Analyse multimodale d'interaction humaine dans le cockpit d'un véhicule

Quentin Portes, Julien Pinquier, Frédéric Lerasle, Jose Mendes-Carlalho

Analyse multimodale d'interaction humaine dans le cockpit d'un véhicule

Quentin Portes[1,2,3] Julien Pinquier[2] Frédéric Lerasle[3] José Mendes Carvalho[1]

[1] Renault Software Lab, Toulouse, France
[2] IRIT, Université Paul Sabatier, CNRS, Toulouse, France
[3] LAAS-CNRS, Université Paul Sabatier, Toulouse, France

quentin.q.portes@renault.com

Résumé

Aujourd'hui, les constructeurs automobiles se concentrent sur l'avenir de la mobilité. Les véhicules électriques, les véhicules autonomes et les véhicules partagés sont les opportunités les plus prometteuses. Le manque d'autorité dans les véhicules partagés soulève différents problèmes comme la sécurité des passagers. Pour garantir cette dernière, il faut concevoir de nouveaux systèmes capables de comprendre les interactions et les conflits éventuels entre les passagers, avant qu'une situation critique ne se produise dans le cockpit. Afin de mieux comprendre les caractéristiques de ces situations d'insécurité, nous avons enregistré un corpus audio-vidéo dans un contexte de véhicule réel. Vingt-deux participants jouant trois différents scénarios (« curieux », « refus argumenté » et « refus non argumenté ») d'interactions entre un conducteur et un passager ont été enregistrés. Notre approche, basée sur de l'apprentissage profond, atteint une précision de 81%. Nous démontrons que la combinaison des modalités vidéo/audio/texte et la temporalité permettent d'améliorer les prédictions de reconnaissance de scénarios in situ.

Mots Clef

Réseau de neurones multimodaux, fusion de données, application automobile, classification supervisée.

Abstract

Nowadays, every car maker is thinking about the future of mobility. Electric vehicles, autonomous vehicles and sharing vehicles are the most promising opportunities. The lack of control authority in autonomous and sharing vehicles raises different issues like the passenger safety. To ensure it, new systems able to understand interactions and possible conflicts between passengers have to be designed. They should be able to predict and trigger with high accuracy, an alert to a remote controller before a critical situation happens in the cockpit. In order to better understand the features of these insecure situations, we recorded an audio-video dataset in real vehicle context. 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 a deep learning approach which achieves a balanced accuracy of 81%. Practically, we highlight that combining multimodality, namely video, audio and text as well as temporality are the keys to perform such accurate predictions in scenario recognition.

Keywords

Multimodal neural network, automotive application, sensor data fusion, supervised classification.

1 Introduction

Les dialogues, les interactions, les émotions et l'analyse des sentiments sont les principaux éléments pour comprendre les interactions humaines. La capture de ces informations, quel que soit le contexte d'application final, pourrait résoudre des problèmes industriels tels que le filtrage des contenus sensibles sur les réseaux sociaux ou l'amélioration de la compréhension des interfaces homme-machine. Pour l'avenir de l'industrie automobile, l'analyse de l'habitacle est une problématique importante. En effet, elle permettra de répondre à différentes questions de sécurité liées aux nouveaux usages de la voiture *i.e.* socialisation, partage de véhicule, voitures autonomes, etc. Plus précisément, deux problèmes de sécurité sont soulevés : le manque d'autorité dû à la non-présence d'un conducteur et le partage d'un véhicule avec un inconnu. Ces circonstances pourraient conduire à des situations telles que des railleries, des brimades ou, dans le pire des cas, des agressions. Ces problèmes de sécurité doivent être anticipés et évités. Les constructeurs automobiles doivent être proactifs dans de telles circonstances.

Pour anticiper ces situations, il est nécessaire d'analyser les interactions entre les passagers à l'aide de caméras et de microphones positionnés dans l'habitacle. Les modalités vidéo, audio et texte fournissent des informations qui, une fois fusionnées, pourraient prédire avec une meilleure précision l'apparition de situations conflictuelles.

Dans cette optique, les récents réseaux de neurones profonds et le succès de l'architecture *Transformer* constituent des avancées majeures. Le modèle BERT [7] (langue an-

glaise), les modèles Roberta et CamemBERT [18] (langue française) ont amélioré les performances globales des tâches de réponse aux questions, de résumé de texte, de traduction, etc. Des travaux récents ont également appliqué le modèle transformateur à l'analyse de dialogues textuels [22, 6]. Ces approches restent basées sur la modalité texte.

Aujourd'hui, les modèles d'analyse vidéo sont capables de capter les informations de mouvement ; citons ici les architectures 3D-CNN (C3D) [23] et 3D-CNN résiduel (R3D) [9].

