

Analyse Automatique des Comportements Multimodaux lors d'Entretiens Vidéo Différés pour le Recrutement

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et technologies de
l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Unité de recherche : Université Paris-Saclay, CNRS, Laboratoire
interdisciplinaire des sciences du numérique, 91405, Orsay, France
Réfèrent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 04 février 2021, par

Léo HEMAMOU

Composition du jury :

Anne VILNAT Professeure, LIMSI-CNRS, France	Présidente
Louis-Philippe MORENCY Professeur, Université de Carnegie Mellon, Etats-Unis	Rapporteur & Examineur
Björn W. SCHULLER Professeur, Université de Augsburg, Allemagne	Rapporteur & Examineur
Ehsan HOQUE Maître de conférence, Université de Rochester, Etats-Unis	Examineur
Dinesh Babu JAYAGOPI Maître de conférence, IIIT Bangalore, Inde	Examineur
Deborah POWELL Maîtresse de conférence, Université de Guelph, Canada	Examinatrice
Jean-Claude MARTIN Professeur des Universités, LIMSI-CNRS	Directeur de thèse
Chloé CLAVEL Professeure, Telecom Paris	Co-directrice de thèse
Arthur GUILLON Responsable R&D,EASYRECRUE	Invité

Remerciements

Tout d'abord, j'aimerais remercier mes encadrants de thèse, Jean-Claude Martin et Chloé Clavel pour m'avoir aiguillé durant ces 3 dernières années. Votre présence a été une énorme énergie durant mon travail que ce soit de la part de votre encadrement (bienveillant et compréhensif), vos connaissances (transversales et pluridisciplinaires) et tout simplement votre savoir-être! Merci d'avoir toujours été dans la communication et l'échange, je n'aurais pu souhaiter un meilleur encadrement. Je retiendrai principalement les échanges de recherche riches, les fous rires et les moments conviviaux qui font partie de l'ADN des équipes du LIMSIS et de Telecom Paris. J'aimerais ensuite remercier mes co-auteurs Ghazi Felhi, Vincent Vandebussche et Arthur Guillon sans quoi rien n'aurait été possible. Merci pour votre regard critique et votre *dankness*! J'aimerais ensuite saluer et remercier les équipes de Telecom Paris et du LIMSIS où une ambiance toute particulière se dégage! En 3 ans, j'ai rencontré un nombre impressionnant de personnalités fascinantes et enrichissantes qui ont influencé ma vision de la recherche, de ma recherche et de ma personnalité. J'aimerais ainsi remercier (liste non exhaustive!) : Valentin Barrière pour sa bienveillance et son mentorship, Brian Ravenet pour ses talents d'écriture, Catherine Pelachaud pour sa transmission de l'amour de la recherche, l'équipe thé LIMSIS (dans l'ordre Lydia, JB, Alya, Delphine, Tom, Matthieu, David T., Yujiro, Jennifer, David R., Amine, Morghane, Sylvain, Florian, Jeremy, Antoine, Valentin), l'équipe des permanents CPU pour son recul et sa pensée pluridisciplinaire qui n'a pas de prix! (Céline, Elise, Nicolas, Vincent et Virginie), l'équipe Deadline Telecom (Atef, Alex, Mohammed, Pierre, Emile, Hamid, Asma, Tanvi, Lucien, Giovanna, Matthieu), la communauté WACAI (Matthieu, Magalie, Nesrine) et finalement l'équipe GRETA pour cette bienveillance ambiante (Guillaume, Caroline, Thomas, Irina, Soumia, Beatrice, Brice, Fajirian, Mireille, Sooraj, Nadine, Reshma).

J'aimerais remercier la société Easyrecrue pour avoir rendu cette recherche possible et particulièrement Mickael Cabrol, Jérémy Langlais et Grégory Wajntrob pour avoir cru en ce projet. J'aimerais ensuite remercier l'équipe RH qui fut mes rats de laboratoire pendant ces trois ans et particulièrement Amandine Reitz qui a répondu à mes demandes les plus relous (trier les postes de commerciaux, classer les compétences, etc). J'aimerais remercier ensuite les collègues d'Easyrecrue qui représentent le dynamisme de la startup nation avec qui j'ai lié une amitié particulière. Merci aux anciens (Roxanne, Quentin, PE, Skander, Hortense, Priscilla) et aux moins anciens (Baptiste, Mathilde, Flo Chauv, Flo Melchior, Quentin, Kamel, Alexis, Mickael, Celine, Marie, Donatien, Carole, etc).

Finalement, j'aimerais remercier mes amis et ma famille pour votre attention et votre amour. Merci à tous mes amis d'ici et d'ailleurs, de l'ultimate (Phoenix, PV, Revos, Amis belges!) de prépa (Matthieu Chap, je vous vois!) et d'école (Borel Team, 7NME, 3/4CD Thomas). Vous êtes l'essence qui permet de recharger les batteries quand on se fait malmener par le reviewer 3. Merci à mes parents Mohcine et France qui m'ont soutenu dans cette aventure sans oublier François, Fabienne, Philippe, Robin, Pauline, Adrien, Julia et merci à Gladys qui a su me supporter pendant mes nombreuses répétitions dans un espace restreint de 28m2 en plein confinement.

Enfin, je dédie cette thèse à mon grand père Paul qui aurait été fier de son petit-fils. Merci encore pour tout Pépé.

Table des matières

1	Introduction	1
1.1	Contexte académique	2
1.2	Contexte applicatif	4
1.3	Contexte législatif et éthique	5
1.4	Contribution	6
1.5	Organisation du manuscrit	8
I	Etat de l’art et matériel	11
2	L’entretien vidéo différé : un outil émergent	12
2.1	L’entretien d’embauche	13
2.2	Une nouvelle modalité : L’entretien vidéo différé	15
2.3	Fiabilité des entretiens d’embauche	16
2.4	Validité des entretiens d’embauche	17
2.5	La validité des attributs	17
2.6	La gestion de l’impression	19
2.7	Discrimination en entretien d’embauche	20
2.8	Acceptabilité des candidats	21
2.9	Comportements non verbaux	21
2.10	Conclusion	22
3	L’analyse automatique appliquée aux entretiens d’embauches	24
3.1	Outils automatiques pour l’entraînement aux entretiens d’embauche	24
3.2	Bases de données pour l’analyse automatique des entretiens d’embauche	25
3.2.1	Le contexte de collecte	25
3.2.2	Le poste ciblé	27
3.2.3	Les dimensions annotées	27

3.3	Les méthodes classiques d'apprentissage automatique	29
3.3.1	Descripteurs	29
3.3.2	Représentations temporelles	32
3.3.3	Représentations multimodales	33
3.3.4	Algorithmes de classification	34
3.3.5	Limites des approches classiques	34
3.4	Les méthodes d'apprentissage profond	34
3.4.1	Apprentissage de représentations	35
3.4.2	Représentation temporelle	36
3.4.3	Représentations multimodales	37
3.5	Interprétabilité et équité pour la confiance humain-machine	37
3.5.1	Méthodes d'interprétabilité	38
3.5.2	Équité dans l'apprentissage automatique	39
3.6	Conclusion	41
4	Matériel	43
4.1	Présentation de la plateforme	43
4.1.1	Le processus de construction de campagne de recrutement	43
4.1.2	Le processus de candidature	47
4.1.3	Le processus d'évaluation des candidats	48
4.2	La sélection des jeux de données	48
4.2.1	Le choix d'un unique type de métier	50
4.2.2	Méthodologie d'attribution des étiquettes	52
4.2.3	Sélection des réponses	55
4.3	Extraction de descripteurs sociaux multimodaux	56
4.3.1	Extraction de descripteurs issus de la vidéo	56
4.3.2	Extraction de descripteurs issus de la voix	58
4.3.3	Extraction de descripteurs issus du contenu verbal	59
4.4	Résumé des jeux de données utilisés dans cette thèse	59
II	Architecture neuronale pour l'analyse automatique d'entretiens d'embauche asynchrones	62
5	HireNet : Un modèle hiérarchique pour l'analyse automatique d'entretiens vidéo différés	63
5.1	HireNet et hypothèses sous-jacentes	63

5.2	Formalisation	64
5.2.1	Gated Recurrent Unit Encoder	66
5.2.2	Encodeur de la réponse	66
5.2.3	Encodeur de la question	67
5.2.4	Attention temporelle au niveau de la réponse	67
5.2.5	Encodeur de l'entretien	67
5.2.6	Encodeur de l'intitulé de poste	68
5.2.7	Attention temporelle au niveau de l'entretien	68
5.2.8	Classification du candidat	69
5.3	Expérimentations	69
5.3.1	Jeu de données	69
5.3.2	Protocoles expérimentaux	69
5.3.3	Extraction de descripteurs sociaux multimodaux	70
5.4	Comparaison de modèles	71
5.5	Modèles multimodaux	73
5.6	Résultats et analyses	73
5.7	Visualisation de l'attention	74
5.8	Conclusion	77
6	Attention et Multimodalité	78
6.1	HireNet monomodal, et mécanisme d'attention contextuelle	80
6.1.1	Limites de l'attention additive	81
6.1.2	Modification des fonctions d'attention	81
6.2	Expériences monomodales	84
6.2.1	Jeu de données et métriques d'évaluations	84
6.2.2	Modèles de références naïfs	84
6.2.3	Modèles de références état de l'art	85
6.2.4	Modèles de références HireNet	85
6.3	Résultats pour les expériences monomodales	86
6.4	Multimodal Hirenet, la multimodalité en pratique	87
6.4.1	Formalisation	88
6.4.2	Représentation des séquences d'entrée multimodales	88
6.4.3	Encodeur de modalité	90
6.4.4	Sous-échantillonnage à pas de temps réguliers	91
6.4.5	Gated Multimodal Unit	91
6.5	Expériences Multimodales	92

6.5.1	Jeu de données et métriques d'évaluations	92
6.5.2	Modèles de référence de l'état de l'art en entretien d'embauche . . .	93
6.5.3	Modèles de références HireNet	93
6.6	Résultats pour les expériences multimodales	98
6.7	Conclusion	98

III Vers une meilleure interprétabilité et équité des modèles de convocabilité **100**

7	Post-analyse des mécanismes d'attention : vers une interprétation des moments clés	101
7.1	Définition des courbes d'attention utilisées	102
7.2	Comment les valeurs d'attention sont-elles distribuées au niveau de l'entretien ?	104
7.3	A quoi correspondent les pics d'attention temporelle au sein d'une réponse ?	104
7.4	Comment les modalités sont-elles fusionnées au sein du GMU pendant les tranches d'attention ?	109
7.5	Le contenu des tranches d'attention est-il différent de celui des tranches aléatoires ?	111
7.6	Est-ce que les tranches d'attention contiennent plus d'information pour inférer la convocabilité que des tranches aléatoires ?	117
7.7	Conclusion	118
8	Équité individuelle pour la convocabilité	121
8.1	Influence du cadre législatif français	122
8.1.1	Collecte de données sensibles	122
8.1.2	Équité individuelle ou équité de groupe ?	123
8.1.3	Des méthodes proscrites	123
8.2	Matériel	124
8.2.1	Description du jeu de données	124
8.2.2	Descripteurs des vidéos monologues	125
8.3	Atténuer les biais dans la prédiction de la convocabilité par un apprentissage adversaire	125
8.3.1	Formalisation	125
8.3.2	Architecture	126
8.3.3	Stratégie d'entraînement pour le réseau adverse	130
8.4	Expériences	130

TABLE DES MATIÈRES

8.4.1	Métriques d'évaluation	130
8.4.2	Performance sur la tâche de convocabilité	131
8.4.3	Entraînement adversaire contre le genre et l'ethnicité.	134
8.4.4	Entraînement adversaire contre la représentation des visages	136
8.5	Conclusion	137
9	Conclusions et perspectives	140
9.1	Apport de notre travail	140
9.2	Perspectives de recherche	144
	Annexes	167
A	Liste des positions références choisies dans le référentiel ROME pour la sélection des postes dans la base de données EASYRECRUE	168
B	Expérimentation sur données simulées pour l'étude des fonction d'attention	172
B.1	Formalisation du cadre d'étude	172
B.2	Premier scénario : un contexte inutile	173
B.3	Deuxième scénario : un contexte indispensable	173
B.4	Architecture	175
B.5	Fonctions d'attention évaluées	176
B.6	Résultats de l'exemple jouet	177
C	Nuages des mots de la représentation multimodale responsables de la discrimination des tranches d'attention	179

Table des figures

1.1	Schéma global de la thèse.	9
2.1	Tableau issu de la méta-analyse de [Schmidt and Hunter, 1998]. Validité des principaux outils d'évaluation en sélection du personnel.	13
2.2	Modèle de la performance des candidats en entretien d'embauche centré sur ses attributs proposé par [Huffcutt et al., 2011].	19
4.1	Diapositive de présentation de l'EVD par la société EASYRECRUE	44
4.2	Spécifications des informations générales de la campagne.	45
4.3	Spécifications des critères d'évaluation de la campagne.	45
4.4	Spécifications des questions de la campagne.	46
4.5	Interface recruteur pour l'évaluation des candidats	49
4.6	Nombre de candidats par type de poste dans la base de données EASYRECRUE	51
4.7	Processus de sélection des campagnes pour le jeu de données	51
4.8	Visualisation du nuage de mots composant les titres d'emploi du premier jeu de données.	53
4.9	Stratégie d'étiquetage pour le jeu de données 2.	54
4.10	Pipeline d'extraction des descripteurs	56
5.1	HireNet. Les blocs en couleur correspondent aux encodeurs de contexte. . .	65
5.2	Valeur d'attention régularisée des 20 mots identifiés comme les plus importants	74
5.3	Exemple de moments saillants détectés grâce aux pics d'attentions pour la modalité vidéo	75
5.4	Questions de l'entretien d'embauche pour la position aléatoire choisie et leurs valeurs d'attentions respectives selon les différentes modalités.	76
6.1	Architecture neuronale HireNet modifiée.	82
6.2	Architecture du Multimodal HireNet	89
6.3	Encodeur multimodal de la réponse	89

6.4	Modèle de référence de la fusion intermédiaire au niveau de la question-réponse	94
6.5	Figures décrivant les modèles de références pour l'encodeur multimodal. . .	95
7.1	Exemple de courbes d'attention temporelle au niveau de l'entretien et la courbe d'attention moyennée résultante.	103
7.2	Exemple de courbes d'attention temporelle au niveau de la réponse et la courbe d'attention moyennée résultante.	103
7.3	Moyenne des scores d'attention temporelle normalisés au niveau de l'entretien, regroupés selon l'ordre des questions.	105
7.4	Un exemple de courbe d'attention temporelle au niveau de la réponse et quelques moments saillants	106
7.5	Densité du nombre de tranches d'attention en fonction de leur moment d'apparition relatif à la longueur totale de la réponse.	108
7.6	Boîtes à moustaches de la norme du vecteur $\sigma_t^m * h_t^m$ pour chaque modalité m (décrite en équation 6.18) pendant les tranches d'attention	110
7.7	Processus d'échantillonnage de moments aléatoires et d'étiquetage par rapport à un pic d'attention.	112
7.8	Nuage de mots des dimensions positivement associés avec les tranches d'attention.	116
7.9	Nuage de mots des dimensions négativement associés avec les tranches d'attention.	116
8.1	Architecture multimodale équitable proposée. Les versions monomodales sont obtenues en utilisant uniquement une modalité en entrée et en retirant le GMU.127	
B.1	Exemple d'individus issus du jeu de données du scénario 1 : un contexte inutile.174	
B.2	Exemple d'individus issus du jeu de données du scénario 2 : un contexte indispensable.	175
C.1	Nuage de mots des dimensions associées avec les tranches d'attention. . . .	180
C.2	Nuage de mots des dimensions associées avec les tranches d'attention. . . .	1

Liste des tableaux

3.1	Résumé des bases de données en lien avec l'analyse automatique d'entretien d'embauche	26
3.2	Tableau regroupant les méthodes d'apprentissage automatique pour chaque travail de l'état de l'art.	30
3.3	Tableau regroupant les méthodes d'apprentissage automatique pour chaque travail de l'état de l'art.	31
3.4	Tableau regroupant les méthodes d'apprentissage profond pour chaque travail de l'état de l'art.	35
4.1	Tableau récapitulatif des jeux de données utilisés au cours de cette thèse. .	61
5.1	Tableau descriptif du jeu de données : nombre de candidats dans chaque ensemble et statistiques globales de l'ensemble du jeu de données.	70
5.2	Résultats pour les modèles monomodaux	70
5.3	Résultats pour les modèles multimodaux naïfs et les modèles basés sur le vote	73
6.1	Tableau descriptif du jeu de données 2 : nombres de candidats et statistiques globales.	86
6.2	Résultats des expériences évaluant HireNet sur le second jeu de données. . .	87
6.3	Comparaison des performances des modèles de références monomodales et multimodales par rapport Multimodal HireNet.	97
7.1	Similarité de Jaccard entre les tranches d'attention de différentes modalités.	107
7.2	Statistiques descriptives des tranches d'attention extraites.	109
7.3	Résultats de classification pour la tâche d'identification des tranches d'attention et aléatoires.	114
7.4	Analyse de l'importance des descripteurs pour les modèles monomodaux. .	119

7.5	Résultats pour la tâche de l'employabilité en utilisant un jeu de données comprenant des tranches aléatoires ou un jeu de données utilisant des tranches d'attention.	120
7.6	Analyse d'importance des descripteurs multimodaux.	120
8.1	Résumé des biais initiaux par rapport au nouveau découpage des données proposée	129
8.2	Différences entre les deux découpages de données pour la prédiction de la convocabilité en utilisant la représentation faciale.	132
8.3	Différences entre les deux découpages de données pour la prédiction de la convocabilité en utilisant les réseaux proposés.	133
8.4	Résultats pour la tâche de la convocabilité et l'effet disparate du système automatique.	134
8.5	Résultat pour l'entraînement adverse contre le genre et l'ethnicité	138
8.6	Effets de l'entraînement adversaire contre la représentation des visages sur les métriques d'évaluation.	139
B.1	Tableau des résultats de justesse en fonction différentes fonctions d'attention sur les jeux de données de l'exemple jouet. D_{S_1} et D_{S_2} consistent en la concaténation des deux jeux de données.	178

Liste des abréviations

càd	C'est à dire
RH	Ressources humaines
EVD	Entretien vidéo différé
AA	Analyse automatique
GRU	Gated Reccurent Unit
GMU	Gated Multimodal Unit
FF	Face à face
BOW	Bag Of Words - Sac de mots
BERT	Bidirectional Encoder Representations from Transformers
LIWC	Linguistic Inquiry Word Count
LR	Régression logistique
SVM	Séparateur à vastes marges
ASR	Automatic speech recognition - reconnaissance automatique de la parole
RGPD	Règlement général sur la protection des données

Section 1

Introduction

Le développement des nouvelles technologies impacte tous les secteurs d'activités, y compris celui des Ressources Humaines, que ce soit durant la recherche de candidats ou dans le processus de sélection.

