Comparison of Two Approaches for Human Tense Situation Analysis in Car Cabin

Quentin Portes
Renault Software Lab
Toulouse, France
quentin.q.portes@renault.com

José Mendes-Carvalho
Renault Software Lab
Toulouse, France
jose.mendes-carvalho@renault.com

Julien Pinquier
IRIT, Paul Sabatier University, CNRS
Toulouse, France
julien.pinquier@irit.fr

Frédéric Lerasle
LAAS-CNRS, Paul Sabatier University
Toulouse, France
lerasle@laas.fr

***Abstract*—The use of audio, video and text modalities to simultaneously analyze human interactions is a recent trend in the field of deep learning. The multimodality tends to create computationally expensive models. Our in-vehicle specific context requires recording a database to validate our approach. Twenty-two participants playing three different scenarios ("curious", "argued refusal" and "not argued refusal") of interactions between a driver and a passenger were recorded. We propose two different models to identify tense situations in a car cabin. One is based on an end-to-end approach and the other one is a hybrid model using handcrafted features for audio and video modalities. We obtain similar results (around 81% of balanced accuracy) with the two architectures but we highlight their complementary. We also provide details regarding the benefits of combining different sensor channels.**

***Keywords*—Human interactions, multimodality, data fusion, audio & video features, end-to-end.**

I. INTRODUCTION

Today, most of the data available on the Internet is saved under a video or an audio format. Sight and hearing are the main channels used by the brain to understand and decode human interactions. Voice is implicitly processed into words by the brain. In case of multiple speakers, improving the process of interaction analysis could lead to increase performances of sentiment, emotion and dialog analysis. These multiple speaker situations are common in the industrial context, *i.e.* the necessity to improve social media filtering, human-machine interaction understanding, brand monitoring, etc. In the automotive context, it will answer safety concerns (*i.e.* taunting, bullying or, in the worst case, aggression) linked to the new usages of cars (*i.e.* socializing, vehicle sharing, autonomous cars, etc.). Our aim is to detect the signals leading to these situations in order to anticipate and avoid them.

To address this issue, we can analyze the passengers' interactions thanks to cameras and microphones on boarded in the car cabin. These two sensors generate three modalities (video, audio and the text transcribed from the audio), which can be combined to significantly improve the performances of human tense situation predictions.

Today, these modalities are usually analyzed with deep learning approaches. We use Bidirectional Encoder Representations from Transformers (BERT) architecture [1] (English language), Roberta and CamemBERT models [2] (French language) for text analysis. They have improved the global performance in question answering, text summarizing tasks, etc. Recent work uses the transformer model for text dialog analysis [3] [4].

For the video modality, 2D and 3D [5] [6] convolutional approaches are the predominant architectures to analyze images and video.

The most common technique regarding audio analysis is the extraction of audio features over a short sliding window with a framework, such as open SMILE [7]. Then, they are usually fed to a sequential model like Long Short-Terme Memory (LSTM) [8].

One way to improve performances of such models is to combine the audio, video and text analysis. This approach contains more information than the video and audio modalities separately [9]–[13].

The automotive context is an embedded system with some associated constraints: execution time, limited computational resources, memory access, etc. Processing and analyzing three modalities with deep learning algorithms tend to induce large models. To deal with the multimodality and the embedded constraints, one solution is to design a hybrid model with one compact model based on handcrafted features running on embedded hardware and one larger model with an end-to-end architecture running on a cloud platform.

The four challenges identified are the following:

- The availability of a public in situ dataset.
- The fusion between non-heterogeneous modalities like video, audio, and text.
- The complexity to model human interactions.
- The embedded constraints.

Actually, to the best of our knowledge, the literature does not deal with all of these issues at the same time. We are addressing them hereafter. This research focuses on recording an exploitable dataset for industrial applications and then

 ISBN: 978-1-61208-942-3

designing two different approaches showing the benefits of the multimodality for detecting tense situations in the car's cabin.

We differ from the literature by our realistic in-situ dataset and our two complementary multimodal models. We also present two different strategies of late fusion.

Section II introduces a literature review on multimodal dialog analysis. In Section III, we detail the protocol used to record our own dataset and its specifications. Section IV provides details and compares our two multimodal approaches for the classification of tense human interactions. Finally, Section V present our results.