Concernant l'analyse audio, l'approche la plus courante consiste à extraire des caractéristiques audio sur une courte fenêtre glissante grâce à des d'outils comme open SMILE [8]. Ces caractéristiques sont classiquement transmises à un modèle séquentiel, type LSTM [25].

Pour améliorer les prédictions des modèles d'analyse de situations, une approche évidente pourrait être l'analyse de l'interaction audio-vidéo ajoutée à l'utilisation du texte transcrit à partir du flux audio. Cette approche semble plus prometteuse que les modalités vidéo et audio prises séparément. Nombreux travaux *e.g.* [19, 11, 1] ont, à ce titre, conclu que la multimodalité est toujours plus performante.

Le domaine de l'automobile, en particulier l'habitacle, est par nature un environnement peu contrôlé : l'exposition au soleil et le bruit audio généré par les vibrations de la route ou des autres voitures en mouvement sont autant d'interférences qui peuvent réduire les performances du système perceptuel. La multimodalité peut alors améliorer les performances globales et accroître la robustesse des modèles.

Cependant, les trois défis automobiles identifiés dans l'analyse des interactions multimodales sont les suivants :

- la disponibilité d'un ensemble de données publiques et *in situ*,
- la fusion entre des modalités non hétérogènes comme la vidéo, l'audio et le texte,
- l'interprétation d'interactions humaines, par définition complexes et évolutives dans le temps.

Aujourd'hui la littérature, contrairement aux travaux présentés ici, n'aborde pas, à notre connaissance, tous ces verrous en même temps.

Nos travaux se concentrent sur l'enregistrement d'un corpus exploitable pour des applications automobiles et sur la conception d'une première approche multimodale. Nous nous différencions de la littérature par notre jeu de données réalistes car *in situ* et notre stratégie de fusion multimodale.

L'article est structuré comme suit. La section 2 présente un état de l'art non exhaustif sur l'analyse du dialogue multimodal. La section 3 détaille le protocole d'enregistrement de notre propre jeu de données et ses spécificités. La section 4 développe notre approche multimodale pour la classification des interactions humaines.

2 État de l'art

Dans la littérature, la plupart des analyses de dialogue, d'interaction et de conversation sont basées sur le texte [13, 16]. Les récentes investigations, avec de nouvelles approches telles que la multimodalité, montrent les avantages de l'utilisation d'informations provenant de différents canaux. Toutes les architectures multimodales surpassent les architectures unimodales [19, 5].

Ces méthodes se fondent sur la fusion de paramètres extraits de chacune des modalités (vidéo, audio et texte). Ensuite, une stratégie de fusion tardive plus ou moins complexe est appliquée. Nous avons identifié quelques travaux récents sur l'analyse de dialogues multimodaux : [14, 17]. Ils se concentrent sur l'analyse des sentiments et des émotions dans les conversations. Tous ces travaux sont basés sur des corpus publics comme MOSI [26].

L'architecture HAN (Hierarchical Attention Network) [24] et les modèles *Transformers* sont aujourd'hui très performants pour l'analyse de documents. Des approches récentes, telles que [22], utilisent le *Transformer* pour l'analyse de dialogues. Notre corpus étant limité en taille et composé essentiellement de texte oral, l'approche HAN semble la plus adaptée.

Dans l'analyse des interactions, les comportements antérieurs du locuteur sont très importants pour comprendre avec plus de précision ses comportements futurs et présents. Aujourd'hui, les modèles d'apprentissage profond ne sont pas en mesure de retenir l'information sur de très longues durées de vidéos. L'utilisation de modèles temporels complets (*statefull* [25]) est une solution pour pallier ce problème.

Dans le domaine de l'analyse des interactions entre passagers dans l'habitacle, les investigations sont rares et restent donc un défi scientifique.

3 Corpus multimodal de dialogue en contexte véhicule

Cette section détaille le protocole utilisé pour enregistrer notre corpus multimodal.

3.1 Caractéristiques du jeu de données

Le corpus vise à enregistrer les interactions entre deux passagers dans l'habitacle d'une voiture. Un conducteur et un passager arrière (côté droit) jouent des scénarios prédéfinis. Les sujets sont des volontaires français sans aucune compétence d'acteur.

Une session d'enregistrement dure 7 minutes par participant et chaque session est découpée en quatre étapes continues :

1. 60s de silence,
2. **180s de rôle/interaction**,
3. 60s de silence,
4. 120s d'interaction avec l'écran du véhicule.

Nos travaux se focalisent sur la phase 2 (rôle/interaction). Pendant cette phase, le conducteur joue toujours le même comportement de vendeur insistant et le passager joue l'un des trois comportements suivants :

— « être curieux de la proposition du conducteur »,
— « refuser la proposition avec argumentation »,
— « refuser catégoriquement la proposition ».