Ainsi, l'entretien vidéo différé permet d'organiser en asynchrone des entretiens avec des candidats et de les évaluer. Les candidats se connectent à une plateforme, se filment pendant qu'ils répondent à des questions définies à l'avance par les recruteurs. La plateforme permet ensuite à plusieurs recruteurs d'évaluer le candidat, d'échanger entre eux et d'inviter éventuellement le candidat à un entretien en face à face. Les recruteurs établissent au préalable un questionnaire de recrutement et y associent des critères d'évaluation. Le/la candidat-e reçoit une invitation pour répondre à ces questions et enregistre en vidéo ses réponses selon ses disponibilités dans un temps limité. Il ou elle n'a pas connaissance des questions à l'avance et ne peut pas se réenregistrer afin de préserver la spontanéité de ses réponses. Le recruteur reçoit les vidéos sur une interface et peut ainsi comparer et évaluer les différents profils avec ses équipes selon les critères définis précédemment. De plus en plus d'entreprises font le choix de ce type d'entretien comme outil de présélection. Le choix d'un tel outil est motivé par l'accès à un plus grand nombre et une plus grande diversité de candidats et par la réduction du temps de traitement et de prise de rendez-vous [Torres and Mejia, 2017].

Le nombre de telles candidatures vidéo devient en conséquence de plus en plus volumineux et difficile à traiter « manuellement » par un ou deux recruteurs. Il devient donc nécessaire d'envisager une aide pour le recruteur devant traiter parfois plusieurs dizaines (voire centaines) d'entretiens vidéo. De plus, le développement d'une telle aide pourra aussi permettre aux candidats de s'entraîner à l'exercice de l'entretien vidéo différé grâce à une évaluation automatique. Très peu de recherches, à notre connaissance, ont porté sur l'étude des entretiens vidéo différés dans un contexte hors laboratoire. Il peut y avoir pourtant

de grandes différences entre une situation réelle d'embauche dans laquelle des candidats sont réellement motivés par une véritable candidature, et des conditions expérimentales contrôlées dans lesquelles les participants simulent un intérêt pour un poste fictif.

Ce contexte soulève plusieurs enjeux applicatifs en termes de recherche : comment aider le recruteur à identifier les candidats pertinents pour le profil de poste et le domaine recherché ? Quels sont les comportements sur lesquels les recruteurs se fondent implicitement pour réaliser cette présélection ?

L'objectif de cette thèse est de développer des recherches en traitement automatique des signaux sociaux multimodaux (expressions faciales, prosodie et contenu verbal) afin d'apporter des réponses à ces questions et fournir ainsi une assistance aux recruteurs et aux candidats.

Cette thèse s'est déroulée dans le cadre d'une convention CIFRE établie entre deux laboratoires de recherches académiques et l'entreprise EASYRECRUE. À la frontière entre plusieurs disciplines (Analyse automatique des signaux sociaux, apprentissage machine, Psychologie du travail), ce partenariat nous a permis de mener à bien nos travaux. D'une part, nous avons profité de l'expertise de la compréhension des comportements et des précédents travaux en psychologie au sein du groupe Cognition, Perception et Usages (CPU) au Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur (LIMSI). D'autre part, nous avons bénéficié de l'expertise en informatique affective et en détection automatique de comportements sociaux émotionnels au sein du Laboratoire du Traitement de l'Information et de la Communication (LTCI). Enfin, ce travail de thèse est ancré dans un cadre industriel particulièrement stimulant pour l'étude des entretiens d'embauche différés au sein de la startup EASYRECRUE. L'entreprise emploie plus de 70 employés, compte 450 clients dont Crédit Agricole, Monoprix, Heineken, UPS ou Sanofi et plus d'un million de candidats tout autour de la planète ont déjà réalisé un entretien vidéo différé au travers de la plateforme EASYRECRUE.

Dès lors, nous pouvons introduire les différents contextes liés à la thèse qu'ils soient de nature académique, applicative ou législative.

1.1 Contexte académique

Cette thèse est par nature pluridisciplinaire. Premièrement, l'objet de nos travaux se concentre sur l'analyse de performances en entretiens vidéo différés. Les entretiens d'embauche sont depuis longtemps un objet très largement étudié en psychologie du travail [Schmitt, 2012]. Ainsi, de nombreux travaux se sont intéressés à l'étude de leur validité

pour évaluer les futures performances au travail en fonction de la façon dont l'entretien d'embauche était construit ou évalué [Levashina et al., 2014] et ce depuis plus d'une cinquantaine d'années [Schmidt and Hunter, 1998]. Ces travaux convergent sur le fait que l'entretien d'embauche demeure un des outils les plus valides pour sélectionner les employés les plus adéquats pour le poste vacant. Dès lors, il n'est pas surprenant que l'entretien d'embauche reste l'outil de sélection le plus utilisé par les recruteurs [Salgado, 2017]. De récents travaux s'intéressent maintenant à la compréhension des éléments influents responsables de l'évaluation de la performance en entretien [Huffcutt et al., 2011]. Ces facteurs latents peuvent être multiples et divers, nous pouvons citer notamment l'état mental du candidat (motivation, anxiété), ses stratégies d'influence (par exemple autopromotion) ou l'influence de ses comportements non verbaux [Schneider et al., 2015]. Contrairement à son homologue face à face, l'entretien vidéo asynchrone ou différé reste une modalité d'entretien émergente apparue dans les années 2010 et peu de travaux ont porté sur celle-ci. Ainsi, un tout nouveau cadre est abordé dans cette thèse.

Deuxièmement, l'informatique affective et sociale a largement bénéficié ces dernières années de l'avancée de l'apprentissage automatique que ce soit pour l'annotation automatique des comportements (prosodie, expressions faciales) ou la retranscription automatique. Ces avancées ont permis d'obtenir des méthodes rapides pour aider à l'annotation et réaliser des études quantitatives en psychologie du travail et plus particulièrement dans le cadre des entretiens d'embauche [Frauendorfer and Mast, 2015, Nguyen, 2015]. Ainsi, de nombreux travaux en informatique affective ont vu le jour pour 1) aider les candidats à s'améliorer en entretien d'embauche ou à la prise de parole en public par exemple grâce à l'aide de recruteurs virtuels [Hoque et al., 2013, Zhao et al., 2017, Batrinca et al., 2013, Anderson et al., 2013, Tanveer et al., 2016], 2) inférer l'employabilité des candidats pour une aide aux recruteurs ou pour comprendre les attributs influents en entretien d'embauche [Rasipuram and Jayagopi, 2018, Chen et al., 2017].

En parallèle de ces travaux, l'étude du comportement humain a largement bénéficié des architectures neuronales pour la prédiction des émotions [Zadeh et al., 2018b], des opinions [Garcia et al., 2019a] ou de la personnalité [Escalante et al., 2020] construites grâce à de larges corpus d'enregistrements vidéos. Ces architectures neuronales ont un avantage singulier dans leur capacité à modéliser la séquentialité ou la multimodalité, capacité nécessaire afin de pouvoir pleinement mettre en exergue des comportements non verbaux influents de l'entretien d'embauche.

L'entretien vidéo différé fournit un cadre favorable pour la recherche en informatique affective et sociale. D'une part, sa nature permet de constituer une base de données beaucoup plus grande que celle que l'on pourrait collecter en entretien face à face grâce à son caractère

asynchrone. D'autre part, aucun réseau de neurones courant n'est à même de modéliser la structure propre de l'entretien d'embauche.

Cependant, les performances de tels systèmes viennent au prix d'une extrême opacité, caractéristique indésirable dans notre cas d'étude. Afin de répondre à ce besoin, de nombreuses recherches en interprétabilité pour les modèles d'apprentissages profonds ont vu le jour et le champ de recherche est en plein essor [Miller, 2019, Gilpin et al., 2019]. Néanmoins, aucune étude n'a porté spécifiquement sur l'analyse des comportements verbaux et non verbaux en entretien d'embauche. L'utilisation de tels systèmes pourrait fournir un outil exploratoire et d'analyse utile aux recherches en psychologie du travail.

1.2 Contexte applicatif

Le milieu des ressources humaines a connu des bouleversements sans précédent ces dernières années et en particulier le recrutement. Ainsi, de nombreux sites d'emploi ou plus communément appelés *jobboard* ont fait leur apparition (Monster.com ou Indeed.com par exemple) matérialisant le marché de l'emploi en connectant virtuellement les recruteurs aux demandeurs d'emploi. À cela s'ajoute l'émergence des réseaux sociaux destinés à la sphère professionnelle comme LinkedIn ou Viadeo.

Cette digitalisation simplifie la collecte de candidatures effectuée par les recruteurs, à un tel point que parfois le nombre de candidatures dépasse les capacités de traitement des équipes de ressources humaines. De plus, la "course au talent" est telle que la réactivité devient une priorité de peur de laisser passer le meilleur candidat. Dès lors, de nombreuses sociétés proposent des outils pour simplifier et aider les recruteurs dans ce processus de sélection via des ATS (Outil de suivi de candidature). Un récent rapport montre que plus de 98 % des compagnies de Fortune 500 utilisent des ATS dans leur processus de sélection. Cette révolution s'est aussi caractérisée par l'usage de l'intelligence artificielle pour effectuer des recommandations spécifiques pour les recruteurs ou les candidats, filtrer automatiquement les CV des candidats, ou générer automatiquement des textes d'offres d'emploi plus attractifs [Sánchez-monederó and Dencik, 2019].

De la même manière, de nombreuses plateformes commerciales d'entretien vidéo différé proposent maintenant l'utilisation de l'intelligence artificielle pour effectuer un classement ou une évaluation de candidats [Raghavan et al., 2019].

Néanmoins, aucune publication scientifique ne résulte de ces solutions commerciales et la validité de ces produits demeure obscure. De plus, la majorité de ces plateformes sont américaines et traitent des entretiens en langue anglaise laissant un questionnement sur la

faisabilité de tels outils pour la langue et la culture française.

Comme évoqué précédemment, cette thèse est portée par le financement de la société EASYRECRUE et s'inscrit dans un contexte industriel. À l'avantage par rapport à précédentes études dans le contexte académique, nous avons accès à une base de données conséquente pour la recherche académique. Cette base de données est constituée de réels entretiens vidéos différés et annotés avec de réelles décisions d'employabilité prises par de vrais recruteurs. De plus, l'accès à une base de données importante telle que celle d'EASYRECRUE nous permet 1) d'accéder à une cohorte pour obtenir des résultats suffisamment significatifs, 2) l'exploration de méthodes d'apprentissages profonds pour l'analyse automatique des entretiens vidéos.

Dans ce contexte, notre principale étude consiste à participer à la conception et à la validation d'un outil visant à prédire la convocabilité du candidat (si le recruteur décide de l'inviter en entretien face à face ou non). Il s'agit donc d'évaluer la faisabilité d'un tel outil. À cela s'ajoute le besoin d'interprétabilité et d'explicabilité afin de construire une relation de confiance entre l'utilisateur du système (candidat ou recruteur) et l'outil automatique [Basch and Melchers, 2019].

1.3 Contexte législatif et éthique

Le recrutement reste un domaine sensible que ce soit d'un point de vue législatif ou éthique. De plus, l'utilisation de l'intelligence artificielle au sein du processus de recrutement a rencontré le scepticisme, tant du législateur [Raji et al., 2020] que du public, qui craignent les comportements injustes et les dérives de sélection de ces algorithmes. Les vendeurs de technologies d'évaluation algorithmique pour la présélection font généralement état de prévisions impartiales, mais ces affirmations sont rarement étayées par des études ou des audits publiés [Raji et al., 2020]. Malgré l'importance cruciale des entretiens tant pour les candidats que pour les services des ressources humaines, les affirmations des fournisseurs doivent être acceptées telles quelles, alors qu'il n'existe aucune procédure standard d'évaluation ou d'impartialité [Sánchez-Monedero et al., 2020]. En outre, ces sociétés ne garantissent l'équité qu'en ce qui concerne l'*égalité de sélection* (c'est-à-dire si le taux de sélection diffère d'un sous-groupe à l'autre, par exemple hommes vs femmes), mais ne fournissent aucune garantie par rapport à l'*égalité de traitement* (c'est-à-dire si les sous-groupes sont traités de la même façon) [Lipton et al., 2018]. Cette problématique devient encore plus compliquée lorsque la collecte de variables sensibles est strictement interdite, et ce même dans le cadre d'un audit interne rendant l'évaluation de l'égalité de sélection

impossible [Lieberman, 2001]. L'éthique et l'influence des biais sont dès lors un sujet primordial marqué pour les discussions multiples de comités d'éthique (par exemple le Comité Consultatif National d'Éthique¹) et de groupes de réflexions (Institut Montaigne²).

Notre travail de thèse fait un pas vers l'acquisition de modèles d'analyse automatique plus justes en proposant une méthodologie afin de limiter les inégalités de traitement qui pourraient apparaître lors de la modélisation d'un tel système.

Enfin, les données que nous manipulons sont encadrées par le Règlement Général sur la Protection des Données (RGPD) au niveau européen. Dans ce sens, plusieurs contraintes sont apparues au cours de cette thèse. Premièrement, les jeux de données obtenus ont été complètement anonymisés afin de respecter la réglementation. Deuxièmement chaque vidéo a une date limite de conservation au-delà de laquelle la vidéo est supprimée. Ainsi, il a fallu constituer un second jeu de données au milieu de notre travail de thèse pour la poursuite de nos travaux.

1.4 Contribution

Les contributions de cette thèse s'inscrivent dans le cadre de l'apprentissage automatique supervisé pour l'analyse des entretiens vidéos différés. Elles répondent aux questions de recherches suivantes :

1. Peut-on proposer une architecture neuronale adaptée à modéliser les entretiens structurés ?
2. Quels sont les comportements influents en entretien d'embauche ?
3. Comment fusionner plusieurs modalités potentiellement bruitées ?
4. Comment limiter les potentiels biais de modèles neuronaux ?

Les contributions de ce manuscrit se résument ainsi :

Architecture neuronale hiérarchique avec mécanisme d'attention adapté aux entretiens structurés : Nous apportons une première contribution à la première question de recherche en proposant une architecture neuronale adaptée aux entretiens structurés nommée HireNet (Chapitre 5). Cette architecture s'appuie sur le caractère hiérarchique de l'entretien d'embauche en le décomposant en une séquence de questions-réponses, elle-même composée d'une séquence de mots ou de comportements non verbaux. De plus, nous mettons en évidence l'importance de la prise en compte du contexte (intitulé des questions

1. <https://www.ccne-ethique.fr/>

2. <https://www.institutmontaigne.org/publications/algorithmes-controle-des-biais-svp>

et du poste). Enfin, l'utilisation de mécanismes d'attention nous permet d'obtenir une interprétabilité locale (individu par individu) permettant partiellement de répondre à la troisième question de recherche.

Ces travaux ont donné lieu à la publication [Hemamou et al., 2019b] : *Hemamou Léo, Felhi Ghazi, Vandebussche Vincent, Martin Jean-claude, Clavel Chloé. HireNet : A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. Proceedings of the AAAI Conference on Artificial Intelligence. 33, (juill. 2019).*

Proposition d'un mécanisme de fusion multimodale interprétable pour modalités bruitées : Nous apportons une réponse à la troisième question de recherche en proposant une méthode de fusion à pas de temps fixe contrôlant la contribution de chacune des modalités (Chapitre 6).

Notre cas d'application s'intéresse à une fusion multimodale avec une retranscription automatique du contenu verbal au contraire de la majorité des précédents travaux opérant sur une retranscription manuelle. Cette particularité remet en question l'absence de bruit dans le contenu verbal, et la fusion au niveau du mot largement adopté dans la communauté. Nous proposons un module qui permet une fusion des modalités à un niveau fin grain, et ce à un pas de temps fixe (toutes les 0.1s). De plus, ce module réside sur des mécanismes d'attentions contrôlant l'importance de chacune des modalités. Enfin, un tel module permet aussi d'obtenir une interprétabilité locale en fournissant une visualisation des modalités les plus influentes en entretien d'embauche répondant partiellement à la deuxième question de recherche.

Ces travaux ont donné lieu à la soumission d'un article journal : *Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Multimodal Hierarchical Attention Neural Network : Looking for Candidates Behaviour which Impact Recruiter's Decision*

Analyse approfondie des mécanismes d'attention : Nous fournissons une méthodologie pour comprendre et analyser les événements locaux importants détectés par les mécanismes d'attentions (Chapitre 7). En ce sens, nous caractérisons par une étude quantitative ce en quoi les instants importants d'un entretien vidéo diffèrent d'instantanés aléatoires. Cette méthode apporte une réponse à la deuxième question de recherche en généralisant les connaissances obtenues par les mécanismes d'interprétabilité locaux des deux premières contributions.

Ces travaux ont donné lieu à la publication de deux articles et à la soumission d'un article journal :

- [Hemamou et al., 2019a] : *Hemamou Léo, Felhi Ghazi, Martin Jean-claude, Clavel Chloé. Slices of Attention in Asynchronous Video Job Interviews. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII) (sept. 2019)*
- [Hemamou et al., 2020] : *Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Attention Slices dans les Entretiens d ' Embauche Vidéo Différés. Workshop sur les "Affects, Compagnons Artificiels et Interactions" (ACAI) (2020)*
- *Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Multimodal Hierarchical Attention Neural Network : Looking for Candidates Behaviour which Impact Recruiter's Decision*

Méthode adversaire pour la suppression d'informations sensibles dans les représentations neuronales : Nous apportons une réponse à la quatrième question de recherche en fournissant et en évaluant une méthodologie pour retirer des informations sensibles dans les représentations neuronales par méthode adversaire (Chapitre 8). À notre connaissance, aucun travail n'a spécifiquement contribué à améliorer l'équité dans les réseaux neuronaux multimodaux appliqués au recrutement. Nous fournissons une première évaluation et étude de faisabilité dans ce domaine. Enfin, nous proposons une nouvelle méthode afin d'assurer cette équité individuelle sans le besoin de collecter de variables sensibles.