II. Related work

The modern dialog, interaction and conversation analyzing models are based on text [14], [15]. Recent investigations, with new approaches such as multimodality, show the benefits of exploiting information from different channels. Every multimodal model on sentiment analysis fields outperforms unimodal architecture ones [9], [16]. Due to the heterogeneity of the modalities (audio, video and text) used in these architectures, the features are extracted per modality. Then, a final, more or less complex, late fusion is applied to obtain better results. The end-to-end models extracting the features tend to be computationally expensive compared to handcrafted approaches. They also need more data to be trained. Most of the time, handcrafted and end-to-end models are compared only on prediction performances. In the context of human behavior understanding, we can capitalize on the full potential of both techniques. Indeed, the study of human interactions, sentiment analysis or emotions represents some knowledge that we can directly inject in a model. Conversely, we can automatically let end-to-end models find features. These two opposite techniques can be complementary in some scenarii.

Our preliminary works are based on a public dataset like MOSI [12] [17]. We identified work on multimodal conversation analysis such as [18] [19] that train on this previous dataset. Additionally, they are only focusing on sentiment and emotion conversation analysis.

Hierarchical Attention Network (HAN) architecture [20] is performing very well as the Transformer [1] on document analyzing. Recent approaches, such as [3], are using Transformer for dialog analysis. As we are working on oral text and a small dataset, the HAN approach seems to be the most appropriated.

Regarding interaction analysis, the speaker's previous behaviors are crucial to hold. Nowadays, the deep learning architecture is not able to process extensive videos. The use of stateful temporal models [8] in our approach will allow us to keep track of the information over scenario duration.

The investigations concerning the car cabin passenger interactions are very scarce and remain a scientific challenge.

III. MULTIMODAL DIALOG CORPUS IN VEHICLE

In this section, we detail the protocol used to record our multimodal dataset. The aim is to classify three different types of interactions. The first one is the "normal/curious" category where two participants have a cordial discussion. The second one is the "argued refusal", where the rear passenger refuses cordially the driver's proposition. The last one is a full refusal of the driver proposition, called: "not argued refusal". The insistent seller scenario has been chosen instead of an aggression scenario for two reasons. The first one is our objective to find discussion resulting in aggression and not physical aggression. The second is due to protocol reliefs reasons. Indeed, willing to play realistic aggression scenarios, obliging to follow a psychological protocol setup for the different subjects would be very restrictive.

A. Purpose of the dataset

We recorded the interactions between two passengers in a car's cabin (see Fig. 1). One driver and one rear seat passenger (right side) are playing predefined non-scripted scenarios. Subjects are French volunteers without any acting skills.

Each pair of participants is recorded for 7 minutes, scheduled in a session of four continuous stages. This paper only focuses on the acting stage:

1) 60s of silence,
2) **180s of acting**,
3) 60s of silence,
4) 120s of interaction with the In-Vehicle Infotainment (IVI).

During the acting stage, the driver always plays the same role of an insistent seller and the passenger plays one of the three following behaviors:

- "be curious about the driver proposition",
- "refuse the proposition with argumentation",
- "refuse categorically the proposition".

We set up a double-blinded scenario. The driver and the passengers never knew the situation that has to be played beforehand. In this configuration, we can say that the driver undergoes the situation.

B. Acquisition setup

We equip a Dacia Duster with six cameras, four microphones and one screen placed on the hood of the car. The screen is in front of the driver view and also visible by the passenger. Its use is motivated to indicate when the subjects have to change the acting phase and stream a video of the road to captivate the driver's attention due to the stationary car. The interactions with the car are available (wheel, gear lever, etc.).

1) Video steaming: All the cameras present in the setup have different resolutions, angles of view and lenses. Our approach privileges the camera #2 because it has the best view and lighting quality. It is a manual-focus camera of recording resolution 1920×1080 pixels. It is placed in order to have a front angle of view, see Fig. 1.

The other cameras will be considered for future investigations.

 ISBN: 978-1-61208-942-3

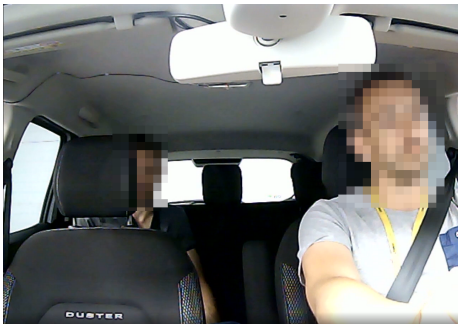

Fig. 1. Field-of-view of the camera #2.