Le conducteur ne connaît pas *a priori* le comportement du passager. Il subit la situation. Le scénario de vendeur insistant a été choisi à la place d'un scénario d'agression pour des raisons protocolaires. En effet, si nous avions voulu jouer des scénarios d'agressions réalistes, nous aurions été obligés de suivre un protocole de suivi psychologique pour chacun des participants.

3.2 Plateforme sensorielle

Le dispositif d'enregistrement multi-capteurs (voir figure 1) est composé de six caméras, quatre microphones et d'un écran posé sur le capot d'un Renault Dacia Duster à l'arrêt. L'écran est en face du conducteur et également visible par le passager. Il a deux objectifs : le premier est d'indiquer quand il doit changer de phase, le second est de diffuser une vidéo de la route pour captiver l'attention du conducteur car la voiture est statique. Toutes les interactions avec la voiture sont autorisées (volant, levier de vitesse, etc.).

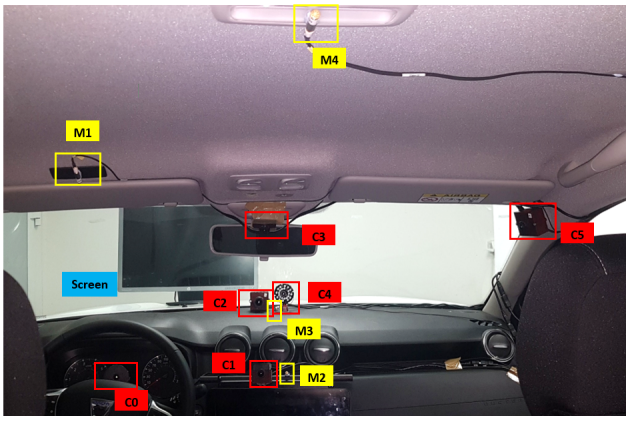

FIGURE 1 – Vue intérieure de l'habitacle de la voiture avec le matériel d'enregistrement et les capteurs.

Flux vidéo. Trois caméras sont déployées. Elles diffèrent par leur résolution, leur angle de vue et leur focale. Notre approche privilégie la caméra #2 car elle présente la meilleure qualité d'image et d'éclairage (voir figure 2). Les autres caméras ne sont pas exploitées dans cette étude. Il s'agit d'une caméra avec mise au point manuelle et de résolution 1920×1080 pixels. Elle est positionnée de manière à avoir un angle de vue frontal (ID = C2 sur la figure 1).

Flux audio. Quatre microphones identiques sont placés à différents emplacements de l'habitacle et enregistrent le flux audio. Notre approche utilise uniquement le microphone plafonnier du conducteur (ID = M1 sur la figure 1).
Remarque : les flux vidéo et audio sont sauvegardés au format brut (pas de compression en direct) pour ne pas perdre en qualité.

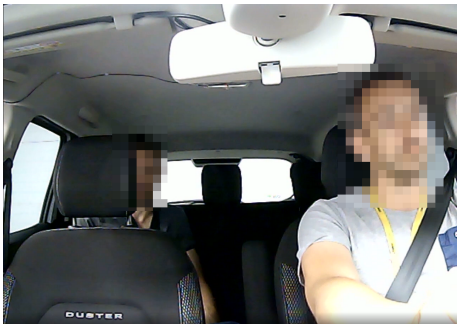

FIGURE 2 – Champ d'observation de la caméra #2.

3.3 Préparation et annotation du corpus

Un post-traitement est inévitable, le processus d'enregistrement générant un décalage temporel entre les flux vidéo et audio. Afin de les synchroniser, nous utilisons Adobe premiere pro.
Pour obtenir la troisième modalité, nous transcrivons le texte à partir du flux audio. Après quelques expérimentations infructueuses (taux d'erreur trop important), nous évitons, pour l'instant, la transcription automatique de la parole (Automatic Speech Transcription, ASR). En effet, la spécificité du contexte oral (répétitions, interjections et mots isolés) ainsi que l'éventuelle mal construction des phrases (sujet-verbe-complément) limitent les performances de l'ASR. Pour transcrire le texte, nous utilisons le logiciel ELAN [1]. Il s'agit d'un outil d'annotation manuelle conçu pour créer, éditer, visualiser des annotations pour des données vidéo et audio. Nous transcrivons le flux audio de chaque acteur en texte, ce qui donne un total de 2026 *tours de parole*. Pour rappel, un *tour de parole* est une unité continue de discours commençant et se concluant par une pause explicite.
Nous annotons la vidéo entière, contrairement à d'autres corpus [26] où les annotations sont faites au niveau du *tour de parole*. Notre choix a pour conséquence de générer des étiquettes erronées si les passagers jouent mal leur rôle. Nous reviendrons sur ces réserves dans l'analyse qualitative des évaluations (section 5.2).