Ces travaux ont donné lieu à la soumission d'un article : *Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Using Adversarial Learning for Removing Sensitive Information in Neural Representation for Hireability Prediction in Monologue Videos*

1.5 Organisation du manuscrit

Le manuscrit s'organise en trois grandes parties. La première partie est consacrée à l'étude préalable, mais non moins nécessaire, de la compréhension des entretiens d'embauche et plus précisément de l'entretien vidéo différé, de l'état de l'art concernant l'analyse automatique appliquée aux entretiens d'embauche et de la présentation de la plateforme EASYRECRUE. La deuxième partie est dédiée à la proposition d'un modèle neuronal multimodal performant pour la prédiction de la convocabilité des candidats en entretien vidéo différé. La troisième partie regroupe les travaux relatifs à l'interprétabilité à l'équité du

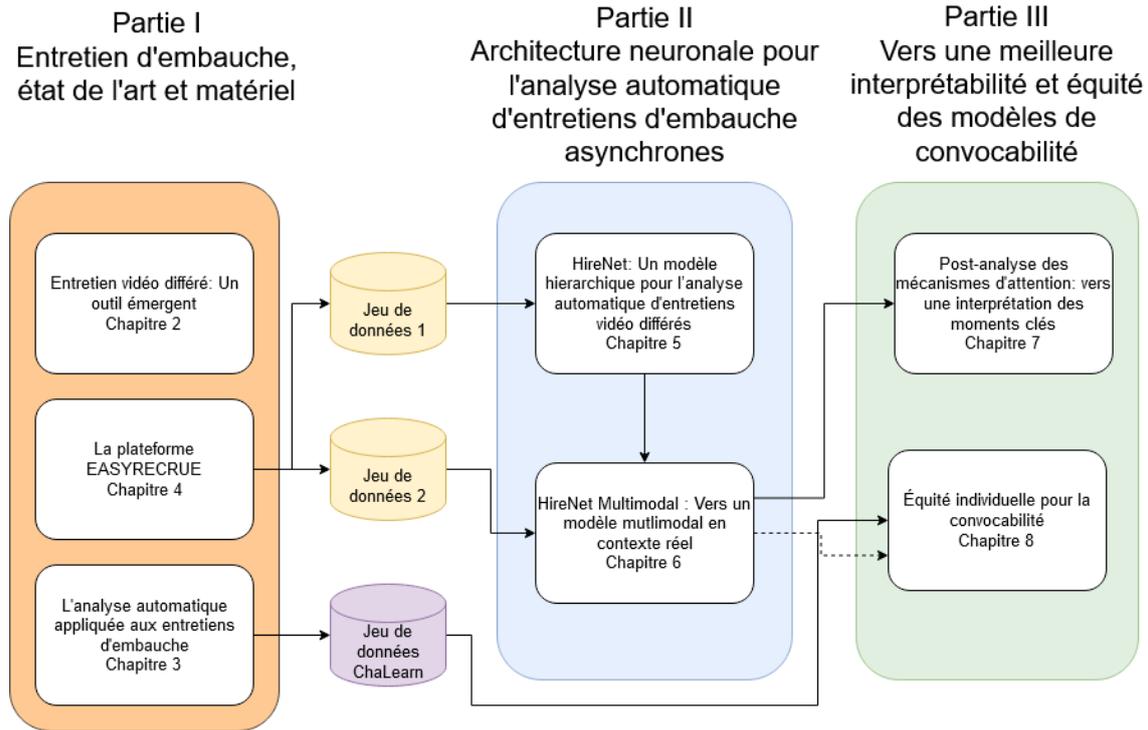


FIGURE 1.1 – Schéma global de la thèse.

modèle.

La figure 1.1 présente un Schéma global de la thèse.

- La partie I présente les entretiens d'embauche, l'état de l'art de l'analyse automatique appliqué aux entretiens d'embauches et le matériel utilisé au cours de cette thèse
 - Le chapitre 2 présente les travaux précédents autour des entretiens d'embauche et comment l'entretien vidéo différé, média émergent, s'inscrit dans cette littérature
 - Le chapitre 3 présente les bases de données et les méthodes utilisées pour l'inférence automatique dans le contexte des entretiens d'embauche qu'ils soient face-à-face ou différés
 - Le chapitre 4 présente la plateforme du partenaire EASYRECRUE grâce à laquelle deux jeux de données ont été constitués. La méthodologie de sélection est décrite et la chaîne de traitement des descripteurs détaillée.
- La partie II *Vers une modélisation plus performante* décrit les expériences et le modèle neuronal multimodal proposé pour l'analyse automatique d'entretien vidéo différé
 - Le chapitre 5 est consacré à la modélisation d'une architecture neuronale pour l'inférence de la convocabilité adaptée au contexte des entretiens vidéo différés.

- Le chapitre 6 est consacré à la proposition d'une nouvelle forme d'attention pour une meilleure prise en compte du contexte (questions et intitulé de poste). De plus, nous proposons un modèle adapté à la prise en compte de la multimodalité lors d'utilisation d'un contenu verbal obtenu par retranscription automatique.
- La partie III décrit les expériences et les méthodes proposées pour l'obtention d'une meilleure interprétabilité et équité du système
- Le chapitre 7 est consacré à une analyse approfondie des mécanismes d'attention des modèles proposés en partie II.
 - Le chapitre 8 est consacré à la proposition d'une méthode adversaire pour la suppression d'information sensible dans les représentations neuronales afin d'assurer un traitement égal pour tous les candidats.

Première partie

Etat de l'art et matériel

Section 2

L'entretien vidéo différé : un outil émergent

Du point de vue du recruteur, il existe plusieurs méthodes pour évaluer l'adéquation d'un candidat à un emploi ou à l'entreprise. Après avoir effectué une analyse de poste et déterminé quelles compétences, connaissances et caractéristiques individuelles sont nécessaires pour le poste, le recruteur choisit les outils d'évaluations les plus adaptés. Ces outils peuvent être des tests classiques (personnalité, connaissances liées au poste), des vérifications de références, des évaluations par des pairs ou des entretiens [Schmidt and Hunter, 1998](voir figure 2.1). Parmi ces outils, l'entretien d'embauche reste le moyen le plus utilisé afin d'évaluer des candidats. Un entretien permet à un recruteur de vérifier des informations, d'évaluer les compétences du candidat, de déterminer une personnalité ou de vérifier l'adéquation du candidat avec la culture de l'entreprise ou le poste.

Dans cette thèse, nous nous intéressons particulièrement à l'entretien d'embauche, outil d'évaluation le plus populaire dans les processus de sélection [Salgado, 2017] et d'une de ces déclinaisons émergentes : l'entretien vidéo différé.

La suite de ce chapitre s'articule de la façon suivante : nous présentons tout d'abord l'entretien d'embauche et l'entretien vidéo différé. Puis nous détaillons par la suite plusieurs thématiques courantes de recherche en psychologie de travail et nous situons comment l'entretien vidéo différé s'inscrit dans ces différentes thématiques. Enfin, nous concluons en positionnant l'analyse automatique des entretiens vidéo différés au travers de ces différentes thématiques.

Table 1
*Predictive Validity for Overall Job Performance of General Mental Ability (GMA) Scores
 Combined With a Second Predictor Using (Standardized) Multiple Regression*

Personnel measures	Validity (<i>r</i>)	Multiple <i>R</i>	Gain in validity from adding supplement	% increase in validity	Standardized regression weights	
					GMA	Supplement
GMA tests ^a	.51					
Work sample tests ^b	.54	.63	.12	24%	.36	.41
Integrity tests ^c	.41	.65	.14	27%	.51	.41
Conscientiousness tests ^d	.31	.60	.09	18%	.51	.31
Employment interviews (structured) ^e	.51	.63	.12	24%	.39	.39
Employment interviews (unstructured) ^f	.38	.55	.04	8%	.43	.22
Job knowledge tests ^g	.48	.58	.07	14%	.36	.31
Job tryout procedure ^h	.44	.58	.07	14%	.40	.20
Peer ratings ⁱ	.49	.58	.07	14%	.35	.31
T & E behavioral consistency method ^j	.45	.58	.07	14%	.39	.31
Reference checks ^k	.26	.57	.06	12%	.51	.26
Job experience (years) ^l	.18	.54	.03	6%	.51	.18
Biographical data measures ^m	.35	.52	.01	2%	.45	.13
Assessment centers ⁿ	.37	.53	.02	4%	.43	.15
T & E point method ^o	.11	.52	.01	2%	.39	.29
Years of education ^p	.10	.52	.01	2%	.51	.10
Interests ^q	.10	.52	.01	2%	.51	.10
Graphology ^r	.02	.51	.00	0%	.51	.02
Age ^s	-.01	.51	.00	0%	.51	-.01

FIGURE 2.1 – Tableau issu de la méta-analyse de [Schmidt and Hunter, 1998]. Validité des principaux outils d'évaluation en sélection du personnel. La validité des outils est évaluée en évaluant la corrélation entre la mesure de l'outil d'évaluation et les performances au travail principalement mesurées qualitativement par les superviseurs. L'entretien d'embauche demeure un outil privilégié pour sa validité notamment lorsqu'il est structuré.

2.1 L'entretien d'embauche

L'entretien d'embauche est défini dans [Levashina et al., 2014] comme *un processus interactif personnel au cours duquel une ou plusieurs personnes posent des questions oralement à une autre personne et évaluent les réponses afin de déterminer les qualifications de cette personne en vue de prendre des décisions en matière d'emploi*. Nous étendons cette définition aux entretiens vidéos différés en modifiant la définition comme suit : *un processus interactif personnel au cours duquel une ou plusieurs personnes posent des questions oralement ou à l'écrit à une autre personne et évaluent ses réponses orales ou écrites afin de déterminer les qualifications de cette personne en vue de prendre des décisions en matière d'emploi*.

Il est important de distinguer deux grandes catégories d'entretien d'embauche nommément l'entretien d'embauche dit structuré et l'entretien d'embauche dit non structuré. En effet, un des points importants soulignés ces dernières années par la psychologie du travail est le fait que les entretiens d'embauches structurés sont plus fiables et valides que leur contrepartie non structurée [Gavand, 2013, Schmidt and Hunter, 1998]. Ainsi plus d'une vingtaine de méta-analyses ont été conduites et marquent une supériorité constante dans

l'usage de l'entretien structuré par rapport à l'entretien non structuré [Levashina et al., 2014]. Au-delà d'une meilleure validité et fiabilité, il a été montré que la structure de l'entretien réduit l'influence des différences interindividuelles (race, genre, et handicaps) sur le jugement du recruteur. L'entretien d'embauche structuré se distingue de l'entretien non structuré par rapport à la présence ou l'absence de 15 composantes [Campion et al., 1997]. Ces composantes se séparent en deux grandes dimensions que sont le *contenu de l'entretien* et *l'évaluation de l'entretien*. Le contenu de l'entretien est dit structuré s'il intègre les composantes suivantes : a) les questions de l'entretien sont construites à la suite d'une analyse du travail, b) les questions posées aux candidats sont toujours les mêmes, c) éviter de poser des sous-questions exploratoires qui ne visent pas à obtenir une clarification de la réponse, d) privilégier de meilleures questions de type situationnel, comportementales, techniques ou liées aux expériences passées, e) poser un plus grand nombre de questions, f) ne pas prendre connaissance d'autres informations avant l'entretien (p. ex. CV, évaluation d'autres examinateurs, prises de référentiels, etc.), g) ne pas autoriser les questions de candidats jusqu'à la fin de l'entretien

L'évaluation de l'entretien est dite structurée si elle intègre les composantes suivantes : a) évaluer chaque question ou utiliser des échelles multiples d'évaluation, b) utiliser des échelles comportementales, c) prendre des notes détaillées pendant l'entretien, d) utiliser plusieurs examinateurs, e) utiliser les mêmes examinateurs pour l'ensemble des candidats, f) ne pas échanger à propos des candidats entre les entretiens, g) entraîner les recruteurs à ce type d'entretien, h) évaluer d'une façon statistique et déterminée préalablement choisie avant l'entretien d'embauche.

On considère un entretien comme structuré s'il contient a minima 6 composantes, les principales étant la présence de l'analyse du travail, les mêmes questions pour tous les candidats, de meilleures questions, l'évaluation de chaque question, l'utilisation d'échelles comportementales et l'entraînement des examinateurs aux entretiens structurés.

De façon intéressante, l'enregistrement des entretiens rajoute possiblement de la structure aux entretiens, limitant le biais de mémoire initialement résolu par la prise de notes pendant l'entretien d'embauche. La littérature en psychologie du travail s'est concentrée sur quatre principales configurations d'entretien d'embauche que sont les entretiens d'embauche face à face, par téléphone, par vidéoconférence et par vidéo différé [Levashina et al., 2014, Torres and Mejia, 2017]. Concernant, l'analyse automatique des entretiens, les deux principales modalités d'entretien étudiées sont respectivement les entretiens face à face et les entretiens vidéos différés. Les conditions dans lesquels le candidat effectue son entretien d'embauche (face à face, par vidéoconférence, etc.) sont importantes, car elles peuvent influencer sur la performance des candidats en entretien d'embauche [Rasipuram and Jaya-

gopi, 2018] et avoir une influence sur le degré de structure de l'entretien [Levashina et al., 2014, Posthuma et al., 2002]. L'entretien face à face consiste en un entretien synchrone où le candidat se trouve en face d'un ou plusieurs examinateurs. Cette configuration d'entretien est la plus courante et la plus étudiée [Levashina et al., 2014]. Au contraire l'entretien vidéo différé est asynchrone : le candidat s'enregistre en vidéo en répondant à une série de questions et le recruteur a la possibilité de regarder et d'évaluer ce candidat lorsqu'il est disponible. Bien que l'entretien vidéo différé devienne de plus en plus utilisé notamment au sein d'entreprises internationales telles que ING, PWC ou Disneyland, il reste un objet peu étudié en psychologie du travail, nous essayons donc de fournir une vue complète de cet objet en le comparant aux entretiens face à face.

2.2 Une nouvelle modalité : L'entretien vidéo différé

L'entretien vidéo différé ou aussi appelé entretien asynchrone ou parfois entretien vidéo se distingue de l'entretien synchrone ou de la vidéoconférence par son caractère à trois étapes. La première étape consiste en la création d'une campagne de recrutement regroupant des questions textuelles ou vidéos de la part du recruteur pour le candidat. Une fois cette campagne de recrutement créée, une invitation est ensuite envoyée au candidat.

Ce candidat a par la suite la possibilité d'effectuer son entretien quand il le souhaite et où il le souhaite dans un délai généralement d'une semaine. Cet entretien consiste principalement à enregistrer ses vidéos-réponses à la suite du visionnage des questions de recrutement pré-enregistrées. Un temps de réflexion généralement de 30s est accordé au candidat avant l'enregistrement de ses réponses afin de préparer sa réponse.

Par la suite, le recruteur regarde, évalue et partage les vidéos réponses des postulants aux autres membres de l'équipe de recrutement. Si un candidat est jugé positivement, il continue le processus d'embauche. L'entretien vidéo différé a principalement pour but d'assurer une étape de présélection et peut être considéré comme similaire dans sa fonction à un entretien téléphonique pour le recruteur.

L'avantage pour les candidats demeure dans le fait qu'ils sont libres dans l'instant et l'endroit pour compléter leur candidature, simplifiant ainsi leur démarche dans le cas où ils ont peu de disponibilités, ou des fuseaux horaires différents. De plus, ils ne sont pas dans l'obligation de se déplacer physiquement comme il serait le cas lors d'un entretien face-à-face. À noter que cette flexibilité s'applique aussi aux recruteurs, chaque recruteur étant libre d'évaluer les entretiens quand il le souhaite. De plus, l'EVD permet de diminuer les coûts en réduisant le temps normalement alloué pour les étapes de présélection [Torres

and Gregory, 2018].

Le premier constat à effectuer est celui de la structure de l'entretien. De par sa construction, l'EVD respecte un grand nombre de critères précédemment évoqués notamment : les mêmes questions sont posées à chaque candidat ; la possibilité de partager l'entretien à plus d'un recruteur ; aucune sous-question exploratoire n'est posée ; les questions du candidat ne sont pas autorisées jusqu'à la fin de l'entretien ; l'entretien est enregistré limitant le biais de mémoire. Le deuxième constat est l'aspect asynchrone de l'outil et son caractère *à sens unique* : le candidat répond aux questions du recruteur sans avoir un retour immédiat sur comment son discours a influencé le recruteur. Ensemble, ces caractéristiques peuvent largement avoir un impact sur la validité de l'outil de présélection, son acceptation par les candidats et les recruteurs ainsi que les stratégies auxquelles le candidat peut avoir recours.

Dans la suite de ce chapitre, nous présentons les thématiques courantes de recherche autour de l'entretien d'embauche et leur application à l'entretien vidéo différé.

2.3 Fiabilité des entretiens d'embauche

La fiabilité des entretiens d'embauche se mesure par rapport au taux d'accord entre annotateurs. La fiabilité inter-annotateur est le plus souvent mesurée au travers du κ de Cohen (dans le cas deux annotateurs), de Fleish (dans le cas de plus de deux annotateurs), du coefficient de corrélation r de Pearson, du α de Krippendorf ou de la corrélation IntraClasse (ICC). Ces mesures nous informent sur la capacité de plusieurs recruteurs à évaluer de la même façon la performance des candidats à travers l'outil de l'entretien. C'est aussi une limite haute de la validité de l'outil. Comme évoqué précédemment, la fiabilité des entretiens d'embauche structurés est nettement supérieure aux entretiens non structurés [Levashina et al., 2014]. [Huffcutt et al., 2013] montre ainsi dans une méta-analyse que la structure de l'entretien augmente le coefficient de fiabilité de 0.36 à 0.76. Cependant, malgré l'aspect inhérent structurel de l'entretien vidéo différé, peu d'études se sont intéressées à quantifier la fiabilité de tels outils. À notre connaissance, les seuls travaux disponibles sont issus des bases de données issues de l'informatique affective répertoriant les taux d'accords entre annotateurs lors de leur processus d'annotation. Ainsi, [Rasipuram and Jayagopi, 2018] répertorient un taux d'accord similaire pour l'annotation des compétences sociales en EVD et en entretien face-à-face et [Chen et al., 2017] obtiennent un coefficient de corrélation intra classe égal à 0.79 pour l'employabilité et à plus de 0.90 pour chacun des traits de personnalité. À noter que ces corpus sont issus d'entretiens simulés (il n'y a pas de réel poste à pourvoir, pour plus d'informations voir tableau 3.1 p. 26), ce qui pourrait potentiellement

influencer ledit score de fiabilité. Enfin, une dernière étude [Singhania et al., 2020], en situation réelle, présente un taux de corrélation de Pearson de 0.67 pour les dimensions de l'engagement et des compétences sociales (mesurées par un questionnaire sur l'expressivité du candidat). Globalement, l'outil de l'EVD semble suffisamment fiable afin d'être utilisé dans un processus de sélection.

2.4 Validité des entretiens d'embauche

La validité des entretiens se mesure par validité selon plusieurs critères. Ainsi, on mesure la corrélation de la note attribuée à la suite d'un entretien d'embauche avec ses futures performances au travail ou les performances actuelles évaluées selon un autre outil (manager, résultat quantitatif, etc.). Les précédents travaux et méta-analyses ont montré largement que les entretiens structurés possèdent une meilleure validité que les entretiens non structurés (voir tableau 2.1). Concernant l'entretien vidéo différé, à notre connaissance, une seule étude s'est intéressée à la validité de l'outil [Gorman et al., 2018]. Cette étude montre que la dimension "connaissances et compétences" évaluée en EVD est corrélée à 0.48 avec la performance au travail auto évaluée. De plus, les dimensions "aptitude intellectuelle", "compétences sociales", "conscienciosité" sont statistiquement significatives avec un taux de corrélation entre 0.26 et 0.36. Dans l'ensemble, l'étude montre que les EVD sont un outil suffisamment valide pour les organisations, notamment pour la phase de pré sélection.