2) Audio streaming: Four identical microphones Brüel&Kjaer prepolarized 1/4 inch Type 4958 are set in different vehicle areas recording the audio stream. Our approach only uses the ceiling driver's microphone because it is the only area used by the car manufacturer.

We recorded all the video and audio streams in RAW format (no live compression) in the concern of not using too much computational power.

C. Preprocessing and annotation of the dataset

Once recorded, a step of post-processing is mandatory. Indeed, the recording process generates a temporal delay between video and audio streams. The six videos and four audio streams are synchronized with Adobe premiere pro. Ultimately, the videos are compressed in MPEG-4 format.

The third modality (text) is obtained by transcribing the audio stream. After some experiments, we avoid the automatic speech transcription (ASR) such as *Amazon transcribe* or *Google speech to text* due to their very high word error rate. In our oral context, repetitions, interjections and isolated words are the most important parts of the dialog. Furthermore, the sentences are potentially poorly constructed (subject-verb-complement). ASR techniques are inefficient in that case.

We use the ELAN software to transcribe the dataset. It is a manual annotation tool for video and audio data. The audio stream is transcribed into utterances for each actor, resulting in a total number of 2026 utterances. An utterance is a continuous unit of speech beginning and ending with an explicit pause. The transcript is reviewed by a peer.

To reduce the annotation time, we annotate at the scenario level in comparison to other datasets [17], where the annotations are at the utterance level. An entire recording sequence is annotated at the beginning of the recording. This choice has some repercussions: for instance, it induces wrong labels if the subjects play their roles not in adequacy with the asked one. We will come back to these issues later in the qualitative analysis (see Section V-B).

D. Specifications and understanding of the corpus

Finally, the dataset contains 44 videos for 22 participants (4 women/18 men). Each pair of participant plays once as a driver and once as a passenger in a random order. The accumulated interactions give on average 46 sentences per video, for a total of 2026 sentences. This represents 21 966 words containing 2082 unique words. We get 54 min for the "curious" class, 27 min for the "argued refusal" class, and 27 min for the "not argued refusal" class, which represent a total of 1h48. An asymmetry in the amount of data recorded is added to take into consideration the fact that in real situations the curious class would be the usual behavior.

By examining the video, we notice that the video modality is less informative than the audio and text ones. Indeed, the passengers are mostly static due to the car context and the belt as well as the driving task restricting the movements. We also detect this outcome in sentiment or dialog analysis based on multimodal datasets [16] [21].

The analysis of the dataset over time shows patterns in the drivers' and passengers' behaviors. Humans are not swapping their emotions or behaviors at a high frequency. Taking this information into account, we decide to plot the features as a function of time for a 15s analyzing window; values higher than 30s result in flat curves with no possible deduction. This Github link[1] makes available the plotted chart. The local descriptor plots are inspired from [22].

After examining the audio-video streams and analyzing the charts, we are able to focus on seven hand-crafted features, as indicated below. Four of them are generated by "the mean talking" and "mean duration" for the two passengers and the three remaining are the "mean silence", the "eye contact" and the "passenger visibility":

- Mean talking - In a normal conversation, the average talking tends to be equitably distributed among the participants.
- Mean duration - It is the average duration of the utterances. Complementary to the mean talking, the length of a speech is a good indicator of who is dominating the conversation and who wants to close the dialog.
- Mean silence - The mean silence is an indicator of the intensity of a dialog. The more silence there is, the more the discussion is poor and tends to be in the refusal situation.
- Eye contact - It is the frequency at which the driver is looking into the interior rear-view mirror. Eye contact is a natural behavior when talking to someone. As the driver is focused on the road and on the driving task, he has no other choice but to look at the rear-view mirror to see its interlocutor.
- Passenger visibility - It is the frequency at which the passenger is seen by the camera. It is a good indicator of the passenger's interest in the conversation. We naturally reduce the distance with our interlocutor when we are engaged in a discussion. In the car discussion context, the rear passenger can move forward between the two front seats. On the video stream, it results in seeing (or not) the rear passenger.

For the text modality, we calculate the frequency distribution of words and the term frequency-inverse document frequency

[1] https://github.com/QuentinPrts/MMEDIA_2021

 ISBN: 978-1-61208-942-3

(TF-IDF) [23] to find if there are specific distributions of words associated with a given scenario. These approaches are very common in text mining and analyzing. The TF-IDF delta between the two opposite classes (”curious” and ”not argued refusal”) exhibit the 10 following most important delta words: *je (I), pas (not), vous (you, second-person plural), ouais (ok), tu (you, second-person singular), non (no), moi (me), oui (yes), donc (so)* and the *ah* interjection. The text modality is not rich (as a reminder, we have 2082 different words).