3.4 Spécifications/compréhension du corpus

Notre corpus se compose de 44 vidéos, pour 22 participants (4 femmes et 18 hommes). Chaque participant joue une fois en tant que conducteur et une fois en tant que passager dans un ordre aléatoire. La somme de toutes les interactions donne 2026 tours de parole ; cela représente environ 22k mots (2082 mots uniques). Nous avons un total de 1h48 de vidéo, soit 54 min pour la classe curieuse, 27 min pour la classe refus argumenté et 27 min pour la classe refus non argumenté.
En pratique, nous observons que la modalité vidéo est moins informative que les modalités audio et texte. Dans un contexte automobile, cela s'explique par les mouvements

1. https://archive.mpi.nl/tla/elan

limités des passagers et par la concentration du conducteur sur la conduite. Ce constat est aussi valable dans d'autres contextes multimodaux comme l'analyse de sentiments ou de dialogues [20, 5, 5].
Une étude préliminaire montre des comportements récurrents. Comme les humains ne changent pas leurs émotions ou leurs comportements toutes les secondes, nous avons tracé les caractéristiques en fonction du temps pour une fenêtre d'analyse de 15s. Ce lien Github [2] regroupe tous les graphiques.
Notre première intuition pour identifier ces descripteurs locaux est inspirée de [3].
Après avoir inspecté les flux audio-vidéo et analysé ces graphiques, nous avons isolé les caractéristiques ci-dessous :

- la prise de parole moyenne. Dans une conversation normale, les durées de prise de parole tendent à être équiréparties entre les participants.
- la durée moyenne du temps de parole est la longueur moyenne d'un *tour de parole*. Dans un contexte d'interaction, la durée d'un discours est un bon indicateur de qui domine la conversation et de qui veut la clore.
- le silence moyen est un indicateur de l'intensité d'un dialogue. Plus il y a de silence, plus la discussion est pauvre et tend vers une situation de refus.
- le contact visuel du conducteur est la fréquence à laquelle le conducteur regarde dans le rétroviseur intérieur. Comme il est concentré sur sa tâche de conduite, il n'a pas d'autre choix que de regarder dans le rétroviseur pour voir son interlocuteur.
- la visibilité du passager est la fréquence à laquelle le passager est vu par la caméra. C'est un bon indicateur pour savoir si le passager est intéressé par la conversation. Nous réduisons naturellement la distance avec notre interlocuteur lorsque nous sommes engagés dans une discussion. Dans le cockpit, le passager arrière se balance (ou non) entre les deux banquettes avant-arrière. Ces balancements corporels sont alors observées par la caméra.

En ce qui concerne la modalité texte, nous traçons la fréquence des mots et le TF-IDF (Term Frequency-Inverse Document Frequency, [12]) pour trouver s'il existe des distributions spécifiques de mots associées à un scénario donné. Ces approches sont très courantes dans l'exploration et l'analyse de textes. Nous calculons aussi le delta TF-IDF absolu entre les deux classes opposées (« curieux » et « refus non argumenté »). Nous obtenons les 10 mots les plus importants suivants : *je, pas, vous, ouais, tu, non, moi, oui, donc* et l'interjection *ah*. Notre modalité textuelle n'est hélas pas riche ; pour rappel nous n'avons que 2082 mots différents.
Les participants n'étant pas de véritables acteurs et leurs dialogues n'étant pas prédéfinis, nous avons observé deux transitions. La première est la mise en place du scénario : les sujets ne pouvaient pas être insistants ou catégoriques dans leur refus dès les 30 premières secondes. La seconde est celle de la fin : les sujets ont manqué d'inspiration, provoquant un essoufflement du scénario pendant les 20 dernières secondes.

2. https://github.com/QuentinPrts/ITSC_2021

4 Analyse multimodale

Nous proposons une approche multimodale basée sur l'analyse vidéo, audio et texte. La tâche consiste à concevoir un modèle capable de classer les flux audio-vidéo en trois classes : « curieux », « refus argumenté », « refus non argumenté ».

4.1 Analyse des flux audio et vidéo

Notre approche consiste à extraire des paramètres de haut niveau pour la modalité audio et vidéo. Le contexte automobile a l'avantage que la position des passagers est statique. Nous pouvons exploiter cette information pour savoir où se trouvent les passagers dans la vidéo. Si nous coupons l'image au milieu de son axe horizontal, nous avons le conducteur sur le côté droit et le passager arrière sur le côté gauche. Pour extraire le paramètre « contact visuel du conducteur », nous utilisons openCV [4] comme détecteur de visage suivi de l'extracteur d'angle d'Euler open source hyperface [21]. Enfin, un algorithme de k-means clustering [2] sur les axes Yaw et Pitch trouve le couple d'angles d'Euler lorsque le conducteur regarde dans le rétroviseur (couleur verte sur la figure 3). L'axe Tilt n'est pas informatif ici.

