *Proc. Interspeech 2011*, pages 3121–3124.

## 9. Language Resource References


M Adda-Decker, C Barras, G Adda, P Paroubek, P Boula de Mareüil, and B Habert. 2008. Annotation and analysis of overlapping speech in political interviews. page 7.

Nicola Ferguson. 1977. Simultaneous speech, interruptions and dominance. *British Journal of Social and Clinical Psychology*, 16(4):295–302. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8260.1977.tb00235.x.

S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri. 2006. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 139–142, Genoa, Italy.

Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The REPERE Corpus : a multimodal corpus for person recognition. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 1102–1107, Istanbul, Turkey.

Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proc. Language Resources and Evaluation Conference (LREC)*, pages 114–118, Istanbul, Turkey.

Marie-Noëlle Guillot. 2005. Revisiting the methodological debate on interruptions: From measurement to classification in the annotation of data for cross-cultural research. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, pages 25–47.

Meysam Shamsi, Anthony Larcher, Loïc Barrault, Sylvain Meignier, Yevhenii Prokopalo, Marie Tahon, Ambuj Mehrish, Simon Petitrenaud, Olivier Galibert, Samuel Gaist, Andre Anjos, Sébastien Marcel, and Marta R. Costa-Jussà. 2022. Towards Lifelong Human Assisted Speaker Diarization. *Computer Speech and Language*. Publisher: Elsevier.

Candace West and Don H. Zimmerman. 1975. Sex roles, interruptions and silences in conversation. In Barrie Thorned and Nancy Henley, editors, *Language and Sex: Difference and Dominance*, page 105–129. Newbury House, Rowley, Mass.