In the chart, we observe two transition phases. The first one is the setting up: the subjects could not be insistent or categorical in their refusal to lead a ”bad acting” in the first 30s of each scenario. The second is at the end: subjects run out of inspiration, causing shortness of breath for the last 20s of each scenario. This changeover is due to the individuals playing the scenarios: they are volunteers and not real actors.

IV. MULTIMODAL ANALYSIS

After analyzing the dataset, we implement two different approaches, one end-to-end model (noted E2E) and one based on handcrafted features (noted H). They have to process data to classify the input stream into three classes corresponding to our three scenarios (”curious”, ”argued refusal”, ”not argued refusal”). The two architectures are detailed in the following sections. Fig. 2 illustrates our two approaches. First, we implement a dedicated model for each modality and evaluate their performances after fusing their outputs. Then, the modalities are converted into a generator of features for a multimodal fusion purpose.

As the basic analysis of the text modality is not performing very well, we decided to implement a deep learning model. It will be used for both approaches presented in the following section.

A. Text analysis

We face a major problem in the text modality. Indeed, every framework and pre-trained models such as Spacy [24], NLTK [25], BERT [1] are well suited for English analysis but perform very badly on the French language. The existing French alternatives are very limited because they are based on old or written French. Thus, we did not obtain sufficient results on the transformers model named Camen-BERT [26] which is trained on Wikipedia text. The poverty of our text makes the basic approach (TF-IDF and embedding + LSTM model) inefficient.

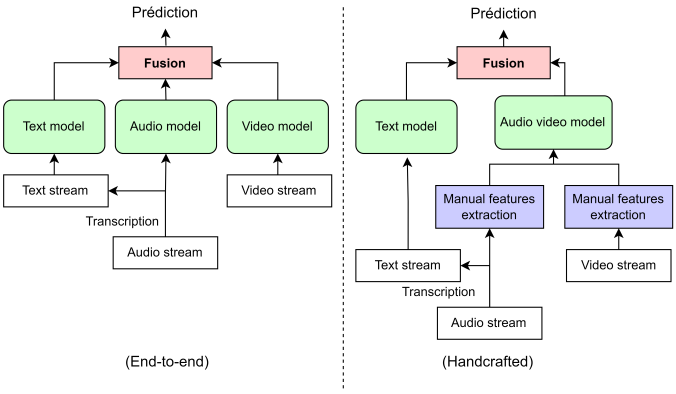

Fig. 2. The two approaches implemented: E2E (left) and H (right).

Ultimately, we implement the Hierarchical Attention Network (HAN) [20], which was originally designed for text document classifiers. We choose this architecture because it has the ability to focus on both word and sentence levels thanks to its attention mechanism.

We modify the original implementation by replacing the basic Gated Recurrent Unit (GRU) layer of the sentence level by a stateful GRU. This modification allows the model to keep track of the hidden state over time, improving the global performances.

The hyper-parameters of this model are tuned empirically. The input of the embedding is of size 500 which is the number of words the most represented in the dataset, and the output is of size 100. Each one of the two GRUs has 16 cells.

B. Handcrafted approach

This first approach consists of combining text and high-level audio-video hand-crafted features. We extract a total of 32 features with the text model and four features from the seven aforementioned raw handcrafted features.

1) Audio-video analysis: Among the seven features, two are extracted from the video stream. The first one, named ”eye contact”, is calculated using the extracted face with Dlib [27] and openCV [28] then hyperface [29] to generate the Euler angles of the head. This process is applied to each frame of the dataset. Finally, a K-means clustering algorithm on the Yaw and Pitch axis determines the couple of Euler angles when the driver is looking in the rear-view mirror. The tilt axis does not provide additional information in the car context.

For the ”passenger visibility”, we use Dlib and openCV to detect the face of the rear passenger on each frame. It is a binary feature, set at 1 if we detect the face of the rear passenger, 0 otherwise.

The five remaining features are the ones detailed in Section III-D: ”the mean talking” and ”the mean duration” for the driver and the passenger, and the ”mean silence” which is common to both of them.

Finally, these seven features fed a Multi-Layer Perceptron (MLP). It is designed with two hidden layers of four neurons each and one output layer generating the prediction.