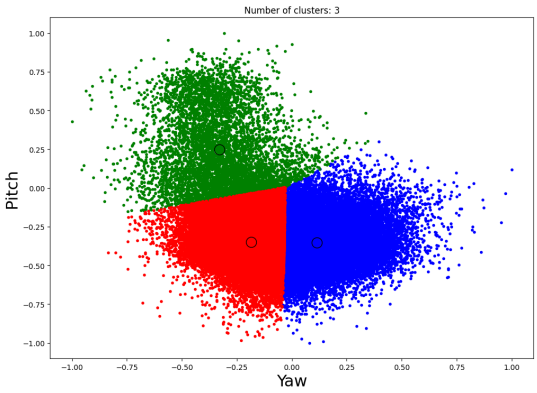

FIGURE 3 – Clustering des angles d'orientation de la tête du conducteur.

Pour la perception du passager sur le siège arrière, nous utilisons à nouveau openCV pour détecter le visage du passager arrière, sur chaque image.
Ensuite, nous réalignons au niveau du *tour de parole* les paramètres audio et vidéo, car le modèle sera alimenté par les trois modalités parfaitement alignées.
Pour vérifier nos choix pour les paramètres sélectionnés précédemment nous calculons la matrice de corrélation de Pearson (1) pour toutes les caractéristiques susmentionnées

avec pour objectif de mettre en évidence des corrélations linéaires entre les paires $(X,\ Y)$ de caractéristiques.

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

où $\sigma_{(.)}$ désigne l'écart-type ; $\mu_{(.)}$ est la moyenne et $\mathbb{E}(.)$ désigne l'espérance mathématique.

Les tableaux 1 et 2 exhibent des corrélations évidentes entre les caractéristiques audio et vidéo. La première est l'augmentation du contact visuel entre le conducteur et le passager avec l'augmentation de la visibilité du passager. La seconde est la moyenne du silence qui diminue avec les contacts visuels. Ces corrélations confirment l'existence d'un lien entre la vidéo et l'audio dans l'interaction humaine.

Les acronymes suivants sont définis pour les sept caractéristiques : Msp désigne la durée moyenne des interactions, Mdur la durée moyenne du temps de parole, eyeC le contact visuel, Msil la moyenne de silence et Pvisi la visibilité du passager.

TABLE 1 – Corrélations de Pearson pour le conducteur.

Conducteur	Msp	Mdur	eyeC	Msil	Pvisi
Msp	1	-	-	-	-
Mdur	0.5	1	-	-	-
eyeC	0.37	-0.03	1	-	-
Msil	-0.56	-0.21	**-0.37**	1	-
Pvisi	0.4	0.07	**0.84**	-0.23	1

TABLE 2 – Corrélations de Pearson pour le passager.

Passager	Msp	Mdur	eyeC	Msil	Pvisi
Msp	1	-	-	-	-
Mdur	0.74	1	-	-	-
eyeC	-0.14	-0.22	1	-	-
Msil	0.22	0.19	**-0.37**	1	-
Pvisi	-0.21	-0.18	**0.84**	-0.23	1

Finalement, ces sept caractéristiques sont envoyées à un perceptron multicouche (MLP). Il est conçu avec deux couches cachées et une couche de sortie générant quatre caractéristiques.

4.2 Analyse du texte

L'analyse du texte nous confronte à trois problèmes majeurs. Le premier est relatif à l'usage de la langue française. En effet, tous les systèmes et modèles pré-entraînés tels Spacy [10], NLTK [15], BERT [7] sont adaptés à l'analyse de l'anglais, mais sont peu performants sur la langue française. Les alternatives existantes pour la langue française sont limitées à des modèles entraînés sur de l'ancien français ou du français écrit. Nous obtenons donc de très mauvais résultats sur le modèle *Transformer* nommé CamenBERT [18] qui est entraîné sur 139 Go de texte Wikipedia. La faible richesse du texte de notre corpus rend inefficace les deux approches de base (TF-IDF et *embedding* + modèle LSTM).

Finalement nous utilisons le réseau d'attention hiérarchique (HAN) [24], qui est initialement conçu pour classer des documents textes. Nous avons choisi cette architecture car, grâce au double mécanisme d'attention, elle a la capacité de se concentrer à la fois sur l'importance des mots et des phrases.