2.5 La validité des attributs

La validité des attributs en entretien d'embauche sert à déterminer quels attributs sont responsables de l'évaluation de la performance en entretien et quels attributs sont effectivement mesurables en entretien. Ces étapes conduisent aussi à comprendre et à théoriser d'une meilleure façon les liens entre dimensions évaluées en entretien et mesurées lors de l'examen de performances [Hamdani et al., 2014]. Tout d'abord je rappelle que les critères mesurés pendant le processus de sélection doivent être spécifiquement choisis après une analyse de poste, et donc seront probablement différents pour chaque position [Voskuijl, 2017]. De plus, des référentiels métiers sont apparus ces dernières années et constituent une aide intéressante à la sélection des compétences à évaluer en entretien d'embauche (par exemple O*NET¹ ou ROME²). Néanmoins, nous pouvons explorer quelles dimensions sont

1. <https://www.onetonline.org/>

2. <https://www.data.gouv.fr/en/datasets/repertoire-operationnel-des-metiers-et-des-emplois-rome/>

le plus souvent évaluées grâce à l'outil de l'entretien d'embauche. C'est ce à quoi s'est intéressé [Huffcutt et al., 2001], démontrant que les entretiens tendent à mesurer généralement l'aptitude intellectuelle, les connaissances et compétences professionnelles, les compétences sociales et les traits de personnalité. Une différence est aussi visible entre entretien structuré et non structuré, les premiers évaluant plus souvent les compétences sociales que les seconds. Au-delà, de ces attributs, la performance de l'entretien peut être influencée par de nombreux facteurs latents. [Huffcutt et al., 2011] propose un cadre pour expliquer les performances en entretien avec une vue centrée sur le candidat. Il s'intéresse, au travers de ce modèle, aux antécédents influençant comment le candidat délivre sa réponse que ce soit les réponses aux questions, la façon dont il délivre son discours (prosodie) ou ses comportements non verbaux (expressions faciales, postures, etc.). Ainsi au-delà des attributs liés au poste, quatre principaux groupes d'influences ont été proposés nommément la dynamique recruteur-candidat (et notamment l'efficacité sociale du candidat), la préparation aux entretiens d'embauche, l'état mental du candidat (motivé ou anxieux [Feiler and Powell, 2016]), et les caractéristiques individuelles du candidat (personnalité, aptitude intellectuelle, éducation, etc.). Enfin, le modèle considère la médiation de facteurs exogènes telle que le design de l'entretien (niveau de structure, modalité de l'entretien) ou les caractéristiques personnelles et démographiques (ethnie, genre, attractivité).

Du point de vue des EVD, il a été supposé que certains attributs étaient plus influents que d'autres, de par le cadre spécial de l'outil d'évaluation. Tout d'abord, les remarques s'appliquant aux entretiens structurés s'appliquent aussi à l'EVD privilégiant probablement l'évaluation des compétences sociales. Par exemple, [Torres and Gregory, 2018] montrent que les attributs de "communication" et de "résolution des problèmes" sont plus influents que les attributs "orientation client" et "aptitude à fournir des résultats" pour un poste de manager dans le secteur hôtelier.

[Torres and Mejia, 2017] émettent l'hypothèse que l'apparence physique aurait une influence plus importante en EVD. Ceci serait justifié à cause de l'accessibilité à l'image du candidat tôt dans le processus de sélection et la possibilité de passer les entretiens rapidement. Les résultats de [Torres and Gregory, 2018] indiquent que l'esthétisme du candidat est un facteur contribuant, mais pas principal. [Suen et al., 2019a] montrent de même que l'esthétisme a une influence, mais qu'elle est moins importante que dans le cadre des entretiens vidéos synchrones.

Enfin, l'aspect structurel et asynchrone de l'EVD influence grandement la dynamique recruteur-candidat. Ainsi, il est probable que l'utilisation des stratégies de la gestion d'impression soit moins efficace qu'en entretien synchrone à cause de l'impossibilité d'avoir accès aux comportements verbaux et non verbaux du recruteur.

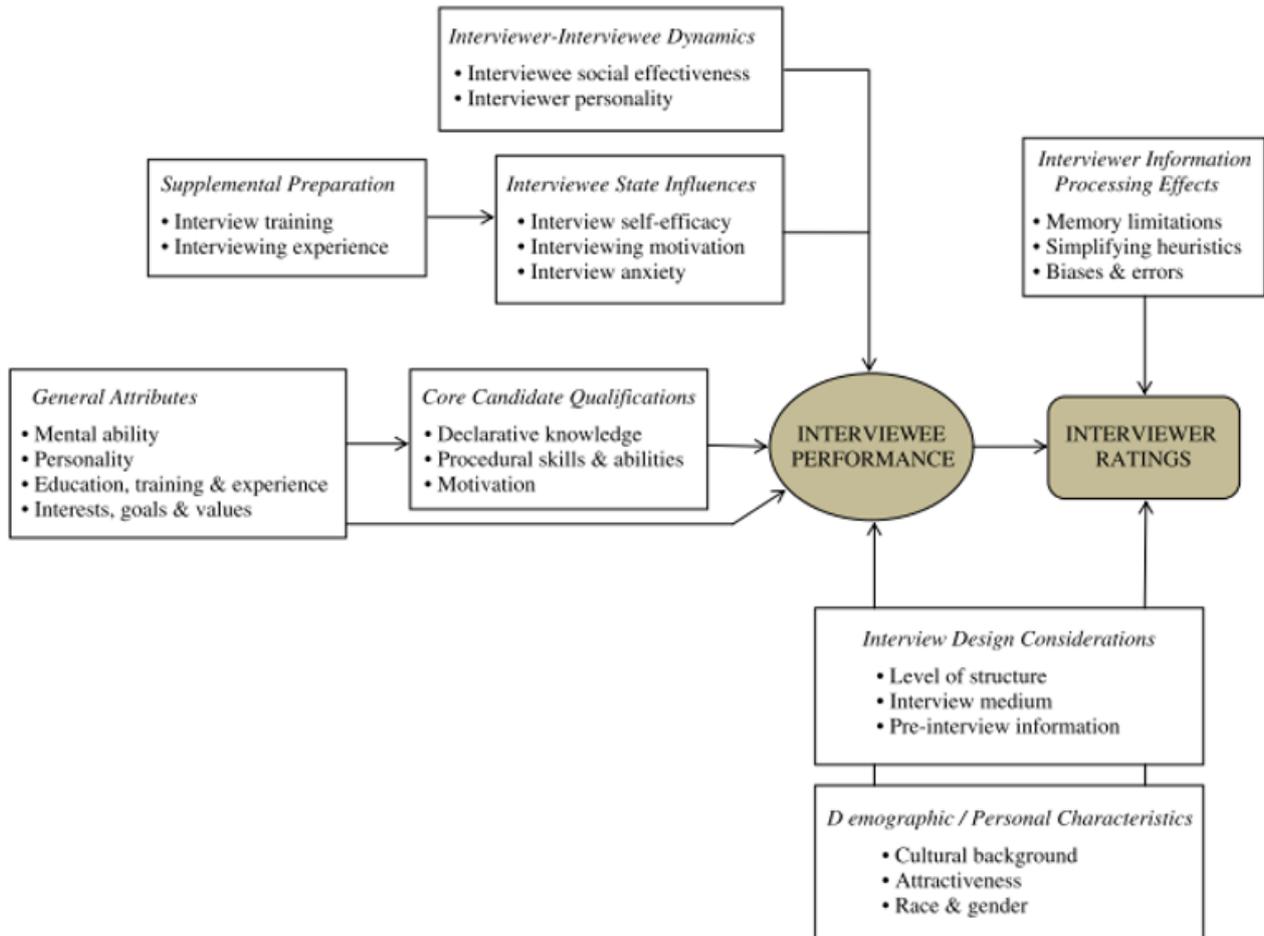


FIGURE 2.2 – Modèle de la performance des candidats en entretien d'embauche centré sur ses attributs proposé par [Huffcutt et al., 2011].

2.6 La gestion de l'impression

La gestion de l'impression se définit par le processus par lequel les candidats essayent de contrôler l'impression qu'ils renvoient. En entretien d'embauche, cette gestion de l'impression est largement utilisée et influence le résultat de l'entretien [Barrick et al., 2009]. Il a été montré que les gestions de l'impression ont une influence plus importante dans les entretiens non structurés que dans les entretiens structurés. De nombreuses gestions de l'impression ont été proposées par exemple : les tactiques de promotion qui consistent à revendiquer des succès ; les tactiques non verbales comme l'utilisation du sourire ou du regard ; les tactiques de flatterie où le candidat essaye d'influencer le recruteur en le flattant ; les tactiques de défense où le candidat se justifie ou s'excuse en cas d'échecs ; les tactiques de protection de l'image où le candidat omet certaines informations. Ces stratégies d'impression peuvent être parfois honnêtes ou fallacieuses [Roulin et al., 2014]. Une interrogation demeure dans

le fait que ces stratégies d'impression sont nuisibles ou non à l'évaluation en entretien d'embauche, notamment dans le cas où elles sont parfois nécessaires pour le poste (consultants ou commercial par exemple). Concernant les entretiens vidéos asynchrones, il semblerait que l'influence de ces stratégies d'impression soit moins importante qu'en entretien d'embauche et ce pour plusieurs raisons. Premièrement, le caractère structuré limite l'utilisation de telles stratégies. Deuxièmement, le candidat n'a pas accès aux réactions du recruteur, limitant sa possibilité d'adapter ses stratégies d'impression à son interlocuteur. Cette hypothèse se confirme à deux reprises où les candidats reportent qu'ils considèrent plus restreint leur utilisation de gestion de l'impression en EVD qu'en entretien d'embauche [Basch et al., 2020, Muralidhar et al., 2020].

2.7 Discrimination en entretien d'embauche

La discrimination lors des entretiens d'embauche a été largement étudiée. Tant d'un point de vue juridique que théorique, il est indiqué que les décisions ne doivent être prises qu'en fonction des dimensions nécessaires pour le poste [Levashina et al., 2014]. Cependant, il a été montré que des facteurs tels que le sexe, l'origine ethnique, l'apparence physique ou l'obésité ont une influence, même si la prise de décision sur ces facteurs est expressément interdite par la loi en Europe et aux États-Unis [Schmitt, 2012]. Il devient donc important, pour des raisons juridiques, de déterminer quand il y a une réelle discrimination, et de ce point de vue, la méthodologie d'évaluation peut différer d'un pays à l'autre [Sánchez-Monedero et al., 2020]. Par exemple, aux États-Unis, la règle des 4/5ème proposée par l'EEOC (Equal Employment Opportunity Commission) stipule que le ratio du groupe le plus favorisé par rapport au groupe le moins favorisé ne doit pas être inférieur à 0,8. En comparaison, cette règle ne peut pas être utilisée dans certains pays européens où il est interdit de collecter des attributs sensibles tels que l'ethnicité [Lieberman, 2001]. Dans le cas où l'analyse statistique sur les groupes protégés est impossible, des pays comme la France utilisent des tests de discrimination : ils comparent le comportement d'un tiers envers deux personnes ayant exactement le même profil pour toutes les caractéristiques pertinentes, sauf celle qui est soupçonnée de donner lieu à une discrimination. La structure de l'entretien semble limiter l'influence des biais liés au genre ou à l'ethnie [Levashina et al., 2014]. Cependant, dans le contexte des EVD, le fait d'avoir accès à l'image des candidats dès le début du processus de sélection peut contribuer à accroître l'influence de l'esthétisme, du sexe ou de l'origine ethnique des candidats. À notre connaissance, une seule étude s'est intéressée à la discrimination dans les EVD. Elle n'a trouvé aucune preuve de discrimination

dans l'évaluation en fonction du sexe ou de l'ethnie [Kroll and Ziegler, 2016]. Globalement, la modalité de l'EVD et son influence sur la discrimination en entretien d'embauche reste une question ouverte.

2.8 Acceptabilité des candidats

L'acceptabilité des candidats est importante dans le sens où elle conditionnera l'applicabilité de l'outil. Ainsi, le processus de sélection influence grandement l'attractivité de la compagnie et le taux d'acceptation des candidats aux offres d'emploi [McCarthy et al., 2017]. L'entretien d'embauche reste l'outil de présélection le plus apprécié des candidats [Salgado, 2017]. Bien que de nombreuses études se soient intéressées aux réactions positives ou négatives des candidats aux entretiens structurés, les résultats ne se sont pas encore montrés consistants [Levashina et al., 2014]. Concernant les EVD, la problématique de l'acceptabilité a été étudiée dans plusieurs études. Globalement, les candidats sont largement défavorables aux entretiens vidéo différés comparés aux entretiens face à face [Brenner et al., 2016, Basch et al., 2020, Suen et al., 2019a]. Néanmoins, les candidats reconnaissent l'utilité et la facilité d'utilisation de l'outil [Brenner et al., 2016, Guchait et al., 2014]. De nombreux points négatifs ont été pointés par les candidats, notamment l'impossibilité de voir les réactions du recruteur, l'impossibilité de montrer leur vraie personnalité, le caractère impersonnel de l'outil, le sentiment étrange de fixer une caméra et la peur de ne pas avoir assez de temps ou d'un problème technique [Guchait et al., 2014]. [Basch et al., 2020] émettent l'hypothèse que les candidats sont largement défavorables envers les EVD en raison de la limitation de leur gestion de l'impression. [Poh and San, 2015] compare l'EVD à l'entretien téléphonique et montre que l'EVD est accepté au même titre que l'entretien téléphonique. [Zibarras et al., 2018] montrent que les EVD sont perçus comme équitables et reportent que les candidats pensent que l'EVD ne devrait être utilisé que pendant la phase de présélection et ne devrait pas remplacer l'entretien face-à-face. Finalement, il a été montré que des explications, à propos des entretiens structurés et la standardisation pour améliorer l'égalité, pouvaient améliorer la favorabilité des candidats envers ces entretiens [Basch and Melchers, 2019].

2.9 Comportements non verbaux

De nombreuses études ont étudié les effets des comportements non verbaux lors des entretiens d'embauche. Ainsi des indices visuels et sonores ont été montrés comme influents afin de prédire les performances lors d'entretiens d'embauche [Forbes and Jackson, 1980],

l'anxiété [Feiler and Powell, 2016], la personnalité des candidats [Degroot and Gooty, 2009] ou les stratégies de gestion d'impressions [Schneider et al., 2015]. De nombreux indices visuels tels que l'attrait physique, les gestes de la main, le sourire, le contact visuel, le hochement de tête, le mouvement de la tête, l'orientation du corps, les indices faciaux, les mouvements des jambes ont été utilisés tout au long des expériences. Par exemple, le mouvement du torse, le toucher du visage, les mouvements des jambes [Feiler and Powell, 2016], une expression neutre et un sourire moins prononcé [Gifford et al., 1985] corréleront négativement avec les performances en entretien, tandis que le contact visuel, les gestes de la main [Feiler and Powell, 2016] ou les hochements de la tête [Schneider et al., 2015] ont une corrélation positive avec les résultats de l'entretien. Des travaux comparables ont été menés pour la voix, en particulier concernant l'impact de l'amplitude et du ton de la voix, les silences ou les dysfluences verbales [Naim et al., 2018, Rasipuram and Jayagopi, 2018]. Enfin, la fréquence de ces comportements non verbaux est influencée par les caractéristiques situationnelles de l'entretien telles que la structure ou le type de questions posées (comportementales ou situationnelles), les traits de personnalité du candidat, son genre et son ethnie [Frauendorfer and Mast, 2015]. Dans le cadre des EVD, très peu de travaux se sont intéressés à l'influence des comportements non verbaux. [Rasipuram and Jayagopi, 2018] s'intéressent aux différentes corrélations des CNV en EVD ou en entretien face à face. Ils montrent que le temps de pause est corrélé négativement aux compétences communicationnelles à l'inverse du temps parlé qui est lui corrélé positivement. La vitesse de diction semble avoir une influence plus importante en EVD comparé aux entretiens face à face. Ces observations sont aussi obtenues par rapport à l'employabilité [Chen et al., 2016b]. Enfin, les comportements issus de l'expression faciale semblent moins influents que ceux issus de la façon dont le candidat s'exprime vocalement.

2.10 Conclusion

L'étude préalable de l'état de l'art de la psychologie du travail concernant l'outil d'évaluation de l'entretien d'embauche est nécessaire pour comprendre et saisir les enjeux de l'analyse automatique appliquée à ce cadre d'étude. Ainsi, cette littérature nous propose un cadre de travail déjà bien identifié pour comprendre les différents types d'entretien (structurés vs non structurés) et situer les caractéristiques de l'entretien vidéo différé.

Deuxièmement, elle nous permet d'évaluer la faisabilité des approches automatiques en nous fournissant des clés de compréhensions par rapport à la difficulté de la tâche (via les mesures de fiabilité), par rapport aux dimensions mesurables durant un entretien (via la

littérature liée à la validité des attributs), et aux comportements et stratégies influentes des candidats qui devront être pris en compte lors de notre modélisation. Enfin, le besoin d'interprétabilité et l'assurance d'une égalité dans l'analyse automatique des candidats sont d'autant plus nécessaire au regard des travaux liés à l'acceptabilité des candidats et à la discrimination en entretien d'embauche.

Section 3

L'analyse automatique appliquée aux entretiens d'embauches

Nous nous intéressons dans ce chapitre aux précédents travaux réalisés pour l'analyse automatique appliquée aux entretiens d'embauche. Nous présentons un rapide aperçu d'outils automatiques pour l'entraînement des candidats aux entretiens d'embauche.

Ensuite, nous répertorions l'ensemble des bases de données et des méthodes d'apprentissage automatique utilisées pour l'inférence automatique dans le cadre des entretiens d'embauche. Enfin, la dernière section s'attaque aux questions de l'interprétabilité et de l'équité au travers de ces travaux.

3.1 Outils automatiques pour l'entraînement aux entretiens d'embauche

L'entraînement aux entretiens d'embauche demeure un facteur influent sur la performance des candidats en entretien (voir figure 2.2 p. 19). Dans ce sens, des agents conversationnels comme MACH [Hoque et al., 2013] et TARDIS [Anderson et al., 2013] ont été proposés afin d'aider des candidats à s'entraîner à passer des entretiens d'embauche. Des agents virtuels ont aussi été construits notamment pour l'entraînement à la prise de parole en public [Batrincea et al., 2013] ou pour améliorer les capacités d'interactions [Ali et al., 2015, Tanaka et al., 2015]. En complémentarité de la construction de ces agents virtuels, des outils de feedback automatique ont vu le jour tel que Automanner [Tanveer et al., 2016], un outil qui extrait automatiquement et avertit l'utilisateur de l'utilisation de gestes parasites, Rhema [Tanveer et al., 2015], un outil aidant des individus durant une présentation à parler à la bonne vitesse et à la bonne intensité, ou ROC Speak [Zhao et al., 2017], une plate-

forme semi-automatisée donnant des retours d'informations lors d'une présentation vidéo grâce à une détection automatique des sourires ou du ton de la voix par exemple. Bien que cette thèse s'intéresse principalement à l'analyse automatique de l'employabilité, l'intégration d'une telle analyse à ces systèmes d'entraînement pourrait avoir un impact social important pour l'aide à l'entraînement des candidats.