2) Temporal fusion of our cues: Adding a temporal late fusion is necessary in our case because the stateful HAN is not sufficient to capture all the temporal information. The Perceptron model has no ability to capture temporality in the data. Furthermore, a late fusion is the usual strategy in case of non-heterogeneous modalities.

The fusion concatenates all features extracted from the three modalities (see Fig. 3). The unimodal models are modified to extract 32 features from the text and four from the audio-video model. It results, after concatenation, in a vector of size 36. Then, they are stacked for each analysing window of 35s to

 ISBN: 978-1-61208-942-3

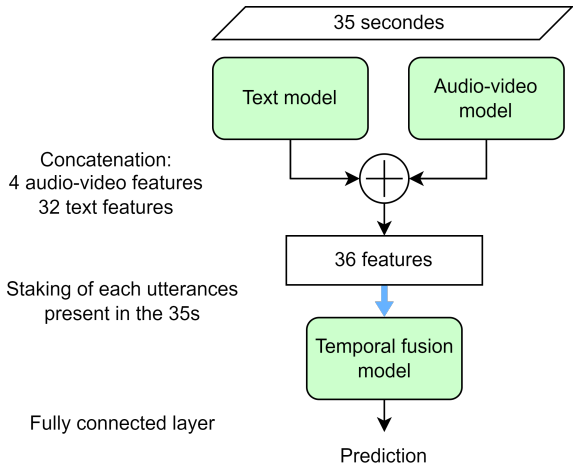

Fig. 3. The temporal fusion for the model H.

finally feed a stack of two stateful temporal Recurrent Neural Network (RNN) named GRU. See [8] for a complete review of the RNN. Finally, a Fully Connected (FC) layer predicts the label. The concept of stateful model is detailed in the Section IV-D.

C. End-to-end approach

This section details the end-to-end approach where the raw data are given in input of the model to directly make the prediction.

1) Audio analysis: In this new approach, we use OpenSMILE [7] with the configuration file emobase2010 [30] to automatically extract the features. We only use the audio stream of the passenger for more than one second. This process filters the interjections like: 'euh', 'ok', 'okay' and the repetitions of words very common in oral language. They are filtered because they do not increase the performances and can even degrade them. Moreover, the stress on the word is very weak in our corpus and more generally in the French language compared to the English language. The stress is more noticeable at an utterance level rather than at the word one in the French language [31]. OpenSMILE allows us to calculate the average value over a period of time. As a result, we have 1581 features per utterance. Finally, these features are stacked for each utterance in the window analysis and sent sequentially to a stateful recurrent network of two layers of 12 cells GRU. Then, a fully connected layer makes the final prediction. The matrix feeding the model is the size of the number of utterances in the 35s of the analysis window by the 1581 features.

2) Video analysis: Recall that the video modality is the least informative feature in our context and is also the least informative in the sentiment and emotion analysis literature [9] [11] [16]. In fact, only the head, the eyelid and the mouth movements give us information. This information is also limited because the driver must not deviate from his driving simulation task. As a reminder, the simulation task is a driving video in the first person view shown on the screen placed on the hood of the car. From our experience acquired with the MOSI corpus use [12], we first experiment the R3D approach [6]. The results were not conclusive. We then implement several other models with the ability to model the temporality:

- convLSTM [32],
- 2D Convolutional Neural Network (CNN) + LSTM,
- R3D + LSTM,
- optical flow [33] + R3D,
- optical flow [33] + 2D CNN + RNN.

All these architectures cited did not give satisfactory results. The models are able to converge during the training phase, but the results collapse during the validation phase. There are several hypotheses explaining this phenomenon: maybe an overfitting problem, an insufficient amount of data, or maybe the models are not able to catch the right features for the classification task.

These results lead us to test two last solutions. The first one uses a vector of 128 features to encode the driver's face in each image. We use the Dlib library [27] to extract them. Then, the features are stacked in 35s windows and sent to a GRU or LSTM model. The results are still not convincing.

The second method extracts the key points of the face. The principle is to retrieve 68 facial landmarks (the contour of the eyes, the face, the nose, etc.) defined by its image coordinates. They are computed for each frame using the Opencv and Dlib libraries. This approach gives good results compared to all the other implementations. We add the head orientation angles as described in Section III-D to improve the performance. Finally, a total of 142 facial features including the three head angles encode the driver's face.