Son implémentation originale est modifiée en remplaçant la couche GRU analysant les phrases par une *statefull* GRU. Cette modification permet au modèle de garder la trace temporelle des phrases précédentes, améliorant ainsi les performances globales.

Les hyper-paramètres de ce modèle sont fixés empiriquement :

- *l'embedding* est constituée des 500 mots les plus représentés dans le corpus. La sortie est un vecteur de taille 100,
- 64 cellules pour le GRU des mots et le GRU des phrases,
- un vecteur de taille 100 pour la sortie de la couche GRU analysant les mots.

4.3 Fusion des modalités

Cette section détaille notre approche de fusion tardive basée sur l'audio, la vidéo, le texte et l'évolution temporelle. La fusion tardive est la stratégie habituelle en cas de modalités hétérogènes. La figure 4 décrit notre modèle. La partie verte fait référence à l'extraction de caractéristiques de chacune des modalités et la partie orange à la fusion temporelle de ces paramètres.

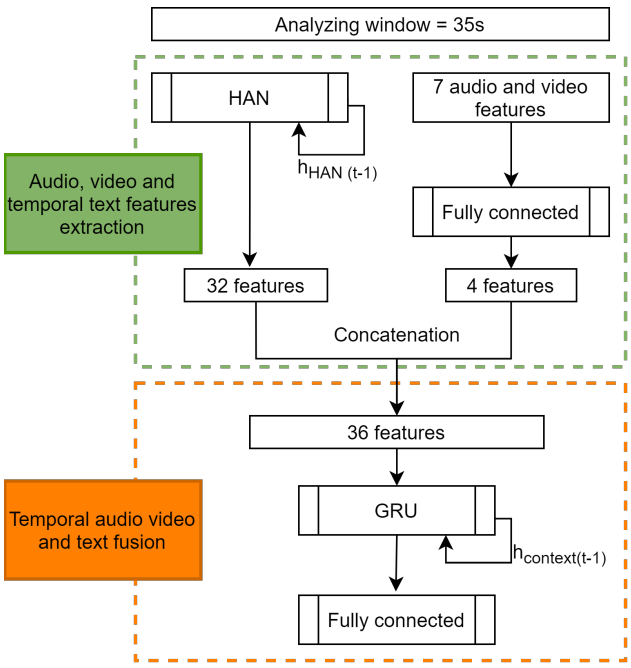

FIGURE 4 – Notre architecture de fusion multimodale.

La fusion combine les paramètres extraits des trois modalités. Les 32 premiers paramètres sont extraits du texte à l'aide du modèle HAN et les 4 restants sont extraits des 7 paramètres définis dans la section 3.4. Il en résulte, après concaténation, un vecteur de taille 36. Ensuite, ce vecteur est envoyé à deux *statefull* GRU empilés. Finalement, nous alimentons une couche entièrement connectée (FC) pour obtenir la prédiction du scénario. Le concept *statefull* est détaillé dans la section suivante.

4.4 Implémentation

Le tuning des hyper-paramètres est vital lorsque nous travaillons sur la multimodalité et le contexte temporel, comme dans notre contexte applicatif.

Dans nos expérimentations, la fenêtre d'analyse glissante est fixée à $T = 35$s. Celle-ci montre empiriquement les meilleurs résultats sur la modalité texte pour différents ensembles de validation.

L'évolution du dialogue donne des informations primordiales qu'il est nécessaire de réussir à capturer. Humainement, il est plus facile de comprendre une situation si nous disposons de plusieurs fenêtres d'analyse chronologiques plutôt que de fenêtres mélangées. Nous mettons en œuvre ce concept en utilisant le *statefull* GRU. L'architecture basique d'un RNN mémorise uniquement les informations dans une séquence. Une séquence peut être une phrase, un ensemble de caractéristiques, etc. À chaque nouvelle séquence, les états cachés sont initialisés à zéro, ce qui signifie qu'il n'y a aucune information précédente. Dans notre approche, nous remplaçons l'initialisation à zéro par l'état caché de l'itération précédente. Appliqué dans la fusion, il garde la trace de l'évolution de toutes les caractéristiques du début à la fin de la vidéo.

Le *statefull* RNN doit être entraîné vidéo par vidéo. Chaque vidéo est découpée à la volée en environ 180/35 = 5 clips. Ensuite, elles sont envoyées chronologiquement une par une au modèle. Cette méthode génère seulement $44 * 5 = 220$ échantillons d'apprentissage. Afin d'augmenter cet ensemble, nous décalons le début de la fenêtre d'analyse pour générer 400 échantillons. Ce décalage consiste à passer quatre fois sur chaque vidéo, à chaque itération le point de départ de la fenêtre d'analyse est décalé de 10s.