3.2 Bases de données pour l'analyse automatique des entretiens d'embauche

Les travaux antérieurs sur l'analyse automatique des entretiens d'embauche font appel à différents corpus pour entraîner et évaluer les systèmes. Nous avons catégorisé ces corpus selon plusieurs dimensions, à savoir : le contexte de collecte et les conditions dans lesquelles le jeu de données a été constitué (p. ex. entretien face à face, asynchrone, etc.), le poste à pourvoir (poste réel ou fictif), l'origine des annotations (recruteurs, annotateurs experts ou naïfs), la population des candidats aux entretiens d'embauche et les dimensions annotées. Dans la suite de cette section, nous effectuons une description de chacune des dimensions. Nous reportons l'ensemble des jeux de données utilisés en apprentissage automatique dans le contexte des entretiens d'embauches en tableau 3.1.

3.2.1 Le contexte de collecte

Comme surligné en chapitre 2, la modalité de l'entretien peut avoir un impact fort sur la perception des candidats. Au-delà de ça, il existe des différences dans les techniques utilisées lors de la collecte de données. Comme on peut le voir dans la première colonne du tableau 3.1, les jeux de données associés aux entretiens face à face sont obtenus majoritairement en laboratoire : les capteurs sont souvent plus qualitatifs et les conditions entre entretiens sont exactement les mêmes. De plus, en dehors des capteurs classiques, il est possible d'ajouter des capteurs plus intrusifs (p. ex., [Finnerty et al., 2016] propose d'utiliser des capteurs électrodermaux pour mesurer le stress en entretien d'embauche). Cependant, l'ajout de ces capteurs en plus de l'utilisation de capteurs vidéo de qualité rend difficile l'obtention d'une base de données suffisamment grande. D'autant plus, que la prise en compte de l'aspect dyadique de l'entretien face à face reste compliquée.

L'EVD, quant à lui, permet d'obtenir des jeux de données plus importants et plus rapidement au prix d'une qualité dégradée des vidéos d'entretiens d'embauche (capteurs peu qualitatifs, problèmes de sons et images, etc.). Parallèlement à ces bases de données d'en-

3.2. BASES DE DONNÉES POUR L'ANALYSE AUTOMATIQUE DES ENTRETIENS D'EMBAUCHE

Travaux	Contexte du jeu de données	Poste ciblé	Annotateurs	Nombre de candidats et type de candidats	Dimensions annotées
[Nguyen et al., 2014, Nguyen and Gatica-Perez, 2015], [Finnerty et al., 2016]	Entretien FF	Mission marketing	AMT	62 étudiants	Employabilité, Stress
[Muralidhar et al., 2016, Muralidhar Idiap et al., 2018], [Muralidhar and Gatica-perez, 2017]	Entretien FF	Simulé	Étudiants en psychologie	169 étudiants	Employabilité, Compétences sociales et communicationnelles, etc
[Naim et al., 2018]	Entretien FF	Simulé	AMT	69 étudiants	Employabilité, Engagement, Stress, etc.
[Nguyen, 2015, Muralidhar et al., 2018]	CV Vidéo	Multiple	AMT	939 chercheurs d'emplois	Compétences professionnelles, sociales et communicationnelles
[Chen et al., 2016a, Chen et al., 2016b]	EVD	Simulé	Experts	36 employés	Employabilité, Personnalité
[Rasipuram and Jayagopi, 2016, Rao S. B et al., 2017], [Rasipuram et al., 2017b, Rasipuram et al., 2017a], [Rasipuram and Jayagopi, 2018]	EVD	Simulé	Observateurs Naïfs	106 étudiants	Compétences communicationnelles
[Muralidhar et al., 2020]	EVD	Simulé	Questionnaires	221 étudiants	Employabilité, Fairness
[Rupasinghe et al., 2017]	FF physique et médié	Simulé	Experts	36 employés	Gestion de l'impression
[Chen et al., 2017, Leong et al., 2019]	EVD	Simulé	Experts	260 AMT	Personnalité
[Suen et al., 2019b]	EVD	RH	Questionnaire	120 candidats	Employabilité, Personnalité
[Escalante et al., 2020]	Vlogs	Aucun	AMT	3060 vidéos coupées en 10 000 clips	Personnalité
[Singhania et al., 2020]	EVD	Multiple	Experts	810 chercheurs d'emploi	Emotion positive, confiance, engagement, compétences sociales

TABLE 3.1 – Résumé des bases de données en lien avec l'analyse automatique d'entretien d'embauche. FF est l'abréviation de face à face. AMT est l'abréviation de Amazon Mechanical Turk.

tretiens d'embauche, nous reportons deux jeux de données qui collectent des monologues dans un aspect un peu plus différent que celui des entretiens d'embauches : les vlogs [Escalante et al., 2020] et les CV vidéo [Nguyen and Gatica-Perez, 2016]. L'utilisation des vlogs permet une collecte importante de vidéos via l'utilisation de plateforme numérique (e.g. YouTube), vidéos reproduisant les conditions techniques de l'enregistrement des EVD. Les CV vidéos quant à eux reproduisent l'aspect monologue des EVD, mais peuvent largement être différents dans les conditions techniques (multiples plans, vidéo animée insérée, etc.).

3.2.2 Le poste ciblé

Il est important de comprendre pour quels postes les candidats postulent au travers des différents jeux de données, car ces postes conditionnent les compétences évaluées en entretien, les questions élaborées pour évaluer ces compétences [Huffcutt, 2011], et les comportements non verbaux utilisés par les candidats [Ruben et al., 2015].

Au-delà de ces éléments, il est aussi important de distinguer le caractère simulé ou non de l'entretien d'embauche qui peut avoir comme effet de ne pas reproduire les conditions réelles d'un entretien. Ainsi, dans le cadre d'un entretien simulé, un candidat pourrait voir sa motivation et son anxiété réduite par rapport aux conditions d'un entretien d'embauche réel [Huffcutt et al., 2011]. De plus, les CV vidéos sont un cas spécial où les candidats peuvent potentiellement utiliser une meilleure gestion de l'impression qu'en entretien face-à-face grâce à la possibilité de s'enregistrer à nouveau jusqu'à satisfaction. Ainsi, comme on peut le voir dans la colonne 3 du tableau 3.1 trois jeux de données constituent des entretiens où le poste à pourvoir est réel [Nguyen et al., 2014, Suen et al., 2019b, Singhania et al., 2020].

3.2.3 Les dimensions annotées

Les attributs évalués en analyse automatique des entretiens d'embauche diffèrent légèrement de ceux évoquées en section 2.5. Tout d'abord, l'*employabilité* demeure la dimension la plus annotée dans les différents jeux de données. Cette dimension réfère à la capacité du candidat à être sélectionné pour le poste, à continuer le processus de sélection ou plus simplement à sa performance en entretien. La personnalité est ensuite la dimension la plus annotée. Le modèle utilisé pour évaluer la personnalité demeure dans l'ensemble des études le modèle OCEAN ou Big 5 [Chen et al., 2017, Rupasinghe et al., 2017, Suen et al., 2019b, Escalante et al., 2020]. Les compétences communicationnelles et sociales des candidats sont la troisième dimension la plus annotée. Il est intéressant de souligner le constat suivant vis à vis du choix de ces dimensions : d'une part, elles demeurent des compétences largement

évaluées en entretien d'embauche [Huffcutt et al., 2001], d'autre part elles sont influentes sur la performance du candidat en dehors des critères orientés postes [Huffcutt et al., 2011] donc quel que soit le type de poste considéré. Enfin, il est aussi courant d'annoter l'état mental du candidat en termes d'émotions, d'anxiété, d'engagement ou de confiance de la motivation du candidat (voir colonne dimension annotée du tableau 3.1).

L'annotation manuelle de chaque segment d'une vidéo prend beaucoup de temps et peut devenir coûteuse. La granularité de l'annotation est donc une question importante à prendre en compte. L'entretien d'embauche étant structuré, il faut donc décider d'une méthode d'annotation, et plus précisément, à quel niveau faut-il annoter les entretiens d'embauche : devons-nous annoter l'entretien du candidat ou chaque réponse de celui-ci ?

De plus, il a été démontré qu'en utilisant seulement une courte quantité d'informations, les gens peuvent déduire correctement les caractéristiques personnelles, les traits ou les états d'un individu [Murphy et al., 2015, Carney et al., 2007]. Cette approche, appelée analyse en tranches fines, a déjà été utilisée dans le cadre d'études sur les interactions sociales [Murphy et al., 2015], les premières impressions [Carney et al., 2007], la prise de parole en public [Chollet and Scherer, 2017]. Récemment, l'annotation en tranches fines a aussi été réalisée dans le cadre des entretiens d'embauches [Nguyen and Gatica-Perez, 2015, Rasipuram and Jayagopi, 2018, Naim et al., 2018]. Ainsi, [Nguyen and Gatica-Perez, 2015] ont montré qu'il était possible d'inférer l'employabilité avec une seule minute d'entretien. Néanmoins, la durée et la stratégie d'échantillonnage pour les tranches fines restent une question ouverte. Les études précédentes se concentrent sur l'échantillonnage de tranches fines de manière aléatoire [Chollet and Scherer, 2017], en utilisant la structure de l'entretien d'embauche (tranches basées sur des questions et des réponses) [Nguyen and Gatica-Perez, 2015], ou au début et à la fin des interactions [Degroot and Gooty, 2009].

Les annotateurs.

L'employabilité est une étiquette compliquée à fournir, il est donc important de savoir qui a étiqueté les données. En effet, la qualité de l'annotation peut être très différente lorsqu'il s'agit d'observateurs naïfs, d'étudiants en psychologie ou experts en ressources humaines, de recruteurs impliqués dans le processus de recrutement ou enfin des décisionnaires pour le poste [Huffcutt et al., 2001]. Il est aussi courant d'obtenir des annotations via des plateformes de crowdsourcing comme Amazon Mechanical Turk. Ces plateformes permettent d'obtenir des annotations plus diverses et d'une façon plus économique, mais sont parfois de qualités moindres. Des stratégies de sélection des annotateurs sont dès lors parfois obligatoires [Naim et al., 2018].

Toutes ces annotations se concentrent sur l'évaluation des dimensions perçues par le recruteur (personnalité, entretien, etc.), mais n'évaluent pas les vérités terrain réelles des

candidats. L'utilisation des questionnaires essaie de combler ce défaut. Ainsi, [Suen et al., 2019b] essaie d'inférer les traits de personnalité de candidats reportés par questionnaire au travers de leurs comportements non verbaux en EVD. À noter qu'un biais de désirabilité sociale, notamment dans le cas d'entretiens réels, peut mener à des problèmes de récolte de données dans le cas des questionnaires [Brenner and DeLamater, 2016].

Le nombre de candidats.

Le nombre de candidats a une incidence sur les méthodes automatiques possiblement utilisables (apprentissage automatique ou apprentissage profond), sur la généralisation des études et sur la validité des analyses. Il est logique de voir que les jeux de données comprenant des monologues (EVD, CV vidéo et Vlogs) comportent un nombre de candidats beaucoup plus important que les entretiens face à face et en vidéoconférence. Ceci s'explique facilement par l'aspect asynchrone des EVD ou l'accessibilité facilitée à un grand nombre de vidéos sur les plateformes numériques telles que YouTube.

3.3 Les méthodes classiques d'apprentissage automatique

Nous présentons ici les méthodes classiques d'apprentissage automatique qui constituent la majorité des travaux précédents. Ces méthodes se découpent en trois grandes étapes : l'extraction de descripteurs, la représentation des descripteurs choisis pour obtenir un vecteur de taille fixe et une méthode de classification. Nous nous intéresserons de plus aux différentes méthodes de fusion utilisées.

3.3.1 Descripteurs

Les récentes avancées en matière de traitement automatique des signaux sociaux ont permis le développement de boîtes à outils pour l'extraction de descripteurs issus des flux audio [Eyben et al., 2016] et vidéo [Baltrusaitis et al., 2018]. Comme les entretiens d'embauche asynchrones sont des vidéos, les caractéristiques de chaque modalité (contenu verbal, audio et vidéo) doivent être extraites à un espace temps régulier afin de construire un modèle de classification. Les indices audio se composent principalement de caractéristiques de la prosodie et de l'activité vocale (fréquence fondamentale, intensité, pauses, silences, énoncés courts, etc.), de coefficients cepstraux de la fréquence mel et de qualité de la voix (jitter, shimmer, HNR, NAQ, etc.) [Nguyen and Gatica-Perez, 2015, Rao S. B et al., 2017]. Les principales boîtes à outils utilisées sont OpenSmile [Eyben et al., 2013a], PyAudioAnalysis [Giannakopoulos, 2015], Praat [Boersma and van Heuven, 2001], MIT Speech Feature Extraction Code [Pentland, 2004] et Social Signal Interpretation [Wagner et al., 2013].

3.3. LES MÉTHODES CLASSIQUES D'APPRENTISSAGE AUTOMATIQUE

Travaux	Modalité	Descripteurs	Représentation	Fusion	Classifieurs	Interprétabilité
[Chen et al., 2016b]	audio	prosodie	fonctions d'agrégations	concaténation	LR, SVM	Corrélation
	vidéo	Emotions, regard, mouvement de la tête				
[Chen et al., 2016a]	langage	transcription	LIWC	concaténation	LR, SVM	Corrélation
	audio	prosodie				
	vidéo	Emotions, regard, mouvement de la tête				
[Rasipuram and Jayagopi, 2016]	langage	transcription	LIWC, Doc2Vec	concaténation	LR, SVM	Corrélation
	audio	activité vocale, prosodie				
[Rasipuram et al., 2017b]	vidéo	AUs, Emotions	fonctions d'agrégations	concaténation	LR, SVM	Corrélation
	audio	activité vocale, prosodie				
[Rao S. B et al., 2017], [Rasipuram et al., 2017a]	langage	transcription, ASR	LIWC, POS, Dictionnaires, Statistiques	concaténation	LR, SVM, RF	Corrélation
	audio	activité vocale, prosodie				
[Rasipuram and Jayagopi, 2018]	vidéo	AUs, Emotions	fonctions d'agrégations, mots visuels	concaténation	LR, SVM, RF	Corrélation
	langage	ASR				
	audio	activité vocale, prosodie				
[Chen et al., 2016a]	vidéo	AUs, Emotions	LIWC, Dictionnaires, Statistiques	concaténation, fusion tardive	LR, SVM, RF	Corrélation, analyse en tranches fines
	langage	transcription, ASR				
[Muralidhar et al., 2020]	audio	prosodie	Mots Audio-Visuels et TF-idf	concaténation, fusion tardive	RF, SVM	Aucune
	vidéo	AUs, regard, mouvement de la tête				
	langage	ASR				
	audio	activité vocale, prosodie	fonctions d'agrégations	concaténation	LR, RF	Corrélation
	vidéo	AUs, regard, mouvement de la tête, énergie de mouvement				

TABLE 3.2 – Tableau regroupant les méthodes d'apprentissage automatique pour chaque travail de l'état de l'art.

Travaux	Modalité	Descripteurs	Représentation	Fusion	Classifieurs	Interprétabilité
[Nguyen et al., 2014], [Nguyen and Gatica-Perez, 2015], [Muralidhar et al., 2016]	audio	activité vocale, prosodie	fonctions d'agrégations	concaténation	LR, RF	Corrélation, Analyse en tranches fines
	vidéo	sourire, hochement de tête, énergie de mouvement				
	multimodale	hochement de tête pendant la prise de parole et backchannel				
[Finnerty et al., 2016]	audio	activité vocale, prosodie	fonctions d'agrégations	concaténation	LR, RF	Corrélation
	vidéo	hochement de tête, énergie de mouvement				
	EDA	Réponse électrodermale				
[Muralidhar Idiap et al., 2018]	audio	activité vocale, prosodie	fonctions d'agrégations	concaténation	LR, RF	Corrélation
	vidéo	hochement de tête, énergie de mouvement, Regard, Emotions				
	multimodale	hochement de tête pendant la prise de parole et backchannel				
[Muralidhar and Gatica-perez, 2017]	audio	activité vocale, prosodie	fonctions d'agrégations	concaténation	LR, RF	Corrélation
	vidéo	hochement de tête, énergie de mouvement, Regard, Emotions				
	multimodale	hochement de tête pendant la prise de parole et backchannel				
[Rupasinghe et al., 2017]	langage	transcription	LIWC			
	audio	prosodie	fonctions d'agrégations	Aucune	Naive Bayes, SVM, RF	Aucune
	vidéo	AUs, Emotions				
[Naim et al., 2018]	audio	prosodie	fonctions d'agrégations	concaténation	LR, SVR	importance des descripteurs, analyse en tranches fines
	vidéo	sourire, distance faciale, mouvement de la tête				
	langage	transcription				
[Nguyen and Gatica-Perez, 2016]	audio	activité vocale, prosodie	fonctions d'agrégations	concaténation	LR, RF	Corrélation
	vidéo	énergie de mouvement totale et faciale				

TABLE 3.3 – Tableau regroupant les méthodes d'apprentissage automatique pour chaque travail de l'état de l'art.

Les caractéristiques dérivées des expressions faciales (unités d'actions faciales, rotation et position de la tête, direction du regard, émotions etc.) constituent les indices visuels les plus extraits grâce à des outils tels que Imotions¹, Facet [Littlewort et al., 2011], OpenFace [Baltrusaitis et al., 2018], Affectiva Affdex [McDuff et al., 2010] ou FaceReader [Stöckli et al., 2018]. À cela s'ajoute l'énergie de mouvement issu de la tête ou de la partie supérieure du corps qui sont reconnus comme des descripteurs intéressants afin de prédire la performance en entretien [Schneider et al., 2015].

Enfin, les progrès du traitement automatique des langues et de la reconnaissance vocale ont permis aux chercheurs d'utiliser le contenu verbal [Muralidhar et al., 2018, Rasipuram et al., 2017b]. Une fois le contenu verbal obtenu (manuellement ou automatiquement), les chercheurs ont largement utilisé des statistiques lexicales (nombre de mots, nombre de mots uniques, etc.), ou des ressources lexicales externes pour extraire des descripteurs. Ainsi l'utilisation de dictionnaires tel que le LIWC (Linguistic Inquiry Word Count) [Rao S. B et al., 2017] ou construits à la main à partir de WordNet [Rao S. B et al., 2017] est courante dans la plupart des études utilisant le contenu verbal. Ces dictionnaires permettent d'obtenir une représentation plus haut niveau en comptant chacun des mots appartenant à des catégories utiles pour l'évaluation du contenu verbal en entretien telles que *l'accomplissement*, *l'utilisation du je* ou *le travail*. L'utilisation d'une représentation par sac de mots a aussi été utilisée afin de déterminer plus finement si l'usage de mots particuliers influe sur la prédiction de l'employabilité [Chen et al., 2017, Leong et al., 2019].