Once the data are processed, they feed a neural network of two layers stateful GRU of four cells each. The matrix feeding the model is the size of the number of frames in the 35s of the analysis window by the 142 features.

3) Fully connected fusion of our cues: In this approach, the temporality is taken into account at the modality level (with the use of stateful GRU) in contrast to the handcrafted approach where it is at the fusion level. We modify the unimodal model to extract features. As each different modality does not have the same impact on the prediction performance, we empirically determine the number of features per modality to obtain the best performances. A total of 10 features are concatenated, four regarding text and audio features and 2 for the video one. Then, a fully connected (FC) layer built of 30 parameters makes the final prediction. See Fig. 4 for a representation of this approach.

D. Implementation Details

Setting the free parameters of the architecture and the training process are really important to deal with the multimodality and temporal context.

Foremost, we empirically set the sliding analyzing window to $T = 35s$.

During a dialog, the situation can evolve and catching this gradation gives a lot of information. As long as videos must be processed by algorithms with smaller analyzing windows, it is

 ISBN: 978-1-61208-942-3

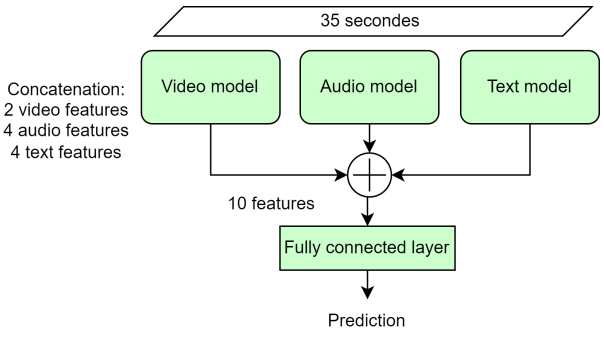

Fig. 4. The fully connected fusion for the model E2E.

important to keep track of the context between each analyzing window. We implement this concept by using stateful GRU. RNN only remembers what happened within a sequence. A sequence can be a set of sentences, a set of features, etc. At the initial time point of every passed sequence, the hidden states are initialized at 0, which means that the previous information is lost. In our approach, we initialize at each iteration the hidden state with the one generated at the previous analyzing window. This process keeps track of the evolution of all the features from the beginning to the end of the video.

Stateful RNN must be trained video by video. Each video is cut on the fly into approximately $180/35 = 5$ subsequence video clips. Then, they are fed chronologically one by one to the model. This training approach generates only $44 * 5 = 220$ training samples. In order to increase the training set, we shift the beginning of the analysis window to generate 400 samples. This consists in passing multiple times over each video. At each iteration, the starting point of the analyzing window is shifted.

Recall that the limit of our dataset forces us to discard the first 30s of our training samples. We delete on the fly these files during the training and validation phases.

The last method used to train the multimodal model is the pre-training techniques. All unimodal models are firstly trained to reach their best accuracy point and be saved. Then, at the beginning of the multimodal training phase, each previously saved model is loaded to initialize the multimodal one. This method is mandatory in our approach, otherwise the multimodal model is not able to converge. Freezing the weight of the loaded model (except for the fusion model) is considered, but it leads to poorer performance results.

On a multi-class problem, we use the cross entropy loss defined as in equation (1).

$$\text{loss}(\hat{y},\ \text{class}\ ) = -\log\left(\frac{\exp(\hat{y}[\ \text{class}\ ])}{\sum_i \exp(\hat{y}[i])}\right) \qquad (1)$$

where $\hat{y}$ is the output score of the model for the corresponding class.

V. Evaluations and associated analysis

In this section, we present the quantitative evaluations for both approaches and a qualitative analysis.

A. Quantitative Evaluations

When we work on behavior or emotion analysis, the speaker dependency is a key point. The idea is to evaluate the abilities of the algorithm to generalize when it deals with a new speaker. For this purpose, we generate five different cross-validation sets by selecting 80% of the speakers for the training phase and 20% for the validation phase. More specifically, there are in the train set: 36 videos representing 1620 utterances generated by 18 speakers and for the validation set: eight videos totaling 405 utterances generated by four speakers.

The balanced accuracy is defined in equation (2). It is mandatory when we do not have a balanced number of samples in each class.

$$\text{balanced-accuracy}\ (y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \sum_i 1\left(\hat{y}_i = y_i\right) \hat{w}_i \quad (2)$$

It is the macro-average of recalled scores per class i with associated weight $\hat{w}_i$ relative to the inverse prevalence of its true class y_i. The $\hat{y}_i$ is the inferred value of the sample i.