Comme évoqué, les limites des jeux d'acteurs nous obligent à ne pas considérer les 30 premières secondes de nos échantillons d'entraînement. Nous les avons donc supprimés lors des phases d'entraînement et de validation.

Afin d'entraîner le modèle multimodal, nous utilisons des techniques de pré-entraînement. Le modèle HAN est d'abord entraîné pendant environ 80 époques sur les données textes. Ensuite, lorsqu'il atteint sa meilleure précision, il est sauvegardé. Enfin, au début de la phase d'entraînement multimodal, le modèle HAN sauvegardé est chargé pour initialiser les poids de l'architecture multimodale.

Pour notre problème multi-classes, nous privilégions la perte d'entropie croisée définie par l'équation (2).

$$\text{loss}(\hat{y},\ \text{class}\ ) = -\log\left(\frac{\exp(\hat{y}[\ \text{class}\ ])}{\sum_i \exp(\hat{y}[i])}\right) \qquad (2)$$

où $\hat{y}$ est la prédiction du modèle pour la classe C.

5 Évaluations et analyses associées

Cette section présente nos évaluations quantitatives puis une analyse qualitative basée sur quelques inférences.

5.1 Évaluations quantitatives

Un point clé, lorsque nous travaillons sur l'analyse du comportement ou des émotions, est la dépendance au locuteur. L'intérêt est d'évaluer la capacité de l'algorithme à généraliser lorsqu'il traite un nouveau locuteur. Nous générons aléatoirement cinq fichiers d'entraînement/validation différents. À chaque fois, nous divisons le jeu de données en 80% (18 participants) et 20% (4 participants). Nous utilisons la micro précision comme métrique pour évaluer notre modèle. La micro précision est définie par l'équation 3. Elle est obligatoire lorsque nous n'avons pas une équipartition d'échantillons dans chaque classe.

$$\text{micro précision}(y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \sum_i 1\left(\hat{y}_i = y_i\right) \hat{w}_i \quad (3)$$

Il s'agit de la micro-moyenne du rappel par classe i à laquelle est associée un poids $\hat{w}_i$ relatif à la prévalence inverse de sa vraie classe y_i. La variable $\hat{y}_i$ est la valeur inférée de l'échantillon i.
Le tableau 3 synthétise nos évaluations. Le modèle audio/vidéo obtient une micro précision de 60%, résultat prometteur compte tenu de la taille du modèle et du nombre restreint de paramètres. Le modèle textuel obtient une micro précision de 70%. Notre approche de fusion donne de très bons résultats : elle améliore la micro précision de 11% pour atteindre une micro précision finale de 81%. L'écart type est la moyenne des cinq écarts types induits par la stratégie de validation croisée.

TABLE 3 – Performances moyennes sur cinq ensembles de validation croisée.

Modalité	Micro précision
Audio + Vidéo	60% ± 1,12
Texte	70% ± 0,8
Audio + Vidéo + Texte	**81%** ± 1,2

La figure 5 illustre un exemple d'un ensemble de validation. La métrique tracée est la micro précision en fonction du temps. Plus précisément, il s'agit de la mico précision sur les fichiers présents au moment t dans la fenêtre $T = 35s$. Lorsque le modèle prend en compte 90% de la vidéo, il est capable de classifier avec 99% de précision.

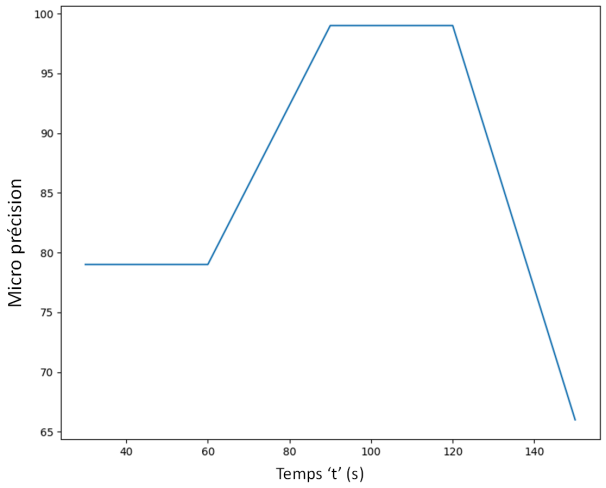

FIGURE 5 – Exemple de micro précision *vs.* temps pris dans une vidéo.

5.2 Évaluations qualitatives

Une étude sur quelques clips aboutit au constat suivant : la principale limite se situe au niveau du jeu de données où les sujets n'ont parfois pas pu jouer leur rôle en adéquation avec le scénario demandé. Si nous ne prenons pas en compte les 30 premières et les 20 dernières secondes, les 130 secondes restantes sont des scénarios comparables à une situation de discussion réelle.