3.3.2 Représentations temporelles

Une fois que les caractéristiques sont extraites à temps régulier, le problème de la temporalité doit être abordé. L'approche la plus courante consiste à simplifier l'aspect temporel en agrégeant la dimension temporelle à l'aide de fonctions statistiques (*par exemple* moyenne, écart-type, etc.). D'autres fonctions d'agrégations ont aussi été proposées comme par exemple compter le nombre de moments où le tour de parole est inférieur à deux secondes (descripteur expert pour le nombre de mots remplisseurs) [Rasipuram and Jayagopi, 2018]. Ces fonctions sont issues d'expertises humaines et varient entre les différentes études. Aussi, dans un but d'unicité, de nombreux travaux utilisent des sets de descripteurs, pour la modalité audio, précédemment utilisés en informatique affective comme le set minimal eGEMAPS [Eyben et al., 2016] ou l'ensemble de descripteurs IS13 ComParE [Schuller et al., 2013]. Néanmoins, l'absence de modélisation des séquences peut entraîner la perte de certains signaux sociaux importants, tels que l'accentuation de la voix en levant les sourcils

1. <https://imotions.com/>

suivi d'un sourire [Janssoone et al., 2016]. En outre, les co-occurrences d'événements ne sont pas prises en compte par cette représentation. Ainsi, une distinction entre un faux sourire (activation de l'unité d'action 12) et un vrai sourire (activation des unités d'action 2, 4 et 12) est impossible [Ekman et al., 1990] sans modélisation des co-occurrences. Pour résoudre le problème des co-occurrences, la représentation de mots visuels, de mots audio ou de mots visuels et audio a été proposée [Chen et al., 2017, Chen et al., 2016a, Rao S. B et al., 2017]. L'idée est de considérer chaque trame comme un mot appartenant à un dictionnaire spécifique. Afin d'obtenir ce dictionnaire, un algorithme d'apprentissage non supervisé est utilisé pour regrouper les trames communes. Une fois que nous avons obtenu ces groupes, nous pouvons lier chacune des trames à un de ces groupes. Dès lors, chaque groupe représente un "mot" et nous pouvons facilement faire correspondre un ensemble de trames extraites à un document composé de ces mots. Ensuite, la tâche est traitée comme une classification de document. Ainsi, [Chen et al., 2017, Leong et al., 2019] représente par la suite les mots visuels grâce à un sac de mots visuels, tandis que [Chen et al., 2016a] obtient une représentation de l'entretien en appliquant l'algorithme Doc2Vec sur l'ensemble des mots visuels.

3.3.3 Représentations multimodales

La communication humaine étant multimodale, le fait d'envisager une représentation multimodale dans l'analyse automatique des entretiens d'embauche vidéo pourrait potentiellement améliorer les performances de ce système. La fusion multimodale peut permettre à un système d'être plus robuste, une exploitation efficace des modalités complémentaires et une solution possible au problème du bruit dans les modalités [Baltrusaitis et al., 2019]. Dans le domaine des entretiens d'embauche automatiques par vidéo, certaines tentatives ont été faites pour concevoir un système multimodal. Par exemple, la fusion précoce (concaténation des caractéristiques) et la fusion tardive (fusion de décision) ont été explorées et ont montré de meilleures performances que les modèles monomodaux pour la tâche de classification de l'employabilité [Chen et al., 2017, Naim et al., 2018, Rasipuram and Jayagopi, 2018]. Il est intéressant de noter que l'intégration d'une modalité visuelle peut dégrader les performances par rapport à la seule fusion audio-textuelle [Chen et al., 2017, Rasipuram and Jayagopi, 2018]. Une autre approche consiste à construire une représentation multimodale par l'ingénierie des descripteurs (par exemple en ajoutant un descripteur indiquant si un candidat a un contact visuel avec le recruteur tout en parlant) [Nguyen and Gatica-Perez, 2015] ou en utilisant une représentation différente [Chen et al., 2016b] (*par exemple* sac de mots audiovisuels).

3.3.4 Algorithmes de classification

Les entretiens d'embauche peuvent être annotés à différents niveaux : [Nguyen and Gatica-Perez, 2015] annotent l'employabilité à chaque réponse du candidat, [Muralidhar et al., 2020] utilise l'annotation de la performance sur tout l'entretien du candidat. D'un point de vue pratique, l'annotation manuelle de chaque réponse prend beaucoup de temps. D'un autre côté, attribuer l'étiquette de chaque réponse comme celle de l'entretien d'embauche peut être problématique. En effet, un candidat ayant obtenu un résultat négatif à un entretien pourrait avoir obtenu de bons résultats à certaines questions, menant à un bruitage des étiquettes d'entraînement. Un choix architectural d'inférence doit être fait en conséquence. Le choix qui est le plus populaire consiste à utiliser toutes les vidéos disponibles indépendamment (et donc d'attribuer l'étiquette de l'entretien) [Chen et al., 2017] pour l'entraînement du système automatique. Après obtention d'une représentation fixe suffisante, un modèle de classification ou de régression est utilisé. La régression logistique régularisée (LASSO ou Ridge), les forêts aléatoires et les séparateurs à vastes marges demeurent les algorithmes les plus utilisés.

3.3.5 Limites des approches classiques

Bien que largement représentées, les méthodes d'apprentissage classiques montrent certaines limites. Tout d'abord, des choix doivent être effectués pour chacune des étapes de construction, que ce soit dans le choix des descripteurs, des fonctions expertes d'agrégation temporelles, ou bien des classifieurs menant possiblement à des choix sous optimaux d'une étape à l'autre. Ainsi, le fait que la représentation ne soit pas apprise conjointement avec les modèles de classification peut entraîner une perte d'informations. En outre, tous les modèles ne tiennent pas compte de la séquentialité des signaux sociaux ou de la structure de l'entretien d'embauche. Ces biais inductifs pourraient largement améliorer les performances des systèmes automatiques. Enfin la fusion des modalités n'est effectuée qu'au niveau des descripteurs ou de la décision limitant la prise en compte des interactions bas niveaux entre modalités.

3.4 Les méthodes d'apprentissage profond

Les réseaux neuronaux ont fait leurs preuves dans de nombreuses tâches de Social Computing. De multiples architectures dans le domaine des réseaux de neurones ont surpassé les descripteurs construits à la main pour la détection des émotions dans les vidéos [Zadeh

Travaux	Modalité	Descripteurs	Représentation	Fusion	Classifieurs	Interprétabilité
[Muralidhar et al., 2018]	langage	transcription, ASR	LIWC, W2V, Doc2Vec, Glove	Aucune	LR, SVM	Aucune
[Leong et al., 2019]	audio	prosodie	CNN et GRU	fusion intermédiaire	MLP	Aucune
	vidéo	AUs, regard, mouvement de la tête	CNN et GRU			
	langage	ASR	TF-idf			
[Suen et al., 2019b]	vidéo	CNN	Une seule image	Absente	MLP	Aucune
[Escalante et al., 2020]	audio	prosodie	fonctions d'agrégation	concaténation	ELM	Étiquette intermédiaire
	Face	CNN	fonctions d'agrégation			
	Scene	CNN	Une seule image			
[Singhania et al., 2020]	audio	activité vocale, prosodie	fonctions d'agrégations	fusion intermédiaire	MLP	Aucune
	vidéo	CNN	CNN 1-D et attention			

TABLE 3.4 – Tableau regroupant les méthodes d'apprentissage profond pour chaque travail de l'état de l'art.

et al., 2018a], ou pour la reconnaissance d'états émotionnels [Yu et al., 2017]. Ces résultats s'expliquent par la capacité des réseaux de neurones à effectuer automatiquement des transformations utiles sur des caractéristiques de bas niveau. Ainsi, certaines architectures sont spécialement conçues pour représenter des séquences ou des structures hiérarchiques. L'utilisation de méthodes neuronales peut répondre à plusieurs enjeux dans notre contexte :

- l'utilisation, comme descripteurs bas niveaux, de descripteurs riches issus de représentation apprise sur des tâches annexes ;
- le design d'architectures neuronales plus performantes ;
- l'apprentissage de représentation permettant d'obtenir automatiquement des représentations pertinentes pour la tâche envisagée ;
- des méthodes de fusion de modalités performantes.

Il est donc naturel que la communauté de l'analyse automatique des entretiens d'embauche ait vu apparaître des travaux utilisant ces méthodes neuronales. Nous reportons l'ensemble des travaux utilisant l'apprentissage profond dans le tableau 3.4. Nous décrivons l'apport de l'apprentissage profond pour chacun des enjeux dans les prochaines sous-sections.

3.4.1 Apprentissage de représentations

L'apprentissage profond a permis l'obtention de descripteurs appris riches et généraux.

Ainsi, dans notre cas d'application, [Singhania et al., 2020, Escalante et al., 2020] ont recours à des réseaux pré entraînés pour la reconnaissance faciale (reconnaissance d'identité) afin d'extraire des descripteurs faciaux. L'utilisation de ces descripteurs permet l'utilisation

d'une information plus riche que les descripteurs experts décrits en section 3.3.1. Cependant, la contrepartie d'une telle représentation réside dans le fait que certaines informations sensibles peuvent être encodées à l'insu du réseau et utilisées par le réseau telles que le genre, l'ethnie, l'âge ou l'apparence du candidat [Bahng et al., 2020]. Il est aussi intéressant de noter que des informations issues de l'environnement du candidat (objets présents dans la chambre, localisation de l'entretien, etc.) sont utilisées dans [Escalante et al., 2020] en encodant les informations visuelles de la scène. Bien qu'il soit possible d'inférer la personnalité de par l'environnement [Gosling et al., 2002], l'utilisation de ces descripteurs peut être discutable par rapport aux possibles biais de première impression issus de cette modalité.

De façon similaire, dans le domaine de l'audio, le spectrogramme brut a été utilisée sans avoir recours à l'extraction de descripteurs [Trigeorgis et al., 2016] et des boîtes à outils pour l'apprentissage non supervisé de représentations ont vu le jour [Freitag et al., 2017]. Néanmoins, l'extraction de descripteurs experts reste la méthode privilégiée dans l'analyse automatique d'entretiens.

L'obtention de représentation textuelle des mots ou de paragraphe a connu une nette avancée avec l'apparition de l'apprentissage profond. Ainsi, il a été montré que le plongement de mots et leur apprentissage par méthode non supervisée [Mikolov et al., 2013] ou par pré entraînement à des tâches courantes [Peters et al., 2018](détection de natures grammaticales, ou d'entités nommées) permettait d'avoir une représentation riche. Dans ce sens, [Muralidhar et al., 2018] s'est intéressé à l'utilisation de ces représentations sur la transcription manuelle ou automatique du contenu verbal du candidat et a montré leur utilité pour l'inférence de compétences sociales.

3.4.2 Représentation temporelle

L'apprentissage profond a permis de fournir des architectures répondant aux besoins de l'apprentissage de représentation notamment pour la prise en compte de la séquentialité et de la temporalité [Cho et al., 2014], des aspects hiérarchiques [Yang et al.,] ou pour la prise en compte du contexte [Yang et al., 2019a]. Ainsi, ces méthodes ont pu être utilisées afin d'apprendre une bonne représentation de l'entretien d'embauche pour l'inférence de la décision grâce à un apprentissage bout à bout résolvant la problématique liée à la gestion de la temporalité.

Ainsi, [Leong et al., 2019] utilise un CNN sur la dimension temporelle avec des fenêtres de temps spécifiques à chacune des modalités, suivi d'un GRU pour modéliser la séquentialité. [Singhania et al., 2020] utilise un CNN suivi d'une fonction d'attention pour intégrer temporellement les descripteurs visuels faciaux du candidat. Finalement, il est toujours

possible d'agréger temporellement les représentations extraites d'apprentissage profond par des fonctions d'agrégation classiques [Escalante et al., 2020].

3.4.3 Représentations multimodales

La multimodalité reste un problème ouvert dans un large éventail d'applications telles que la réponse aux questions visuelles (c-à-d. répondre à une question portant sur une image) [Ben-younes et al., 2017], la reconnaissance d'actions (c-à-d. reconnaître des actions spécifiques dans une vidéo comme la danse par exemple) [Garcia et al., 2019b], la génération de descriptions vidéo (c-à-d. décrire ce qu'il se passe dans une vidéo) [Aafaq et al., 2019] ou la production automatique de résumé vidéo [Palaskar et al., 2019]. Au-delà des fusions précoces et tardives, de nombreux modèles ont été proposés pour fusionner les modalités en dehors de l'analyse automatique de l'entretien d'embauche. Ainsi, la fusion multimodale à travers les réseaux de neurones a gagné un intérêt énorme ces dernières années, en particulier dans le domaine de la reconnaissance des émotions. Des modèles basés sur les cellules de mémoire [Zadeh et al., 2018b], les mécanismes d'attention [Zadeh et al., 2018a], l'architecture des transformateurs [Tsai et al., 2019], le plongement de mots multimodaux [Wang et al., 2019b] ou la représentation jointe des modalités [Pham et al., 2019] ont montré des performances accrues sur les tâches de prédiction des émotions. Cependant, les méthodes de fusion des modalités dans le cadre d'entretiens d'embauche demeurent limitées à la concaténation des représentations cachées avant la dernière couche de décision [Escalante et al., 2020, Leong et al., 2019, Singhanian et al., 2020] limitant la prise en compte des comportements multimodaux bas niveaux.

3.5 Interprétabilité et équité pour la confiance humain-machine

La confiance entre les utilisateurs et les systèmes d'apprentissage machine est un élément crucial, en particulier lorsque leur utilisation est destinée à des applications critiques dans les domaines de la santé, de la justice ou des ressources humaines. Dans ce sens, il est crucial de comprendre l'inférence des systèmes automatiques et de s'assurer que le système automatique traite de la même façon chaque candidat sans les discriminer par rapport à des informations qu'ils ne devraient pas utiliser.

3.5.1 Méthodes d'interprétabilité

Les notions de transparence et d'interprétabilité sont parfois encore peu claires [Gilpin et al., 2019], en particulier dans notre cas, car l'étude sur l'influence des indices sociaux sur les décisions d'embauche est toujours en cours [Levashina et al., 2014]. Ainsi à la différence des travaux en interprétabilité en vision par ordinateur où la vérité terrain est connue (c-à-d on sait quels pixels correspondent à un chat), la vérité terrain pour l'analyse de l'influence des comportements verbaux et non verbaux en entretien d'embauche n'est pas établie (c-à-d les comportements influents n'ont pas été annotés par les experts). Ce qui place notre recherche dans un cadre d'étude spécifique où la vérité terrain n'est pas connue [Yang et al., 2019b].

Concernant les méthodes d'apprentissage automatique pour l'interprétabilité, nous pouvons les classer selon deux dimensions [Yang et al., 2019b] :

- la dimension de l'interprétation : le modèle fournit-il une interprétation locale ou une interprétation globale ? Autrement dit, le modèle fournit-il des clés d'interprétabilité pour chacune des décisions ou le modèle peut-il être interprété dans son ensemble, dans son fonctionnement global ?
- la méthode d'interprétation : intrinsèque au modèle, ou par une méthode posthoc. Autrement dit, le modèle est soit interprétable directement par ses paramètres, soit une méthode à posteriori indépendante du modèle tente de fournir une interprétation.

Au-delà des méthodes d'analyse par corrélation qui demeurent l'analyse la plus populaire, les modèles précédents d'analyse automatique des entretiens d'embauche étaient des modèles interprétés dans leur ensemble avec une interprétation intrinsèque. Plus précisément, ces méthodes consistent à analyser les poids appris d'une regression logistique ou d'un séparateur à vaste marge (SVM) et à regarder les descripteurs les plus influents pour ces modèles. Bien que ces modèles ont l'avantage d'être complètement interprétables, ils montrent de moins bonnes performances que les modèles neuronaux.

Cependant, l'utilisation de réseaux de neurones se fait au prix d'une opacité extrême. Plusieurs méthodes ont donc été proposées pour obtenir une décision interprétable à partir des réseaux de neurones. Une grande partie des travaux s'articule autour de méthodes posthoc pour interpréter le réseau. Les méthodes locales s'appuient principalement sur des méthodes de sensibilité ou d'analyse des gradients intégrés pour présenter les éléments les plus influents sur une image ou dans un texte pour la décision [Ancona et al., 2017]. L'utilisation de classifieurs locaux a aussi été proposée pour expliquer la classification d'une instance : un classifieur intrinsèquement interprétable est entraîné dans une région voisine au point à interpréter et est ensuite utilisé pour expliquer pourquoi le point est classé d'une

telle façon [Ribeiro et al., 2016]. Enfin, l'apprentissage mimétique (un modèle interprétable entraîné sur la sortie du modèle d'apprentissage profond) [Liu et al., 2019] ou l'exploration des états cachés [Yosinski et al., 2015] ont été proposés pour interpréter de façon globale un modèle d'apprentissage profond.

La dernière catégorie s'intéresse à la construction de mécanismes intrinsèques pour l'interprétabilité locale. Les modèles sont construits pour intégrer des mécanismes d'interprétabilité au sein des modèles. [Chen et al., 2018] exhibe les parties d'une image les plus en lien avec l'inférence de la classe, [Murdoch et al., 2018] décompose les composantes du LSTM par rapport à l'importance intrinsèque du mot ou du mot et de son contexte. Enfin, les mécanismes d'attention se sont avérés être efficaces pour mettre en évidence les informations importantes, ce qui améliore les performances et l'interprétabilité des réseaux neuronaux. Par exemple, dans la détection des états émotionnels, les mécanismes d'attention permettent de se concentrer uniquement sur les moments importants au cours de conversations dyadiques [Yu et al., 2017]. Toutefois, la plupart des études limitent l'analyse des mécanismes d'attention à la présentation d'exemples et ne procèdent pas à une analyse approfondie. En outre, la validité des courbes d'attention en tant qu'explication a récemment été remise en question [Jain and Wallace, 2019].

3.5.2 Équité dans l'apprentissage automatique

En améliorant l'interprétabilité, les concepteurs de systèmes automatiques visent à fournir une meilleure visibilité pour aider à la prise de décision et à l'expliquer.