We obtain the following results (see Table I). This is the mean of the five cross-validation sets.

The results obtained with these two approaches are quite similar. Indeed, we obtain 81.6% of balanced accuracy with the end-to-end model and 81% with the handcrafted approach. The approach using the handcrafted (H) features is more consistent with a standard deviation below the end-to-end approach (E2E). The (H) architecture does not contain enough parameters on the video or audio alone to allow a classification by modality. The standard deviation of (E2E) is likely due to the audio and video modalities which are difficult to exploit. As a reminder, we were able to obtain convincing results by using only the rear passenger data for audio and the driver's face for video. Additionally, the literature [34] shows the existence of a threshold in the quantity of data for which the end-to-end approaches outperform classical approaches (statistical, machine learning, etc.). Below this threshold, the classical techniques obtain the same or better performances as the end-to-end one. Our results and the size of our corpus seem to indicate that we are in this situation. Increasing the amount of the data could therefore solve the issue.

The case of speaker dependency in the selection of train/test data is the least favorable for a neural network and not representative of real-world applications. Indeed, in the case

TABLE I
Results and comparison of the two approaches. SD refers to Speaker Dependant.

Model	Modalities	Balanced accuracy
End-to-end (E2E)	Video	65.6% ± 4
	Audio	70.6% ± 4,9
	Text	70% ± 0.8
	Audio + video	61% ± 3.9
	Video + Audio + Text	**81,6% ± 5.9**
	Video + Audio + Text SD	***88.2%***
Handcrafted (H)	Text	70% ± 0.8
	Audio + Video	60% ± 1.12
	Video + Audio + Text	**81% ± 1.2**

of a smartphone assistant, a "world" model is specified to a new user with a new training phase based on some samples of his voice. For example, on a new Android device, when the user starts to use Google assistant, it asks the user to repeat a few times "Ok Google". To reproduce this configuration, we train our end-to-end model on the first 90s of each video and test it on the remaining 90s. This approach gives a balanced accuracy of 88.2%. It shows the benefit of a specialization phase and partially shows our shortage of data.

B. Quantitative evaluations and Juxtaposition of the two approaches

After analyzing the miss-classified files, we observe some issues leading to these miss-classification. A few participants did not play their role in adequacy with the asked scenario or they took a very long time to engage in the discussion. Two specific issues also lead to bad results: (i) on one video the voices of the two speakers are very low compared to the other's recording; (ii) on another video, the driver has a very bad posture resulting in a half-visible face.

Typically, in literature, the performances of the different approaches are compared. We argue that the two presented approaches are complementary. If we examine the wrong/right classifications of each model for the same test file, we notice that errors are not made on the same analysis window. These two models can be complementary in their decision-making.

Fig. 5 shows the confusion matrices for our two models on the same cross-validation file. The end-to-end approach has a better ability to classify the videos of the "curious" and "categorical refusal" classes which are the most opposed classes. The handcrafted based model performs well on the "argued refusal" class.

In our application context, one hybrid solution could be to use the embedded model, *i.e.* Handcrafted (H), to establish a first diagnosis of the situation and then send the video data to a cloud platform to inference with the end-to-end model. This choice lead to reduce the cost of data transmission to a cloud platform.

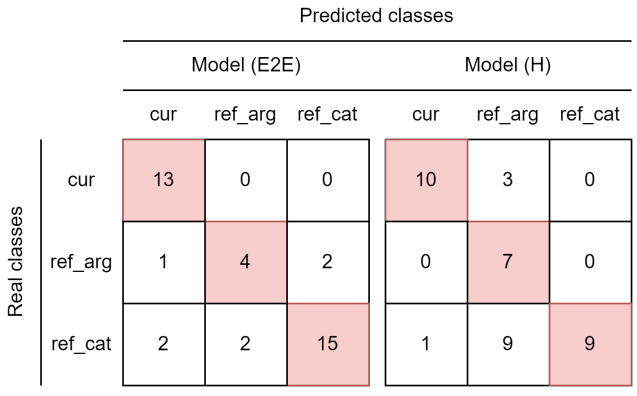

Fig. 5. Comparison of the two confusion matrix for a same cross-validation set. (H) refers to the handcrafted model and (E2E) to the end-to-end one. cur denotes the "curious" class, ref_arg describe the "argued refusal" and ref_cat stands for the "not argued refusal" class.