Les erreurs restantes sont dues à de mauvaises classifications du modèle. Nous nous attendons à ce que la distribution des données de la catégorie « refus argumenté » se situe au milieu des deux autres. Il est parfois difficile pour le modèle de classer le « refus argumenté » dans la bonne classe, comme illustré sur la figure 6.

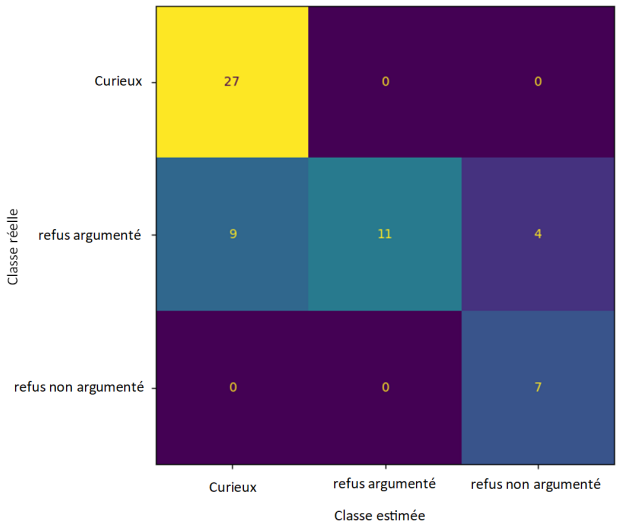

FIGURE 6 – Exemple de matrice de confusion.

6 Conclusion et perspectives

Nos travaux s'appuient sur un corpus d'interaction multimodale acquise dans un cockpit de voiture. Les performances obtenues avec notre modèle sont très prometteuses. La multimodalité et la méthode *statefull* RNN améliorent considérablement les performances. Nous obtenons une micro précision finale de 81% sur cinq ensembles de validation croisée différents.

Plusieurs pistes restent à investiguer. Un nouveau modèle de bout en bout (*i.e. end-to-end*) sera conçu pour ingérer les données vidéo et audio sans processus d'extraction manuelle. Ensuite, nous implémenterons la meilleure de nos deux approches dans un véhicule en tenant compte des ressources de calcul embarquées.

Remerciements

Ce travail a été partiellement financé via une bourse de doctorat industriel de l'Association Nationale de la Recherche et de la Technologie (ANRT), France.

Références

[1] 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)*, pages 19–23, 2019.

[2] David Arthur and Sergei Vassilvitskii. K-means++ : The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[3] Benjamin Bigot, Julien Pinquier, Isabelle Ferrané, and Régine André-Obrecht. Looking for relevant features for speaker role recognition. In *INTERSPEECH, Makuhari, Japan, 26/09/10-30/09/10*, pages 1057–1060. International Speech Communication Association (ISCA), 2010.

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

[5] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and R. B. V. Subramaanyam. Benchmarking multimodal sentiment analysis. *arXiv :1707.09538 [cs]*, 2017.

[6] Derek Chen, Howard Chen, Yi Yang, Alex Lin, and Zhou Yu. Action-Based Conversations Dataset : A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. *arXiv :2104.00783 [cs]*, April 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186,

Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. opensmile – the munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462, 01 2010.

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. *arXiv :1708.07632 [cs]*, 2017.

[10] Matthew Honnibal and Ines Montani. spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[11] 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)*, pages 145–150, 2018.

[12] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 :11–21, 1972.

[13] Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. BiERU : Bidirectional Emotional Recurrent Unit for Conversational Sentiment Analysis. *arXiv :2006.00492 [cs]*, February 2021.

[15] Edward Loper and Steven Bird. Nltk : The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia : Association for Computational Linguistics*, 2002.

[16] Yi Luan, Yangfeng Ji, and Mari Ostendorf. Lstm based conversation models, 2016.

[17] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn : An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 :6818–6825, July 2019.

[18] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Eric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.

[19] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe 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)*, pages 873–883. Association for Computational Linguistics, 2017.

[20] Quentin Portes., José Carvalho., Julien Pinquier., and Frédéric 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,*, pages 387–398. INSTICC, SciTePress, 2021.

[21] 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*, 41(1) :121–135, 2019.

[22] Bishal Santra, Potnuru Anusha, and Pawan Goyal. Hierarchical Transformer for Task Oriented Dialog Systems. *arXiv :2011.08067 [cs]*, March 2021.

[23] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[24] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard 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*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[25] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A Review of Recurrent Neural Networks : LSTM Cells and Network Architectures. *Neural Computation*, 31(7) :1235–1270, July 2019.

[26] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi : Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016.