Néanmoins, ces systèmes pourraient encore produire des résultats biaisés en défaveur des groupes minoritaires. En effet, un biais systématique pourrait être le résultat de biais réels existant initialement dans l'ensemble de données, d'une représentation inadéquate dans la construction du modèle, ou d'une sous-représentation des populations minoritaires. Récemment, l'équité dans l'apprentissage automatique ou aussi appelé *fairness* a connu une popularité croissante [Hutchinson and Mitchell, 2019] et de nombreuses initiatives ont vu le jour (Initiative européenne sur Human-Centric Machine Learning², workshop international³, conférence⁴).

Plusieurs chercheurs en analyse automatique des entretiens d'embauche ont étudié les biais qui pourraient exister dans l'ensemble de données [Escalante et al., 2020, Leong et al., 2019] ou ont évalué l'équité des résultats produits par le système automatique [Singhania et al., 2020]. Toutefois, ces études se limitent à garantir l'absence de biais dans les résultats

2. <https://ellis.eu/>

3. <https://sites.google.com/view/hcml-2019>

4. <https://facctconference.org/index.html>

du système ou dans l'ensemble des données et ne proposent aucune méthode dans le cas d'un pipeline biaisé. En outre, elles ne garantissent pas que des informations sensibles n'influencent pas la décision d'embauche du système, ce qui serait contraire aux lois sur l'égalité de traitement section 2.7.

De nombreuses mesures de l'équité ont été proposées et il n'y a pas de consensus par rapport au choix de ces métriques [Narayanan, 2018]. Le choix de ces métriques dépend largement du contexte et de l'application visée. D'une façon intéressante, les politiques de mesure de discrimination évoquées en section 2.7 p. 20 ont un impact sur les métriques choisies pour mesurer l'équité dans la littérature en apprentissage automatique : respectivement (1) *parité de classification* qui garantit que les mesures de performance sont égales entre chacun des groupes protégés et (2) *anti-classification* qui garantit qu'aucune variable protégée (race, sexe, etc.) n'est utilisée pour prendre une décision. De la même manière que les décisions humaines, les systèmes doivent être évalués [Raghavan et al., 2019].

Les approches d'équité ont été conçues pour améliorer diverses parties des systèmes d'apprentissage machine, allant du prétraitement des données ou de l'apprentissage de représentation au post-traitement des résultats. L'approche la plus simple est la repondération et le ré-étiquetage d'instances spécifiques de l'ensemble de données, mais cette approche conduit souvent à des résultats incohérents [Agarwal et al., 2018]. Une autre approche consiste à prétraiter les données afin de supprimer les informations corrélées aux attributs sensibles tout en préservant les informations de la représentation originale [Zemel et al., 2013, Calmon et al., 2017, Louizos et al., 2015]. Cependant, ces méthodes sont la plupart du temps appliquées sur une représentation fixe, ce qui conduit à une utilisation difficile sur les séquences multimodales qui sont typiques dans l'analyse des EVD. Enfin, les méthodes de post-traitement les plus répandues se concentrent sur la modification du seuil des classifieurs en fonction des classes minoritaires. Toutefois, l'application de ces méthodes peut être limitée, car la discrimination positive est interdite dans plusieurs pays (par exemple, en France, au Royaume-Uni ou en Allemagne). Récemment, un type d'approche prometteur a été proposé, qui se concentre sur la suppression des informations sensibles de la représentation neuronale pendant l'étape d'entraînement du réseau par méthodes adversaires [Louppe et al., 2017, Madras et al., 2018, Delobelle et al., 2020]. L'idée est d'obtenir une représentation sans aucune information sensible identifiable assurant le système ne pas prendre en compte cette information pendant l'inférence. Il est intéressant de noter que ces méthodologies sont également liées à des méthodes de protection de données privées. Les concepteurs de réseaux ont essayé de protéger leur système contre des attaquants qui essaieraient de récupérer des informations personnelles à partir d'une représentation latente [Tripathy et al., 2019]. Ainsi, l'apprentissage adversaire a été utilisé dans le contexte de

la reconnaissance automatique de la parole [Aloufi et al., 2020, Srivastava et al., 2019], de la détection de sentiments dans les tweets [Elazar and Goldberg, 2018], de reconnaissance visuelle [Wang et al., 2019a], ou de détection multimodale (prosodie et contenu verbal) des émotions [Jaiswal and Mower Provost, 2020].

3.6 Conclusion

Au travers de ce chapitre, nous avons synthétisé l'état de la recherche portant sur l'analyse automatique des entretiens d'embauche. En ce sens, nous avons décrit les bases de données, les méthodes classiques d'apprentissage automatique et les méthodes émergentes d'apprentissage profond utilisées pour la prédiction automatique de compétences en entretien d'embauche. Cette vue d'ensemble des principaux travaux menés nous a permis de dresser un constat de l'état de l'art sur plusieurs points :

1. Tout d'abord, la majorité des bases de données sont relativement petites en termes de nombre de candidats limitant l'utilisation d'approches d'apprentissage profond. De plus, peu de ces bases de données ont été collectées dans un contexte d'emploi : avec un vrai poste à la clé, annoté par les recruteurs en charge du recrutement et des vrais candidats au poste en question. L'utilisation de l'entretien vidéo différé pourrait permettre l'obtention d'une base de données conséquente en nombre de candidats et dans des conditions réelles.
2. Au contraire des travaux utilisant les méthodes classiques d'apprentissage automatique, les quelques approches d'apprentissage profond ne fournissent aucune interprétabilité de leur système.
3. Aucune des méthodes présentées ne considère le contexte de l'entretien d'embauche et plus précisément sa structure et les questions posées lors de l'entretien.
4. Les méthodes de fusion se résument souvent à une fusion précoce ou tardive des représentations de chacune des modalités.
5. Les seuls travaux s'intéressant à l'égalité de jugement des candidats se limitent à l'évaluation de leur système sans la proposition d'une méthodologie pour l'amélioration de leur système.

À la lumière de ces points, notre travail consiste à apporter de premières réponses à ces enjeux. Nous proposons tout d'abord la sélection d'un jeu de données conséquent dans un contexte réel en collaboration avec un acteur industriel en chapitre 4. Nous proposons un modèle d'apprentissage profond adapté à la structure des entretiens vidéo différés en

chapitre 5. En chapitre 6, nous améliorons notre précédent modèle en intégrant une fusion multimodale adaptée à l'utilisation d'un contenu verbal obtenu par retranscription automatique. Le chapitre 7 s'intéresse à fournir des éléments d'interprétation liés aux modèles d'apprentissage profond entraînés pour la tâche de convocabilité. Enfin, nous proposons dans le chapitre 8 une méthodologie pour l'obtention d'une analyse automatique plus égalitaire entre candidats.

Section 4

Matériel

Dans ce chapitre, nous présentons la plateforme du partenaire industriel EASYRECRUE utilisée pour constituer nos deux jeux de données. Nous décrivons ensuite nos choix pour la sélection d'un type de poste particulier. Nous justifions notre choix méthodologique quant à l'attribution des étiquettes de convocabilité pour chacun des candidats. Enfin, nous présentons notre motivation dans le choix des descripteurs à extraire et le traitement opérationnel de cette étape.

4.1 Présentation de la plateforme

L'utilisation de l'EVD se constitue de trois principales étapes. Premièrement, le recruteur construit sa campagne de recrutement (définition des questions de l'entretien, temps alloué au candidat par réponse, etc.). Puis, les candidats invités à le faire complètent leur EVD. La dernière étape consiste en l'évaluation des candidats par le recruteur sur la plateforme(Figure 4.1). Nous présentons chacune de ses étapes en détail ci-dessous en sous-section 4.1.1, 4.1.2, 4.1.3.

4.1.1 Le processus de construction de campagne de recrutement

La construction du guide d'un entretien vidéo différé n'est pas très différente de celui d'un entretien structuré. Suite à une précédente analyse de poste, le recruteur a mis en exergue les différentes compétences requises pour le poste concerné (voir section 2.5 p. 17). Une fois cette étape effectuée, le recruteur précise les différents critères de sélection, et les questions qui seront posées pendant l'entretien. Il en est de même pour la création d'une campagne de recrutement sur la plateforme EASYRECRUE. Ainsi, le créateur d'une campagne doit spécifier les informations suivantes :

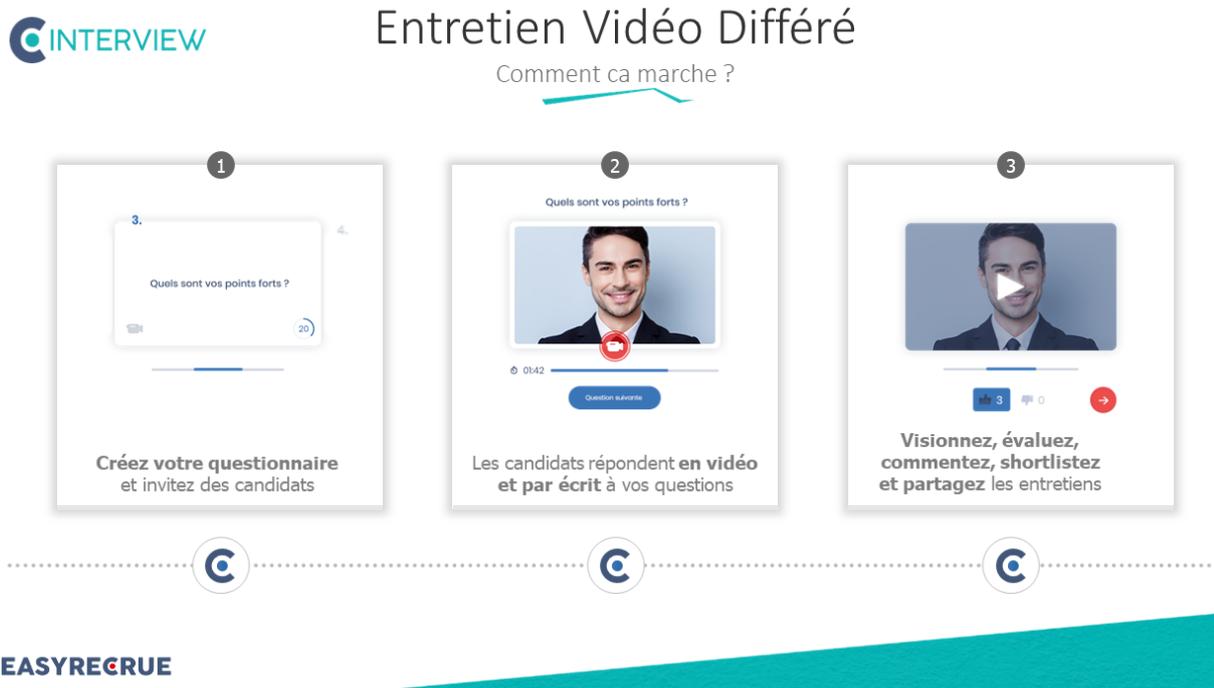


FIGURE 4.1 – Diapositive de présentation de l'EVD par la société EASYRECRUE

1. Le nom de la campagne de recrutement
2. Les critères d'évaluation (p. ex. des dimensions choisies par le recruteur au travers desquelles il évaluera ce candidat)
3. La liste des questions de l'entretien (p. ex. À quelles questions le candidat devra-t-il répondre ?)

Nous décrivons dans les sous-sections chacune des étapes de création de la campagne.

Caractéristique générale de la campagne

La première étape de la construction de la campagne consiste à définir l'intitulé du poste pour lequel la campagne de recrutement est créée. Ce titre se retrouve sur l'email automatique lorsqu'un candidat est invité et sur l'interface utilisateur candidat lorsque celui répond à l'entretien. Il est donc nécessaire que le titre choisi représente assez bien la position décrite dans l'offre d'emploi. Lors de cette étape, le recruteur peut aussi exiger des documents nécessaires tels qu'un Curriculum Vitae ou une lettre de motivation. Finalement, la langue pour l'interface utilisateur candidat est sélectionnée.

The screenshot shows the 'EASYRECRUE' interface. At the top, there is a search bar with the text 'Rechercher un candidat, une campagne de recrutement...' and a user profile 'Bonjour leo'. Below the search bar, there are navigation tabs: 'Tableau de bord', 'Engage', 'Interview', 'Assess', 'Aide', and 'Réglages'. The main content area is titled 'Thèse Informatique Affective' with the identifier '#6196 par Leo H'. There are three tabs: 'Général' (selected), 'Différé', and 'Live'. Under the 'Général' tab, there are three sections: 'Titre' with a text input field containing 'Thèse Informatique Affective'; 'Langue du candidat' with a dropdown menu set to 'Français'; and 'Documents demandés' with an '+ Ajouter' button.

FIGURE 4.2 – Spécifications des informations générales de la campagne.

The screenshot shows the 'EASYRECRUE' interface for the 'Thèse Informatique Affective' campaign. The title and identifier '#6196 par Leo H' are at the top. There are three tabs: 'Général', 'Différé' (selected), and 'Live'. Below the tabs, there are three sub-sections: 'Questionnaire', 'Critères d'évaluation' (selected), and 'Invitation et accueil'. Under 'Critères d'évaluation', there are four criteria listed in a table-like structure:

Critère n°1	Cohérence du parcours	
Critère n°2	Motivation pour le poste	x
Critère n°3	Connaissances informatique affective	x
Critère n°4	Compétences sociales	x

Below the table is a button '+ Ajouter un critère'. At the bottom, there are two navigation buttons: 'Précédent' and 'Suivant'.

FIGURE 4.3 – Spécifications des critères d'évaluation de la campagne.

Critères d'évaluation

Tout comme pour un entretien structuré, une liste de critères d'évaluation peut être choisie. Un recruteur peut enregistrer une liste d'une taille maximale de 10 critères (voir figure 4.3). Chacun des critères est entré manuellement par le recruteur en remplissant un champ de texte libre.

Questions de l'entretien

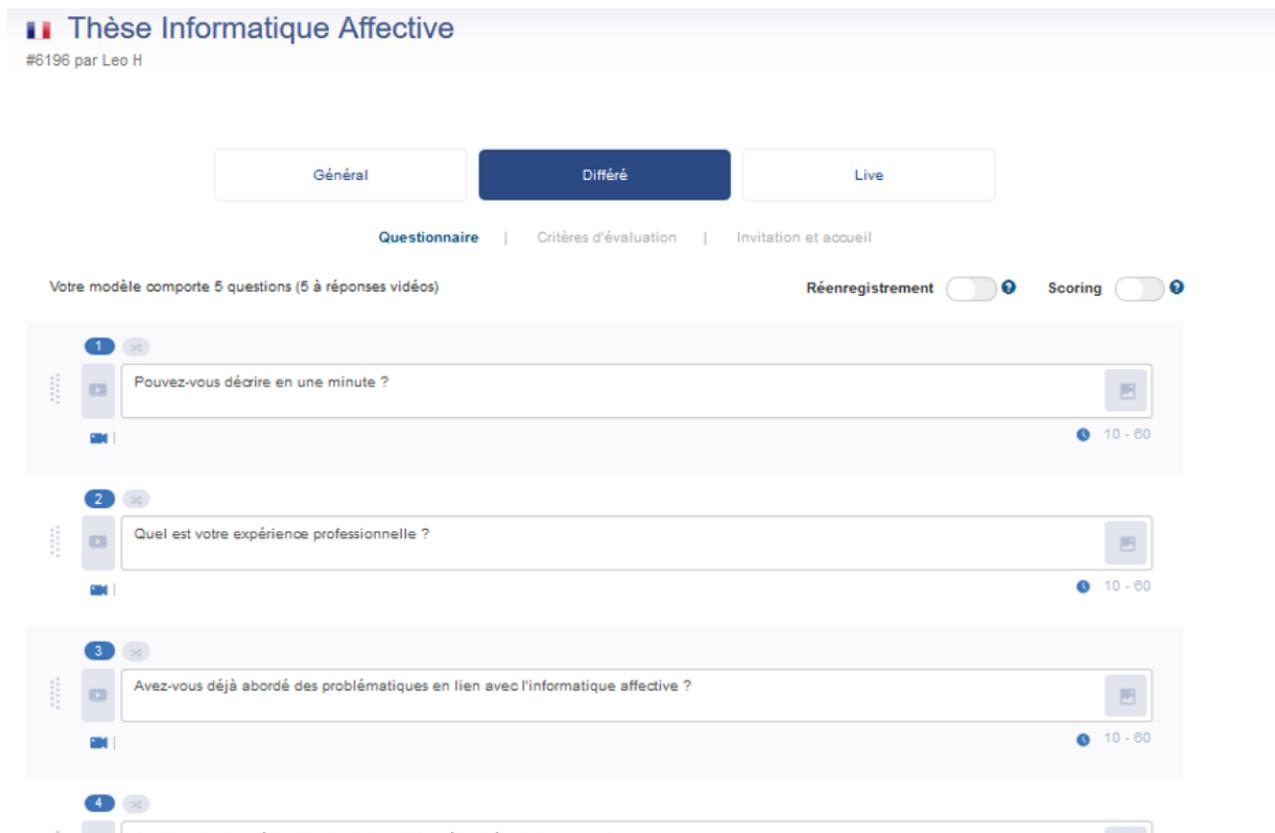


FIGURE 4.4 – Spécifications des questions de la campagne.

Une fois les critères d'évaluation choisis, le recruteur crée la liste des questions constituant l'entretien. L'interface utilisateur de la création des questions est disponible en Figure 4.4.

Concernant les questions, le recruteur peut choisir entre plusieurs types de questions. Elles sont au nombre de quatre.

- Questions à réponses vidéos : les candidats répondent par une réponse vidéo en s'enregistrant via une webcam, un téléphone ou une tablette.
- Questions à réponses à choix multiples : les candidates répondent en cochant les réponses qu'ils trouvent les plus pertinentes à la question posée.

- Questions à réponse champ libre : les candidats répondent avec un champ texte libre à la question posée.
- Questions à réponse calendrier : les candidats indiquent une date via un calendrier. Ce type de réponse est le plus souvent lié à des questions temporelles (disponibilité du candidat, fin d'un diplôme, etc.).

Chaque question peut être accompagnée d'une photo ou d'une vidéo, permettant d'ajouter du contenu contextuel à la question (pouvez m'expliquer ce graphique ? Le schéma électrique présenté est-il correct ? etc.). La question peut aussi être posée en vidéo (Vidéo du recruteur posant la question) pour humaniser l'entretien [Lukacik et al., 2020].

Pour chaque question, le recruteur configure deux durées, nommément la durée dite de préparation et la durée dite de réponse. La durée de préparation consiste au temps pour lequel le candidat est autorisé à prendre des notes et à réfléchir avant de répondre à la question. La durée de réponse est le temps maximal du candidat pour répondre à la question posée. Il n'y a pas de limites dans le nombre de questions.

L'ordre des questions peut être choisi par le recruteur ou randomisé selon ses préférences. Enfin, le recruteur peut autoriser ou non le candidat à réenregistrer ses réponses.

4.1.2 Le processus de candidature

Une fois l'invitation à l'entretien reçue par email, le candidat a la possibilité de passer son entretien d'embauche quand il le souhaite sous couvert de la date d'expiration de l'entretien (en général 7 jours). Une liste de conseils est fournie au candidat pour passer l'entretien dans de bonnes conditions comme : vérifier que l'endroit soit calme et bien éclairé, opter pour un fond neutre et une tenue vestimentaire adaptée.