VI. Conclusion

This paper compares two multimodal approaches for the analysis of interaction in a real vehicle context. The performances obtained with these models are promising with 81% of balanced accuracy for the handcrafted model and 81.6% for the end-to-end approach. We also show the benefits of the fusion of different modalities and the complementary of our two approaches.

The embeddability capability of neural networks and the real application context is often omitted in the literature and even more in multimodal systems. Our future work will focus on integrating the two detailed architectures on a specific automotive platform to evaluate the embedding performances.

Acknowledgment

This work has been carried out under the funding of an industrial doctorates fellowship from the National Association for Research and Technology (ANRT), France.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[2] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, E. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a Tasty French Language Model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7203–7219.

[3] B. Santra, P. Anusha, and P. Goyal, "Hierarchical Transformer for Task Oriented Dialog Systems," *arXiv:2011.08067 [cs]*, Mar. 2021.

[4] D. Chen, H. Chen, Y. Yang, A. Lin, and Z. Yu, "Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems," *arXiv:2104.00783 [cs]*, Apr. 2021.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015.

[6] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," *arXiv:1708.07632 [cs]*, 2017.

[7] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile – the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 01 2010, pp. 1459–1462.

[8] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.

[9] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 873–883. [Online]. Available: http://aclweb.org/anthology/P17-1081

[10] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "An ensemble approach to utterance level multimodal sentiment analysis," in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018, pp. 145–150.

[11] A. Agarwal, A. Yadav, and D. K. Vishwakarma, "Multimodal sentiment analysis via rnn variants," in *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science Engineering (BCD)*, 2019, pp. 19–23.

[12] Q. Portes., J. Carvalho., J. Pinquier., and F. Lerasle., "Multimodal neural network for sentiment analysis in embedded systems," in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, INSTICC. SciTePress, 2021, pp. 387–398.

 ISBN: 978-1-61208-942-3

[13] Q. Portes, J. Pinquier, F. Lerasle, and J. M. Carvalho, "Multimodal human interaction analysis in vehicle cockpit," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 2118–2124.

[14] R. Li, C. Lin, M. Collinson, X. Li, and G. Chen, "A dual-attention hierarchical recurrent neural network for dialogue act classification," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 383–392.

[15] Y. Luan, Y. Ji, and M. Ostendorf, "Lstm based conversation models," 2016.

[16] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. B. V. Subramaanyam, "Benchmarking multimodal sentiment analysis," *arXiv:1707.09538 [cs]*, 2017.

[17] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016.

[18] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional Emotional Recurrent Unit for Conversational Sentiment Analysis," *arXiv:2006.00492 [cs]*, Feb. 2021.

[19] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6818–6825, Jul. 2019.

[20] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1480–1489.

[21] Q. Portes, J. Carvalho, J. Pinquier, and F. Lerasle, "Multimodal neural network for sentiment analysis in embedded systems," in *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, INSTICC. SciTePress, 2021, pp. 387–398.

[22] B. Bigot, J. Pinquier, I. Ferrané, and R. André-Obrecht, "Looking for relevant features for speaker role recognition (regular paper)," in *INTERSPEECH, Makuhari, Japan, 26/09/10-30/09/10*. http://www.isca-speech.org/: International Speech Communication Association (ISCA), 2010, pp. 1057–1060.

[23] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

[24] M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1373–1378. [Online]. Available: https://aclweb.org/anthology/D/D15/D15-1162

[25] E. Loper and S. Bird, "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ser. ETMTNLP '02. Association for Computational Linguistics, 2002, p. 63–70. [Online]. Available: https://doi.org/10.3115/1118108.1118117

[26] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 7203–7219.

[27] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, p. 1755–1758, dec 2009.

[28] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[29] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[30] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," *INTERSPEECH 2010*, p. 4, 2010.

[31] J. Vaissière, "Cross-linguistic prosodic transcription: French vs. English," in *Problems and methods of experimental phonetics. In honour of the 70th anniversary of Pr. L.V. Bondarko*, N. Volskaya, N. Svetozarova, and P. Skrelin, Eds. St Petersburg State University Press, 2002, pp. 147–164.

[32] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. kin Wong, and W. chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," 2015.

[33] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2017/IMKDB17

[34] M. Z. Alom, T. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. Nasrin, M. Hasan, B. Essen, A. Awwal, and V. Asari, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, 03 2019.