Après l'acceptation des conditions d'utilisation des données et du processus de traitement, le candidat accède à une plateforme d'entraînement. Le candidat est amené à vérifier techniquement si son matériel est fonctionnel (son trop bas, visualisation de ce que voit le recruteur, etc.). À la suite de cette vérification, les candidats peuvent s'entraîner à l'exercice de l'EVD et mieux comprendre le rendu final accessible au recruteur. Cet entraînement peut être effectué autant de fois qu'il le souhaite. Le candidat effectue ensuite son entretien d'embauche comme configuré précédemment. À noter qu'il est possible d'effectuer des pauses entre les questions et qu'en cas de problèmes techniques, un centre d'appel est disponible. Les entretiens peuvent être enregistrés à partir d'une webcam, d'un smartphone ou d'une tablette. Ainsi, les environnements bruyants et les équipements de mauvaise qualité sont fréquents et rendent la qualité des vidéos variable.

4.1.3 Le processus d'évaluation des candidats

Une fois que le candidat a effectué son entretien d'embauche, le recruteur peut le visionner sur la plateforme EASYRECRUE. L'interface utilisateur est disponible en Figure 4.5. Les différentes questions sont présentes sur la partie gauche de l'écran (visible en bleu sur la figure) et les réponses vidéos sur la partie droite (visible en rouge sur la figure). Ensuite, le recruteur a la possibilité d'évaluer de plusieurs façons le candidat, il peut :

- Déplacer le candidat vers un autre dossier ('Shortlist' ou 'Archive'). Ces dossiers servent souvent à séparer les candidats avec lesquels le recruteur veut continuer le processus de recrutement ou d'archiver des candidats qui ne sont plus disponibles ou qui n'ont pas convenu pour le poste concerné. Visible en violet sur la figure.
- Évaluer selon les critères d'évaluations précédemment choisis avec une note allant de 0 à 5 étoiles. Visible en vert sur la figure.
- Cliquer sur un bouton "j'aime" ou "je n'aime pas". Visible en orange sur la figure.
- Envoyer un mail d'acceptation ou de refus aux candidats. Visible en jaune sur la figure.
- Échanger avec les autres collaborateurs sur la plateforme via commentaires. Visible en gris sur la figure.

Chacun des collaborateurs peut évaluer individuellement le candidat. Seul le dossier où se trouve le candidat (archive, courant ou shortlist) est commun entre tous les évaluateurs.

4.2 La sélection des jeux de données

Nous avons décrit précédemment la plateforme d'EASYRECRUE et plus précisément comment les recruteurs construisent leur entretien, comment les candidats complètent leur entretien et comment les recruteurs les évaluent. L'un des objectifs de cette thèse est d'examiner l'influence des signaux sociaux de façon monomodale ou multimodale sur la convocabilité. Dans ce sens, nous avons construit deux jeux de données, le deuxième étant spécialement adapté aux expériences multimodales.

Dans cette section, nous présentons la procédure de collection de ces deux jeux de données. Pour cela, nous présentons dans les sous-sections la méthodologie mise en place pour 1) sélectionner le corpus de réponses vidéos, 2) effectuer l'étiquetage des données et 3) extraire les séquences de descripteurs.

Leo Hemamo
 | 1.hemamou+4@gmail.com | |

POSTE: Internship machine learning | STATUT: En cours | DERNIÈRE ACTION: 11/06/2019 à 18:49

DIFFÉRÉ LIVE

PROGRESSION CANDIDAT ▶
 Fin entretien 11/06/2019 à 18:32

QUESTION N°2 | Réponse vidéo
 Tell us about a challenge you have overcome and how you did it.

QUESTION N°3 | Réponse vidéo
 Why are you applying for this job?

QUESTION N°4 | Réponse vidéo
 What do you think define good customer service? How about a successful sale?

QUESTION N°5 | Réponse vidéo
 How do your friends describe you?

QUESTION N°6 | Réponse vidéo
 Is there anything else we should know about you?

Évaluation de l'entretien | Note moyenne : 4 / 5

Communication skills ★★★★☆	Technical knowledge ★★★★☆
Social skills ★★★★★	

AVIS ET COMMENTAIRES

Leo H 02/07/2020 à 11:25
 Je pense que c'est un candidat intéressant pour le poste. A voir, s'il réussit le business case.

Publier

Archiver
 Shortlist
 Inviter à un live
 Partager la fiche
 Obtenir un lien
 Écrire un email
 Inviter

FIGURE 4.5 – Interface recruteur pour l'évaluation des candidats

4.2.1 Le choix d'un unique type de métier

Au-delà des études de l'état de l'art, l'EVD a été utilisé en pratique par de nombreuses sociétés et pour recruter des candidats sur des positions très diverses [Lukacik et al., 2020]. Dans un souci d'unicité et de simplification des travaux de recherche, nous avons préféré nous intéresser à un seul type de poste. En effet, une base de données contenant de nombreuses positions pourrait potentiellement mener à des problèmes de domaine d'adaptation : un parcours académique peut très bien convenir pour un poste de marketing, mais pas du tout pour un poste de RH par exemple. De plus, les critères d'évaluation sont très différents d'un poste à l'autre, on cherchera par exemple une excellente communication pour un commercial, compétence qui sera moins prioritaire pour un poste de développeur. Enfin, les signaux sociaux influents peuvent être différents selon la position choisie (Blue collar vs White Collar, poste junior vs poste de manager) [Ruben et al., 2015]. Afin de réduire la complexité de la tâche, nous avons choisi de nous limiter à un seul type de poste.

Dans ce but, nous avons mené une étude exploratoire de la base de données et différentes discussions avec les responsables relations clients. Il est apparu que trois grandes catégories de postes existaient au sein de la base de données nommément : les postes commerciaux / relations clients, les postes liés à des positions marketing et les postes liés à des positions en ressources humaines. Nous avons choisi de nous focaliser sur le premier type de poste : les postes commerciaux / relations clients. Ce choix a été motivé par plusieurs raisons. Premièrement, c'est un poste transverse à de nombreux secteurs d'activités et/ou sociétés, ce qui en fait l'un des postes les plus représentés dans la base de données d'EASYRECRUE (voir Figure 4.6). Deuxièmement, c'est un poste où les compétences sociales sont recherchées ; compétences pour lesquelles, l'EVD est un outil d'évaluation adapté (section 2.4 et section 2.5). Troisièmement, cette position est aussi largement évoquée dans les états de l'art [Torres and Gregory, 2018, Nguyen et al., 2014] facilitant une comparaison avec l'existant.

Méthodologie pour la sélection des campagnes de recrutement

Comme évoqué en section 4.1.1, les titres des postes pour les campagnes de recrutement sont enregistrés à la main par le recruteur. La première étape consiste à sélectionner les campagnes de recrutement correspondant à une position commerciale en fonction de leur titre.

Afin d'effectuer cette sélection des campagnes de recrutement, nous procédons en 3 étapes (voir Figure 4.7).

i) Sélection des positions. Nous nous appuyons sur le référentiel ROME, ressource

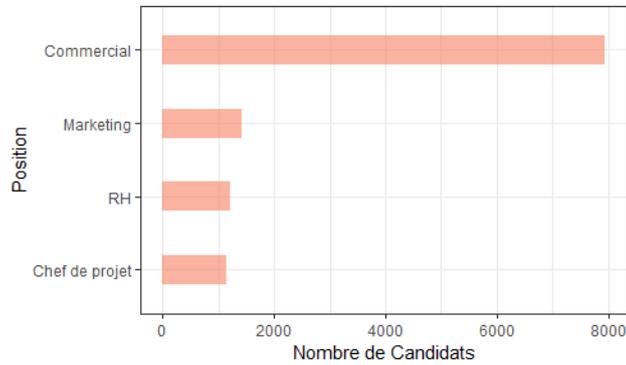


FIGURE 4.6 – Nombre de candidats par type de poste dans la base de données EASY-RECRUE. La position "Commercial" est la plus représentée, ce qui en fait une position privilégiée pour l'analyse automatique.

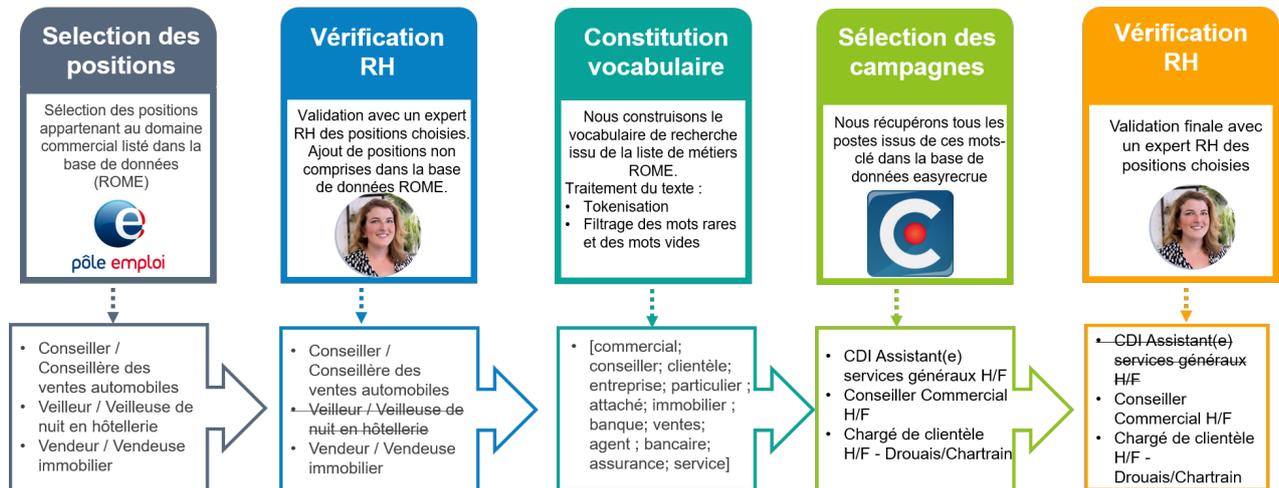


FIGURE 4.7 – Processus de sélection des campagnes pour le jeu de données. La partie inférieure montre des exemples aux différentes étapes de cette constitution.

externe constituant le référentiel français des métiers. Grâce à ce référentiel, nous choisissons à la main une liste de métiers en lien soit avec la relation client, soit avec une compétence commerciale. Cette liste est disponible en appendices A. Une fois cette liste établie, un expert RH vérifie que cette liste est bien en adéquation avec des postes à compétence commerciale ou relation clientèle.

ii) Constitution du vocabulaire de recherche. Une fois cette liste construite, nous construisons le vocabulaire nécessaire pour isoler les campagnes voulues. Dans ce sens nous effectuons les démarches suivantes sur la liste de métiers sélectionnée :

1. nous tokenisons chacun des mots contenus dans la liste des métiers
2. Nous enlevons tous les caractères spéciaux et numériques
3. Nous enlevons les mots dont le nombre de caractères est inférieur à trois (H, F, vie, car, etc)
4. Nous enlevons les mots vides¹ (de, le , à, etc.)

Cette liste de mots constitue par la suite le vocabulaire nécessaire afin de trouver les postes recherchés. À noter que nous gardons un bon nombre de mots que l'on peut juger non utile (tel que veilleur, tournante, fluviale, etc.). Ceci est un choix, nous avons préféré conserver un vocabulaire assez large afin d'avoir un rappel assez grand sur les postes à sélectionner.

iii) Sélection des campagnes. Une fois ce vocabulaire constitué, nous n'avons plus qu'à extraire tous les postes contenant au moins un mot dans ce vocabulaire. Nous vérifions ensuite à la main si le poste correspond bien à l'un des métiers initiaux. Finalement, notre expert RH vérifie et valide l'ensemble des positions finales dans le jeu de données.

Un nuage de mots de l'ensemble des positions finales est disponible en Figure 4.8

4.2.2 Méthodologie d'attribution des étiquettes

Pour rappel, en regardant les vidéos des candidats, les recruteurs peuvent aimer, ne pas aimer, shortlister des candidats, les évaluer sur des critères prédéfinis ou écrire des commentaires (voir section 4.1.3 p. 48). Une fois les campagnes relatives aux positions commerciales extraites, nous effectuons une stratégie d'attribution des étiquettes pour chacun des entretiens. Tout d'abord, nous excluons de tous les jeux de données, les candidats n'ayant pas encore été vus par au moins un recruteur.

Nous choisissons de créer une tâche de classification binaire. Nous séparons les candidats en deux classes nommément "Convocable" et "Non Convocable". Ces étiquettes traduisent

1. stopwords en anglais

faiblesses quant à la définition des classes négatives. Plusieurs cas problématiques existent notamment :

- Il se peut que des candidats pour lesquels le recruteur ne s’est pas exprimé appartiennent en effet à la classe positive.
- Les évaluations par critères n’étaient pas prises en compte pour l’attribution des étiquettes, cette méthode d’étiquetage implique pour notre jeu de données un nombre important de candidats sans annotations.
- Le dossier "shortlist" est parfois mal utilisé par les recruteurs (certains recruteurs s’en servent comme dossier "candidats traités").

Pour combler ces lacunes, nous proposons lors de la constitution du deuxième jeu de données d’effectuer une modification de la stratégie d’étiquetage.

Stratégie d’étiquetage pour le JDD2

Pour chaque recruteur qui a au moins émis une opinion (pouce haut ou bas), évalué un candidat sur un critère ou déplacé le candidat dans le dossier shortlist, le label de la paire $\{candidat; recruteur\}$ est défini par la procédure visible sur la figure 4.9. Nous avons choisi d’étiqueter à nouveau les candidats pour une tâche de classification binaire et non une tâche de régression. Ce choix est motivé par le fait que les opinions demeurent la modalité d’annotations la plus utilisée par les recruteurs, et qu’il était préférable de conserver une tâche de classification pour des raisons de comparaisons.

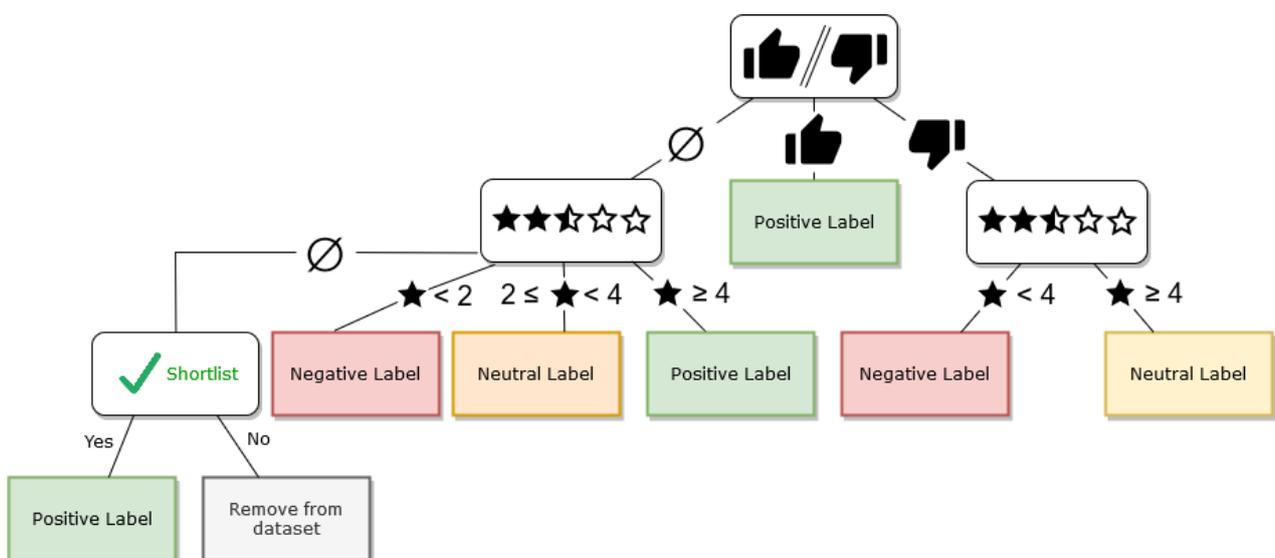


FIGURE 4.9 – Stratégie d’étiquetage pour le jeu de données 2.

Plusieurs critères étant disponibles lors de l'évaluation du candidat, nous considérons la note moyenne des scores d'évaluation de tous les critères comme une note représentative de la convocabilité du candidat. La procédure d'annotation proposée a été choisie à la suite de discussions avec des utilisateurs et des observations de leur expérience utilisateur sur la plate-forme. Par exemple, le bouton "shortlist" est souvent utilisé pour poursuivre le processus de sélection même si le candidat n'est pas considéré comme un bon candidat (en particulier pour les postes en pénurie). Ainsi, il est naturel que nous n'utilisions cette annotation que lorsque d'autres manquent. Nous avons ainsi défini un ordre dans le choix des annotations : nous considérons l'opinion (pouce haut ou bas) comme la plus valide, s'en suit la moyenne des évaluations et le dossier shortlist. Si plusieurs annotateurs ont annoté le même candidat (c'est-à-dire plusieurs paires de $\{candidat; recruteur\}$ pour le même *candidat*), nous effectuons une agrégation d'étiquettes avec un vote majoritaire.

En cas d'égalité, nous procédons à l'étiquetage suivant :

- s'il y a autant d'étiquettes négatives que d'étiquettes neutre, le candidat est considéré comme négatif
- s'il y a autant d'étiquettes positives que d'étiquettes négatives, le candidat est considéré comme neutre
- s'il y a autant d'étiquettes positives, négatives et neutres, le candidat est considéré comme neutre
- s'il y a autant d'étiquettes positives que d'étiquettes neutres, le candidat est considéré comme positif

Finalement nous considérons les candidats associés à l'étiquette neutre comme non convocables. À noter qu'il serait possible de constituer trois classes comme il est courant de le faire en classification d'opinions. Nous avons préféré simplifier à une tâche de classification binaire pour comparer plus facilement les résultats obtenus entre nos jeux de données et avec l'état de l'art.

4.2.3 Sélection des réponses

Le recruteur a la possibilité de poser plusieurs questions dont le type de réponses diffère (voir section 4.1.1 p. 46). Dans cette thèse, nous nous intéressons principalement à l'influence des comportements verbaux et non verbaux en entretien vidéo différé. Nous excluons donc de notre analyse les questions autres que celles à réponses vidéos. De ce fait, il est possible qu'une partie des informations soit manquante afin d'inférer correctement la convocabilité d'un candidat. Ainsi, il nous sera impossible de prédire la convocabilité si par

exemple le critère d'orthographe est une dimension importante dans certaines campagnes de recrutement. Néanmoins, les questions à réponses vidéos demeurent le type de questions le plus utilisé au sein de la plateforme.

4.3 Extraction de descripteurs sociaux multimodaux

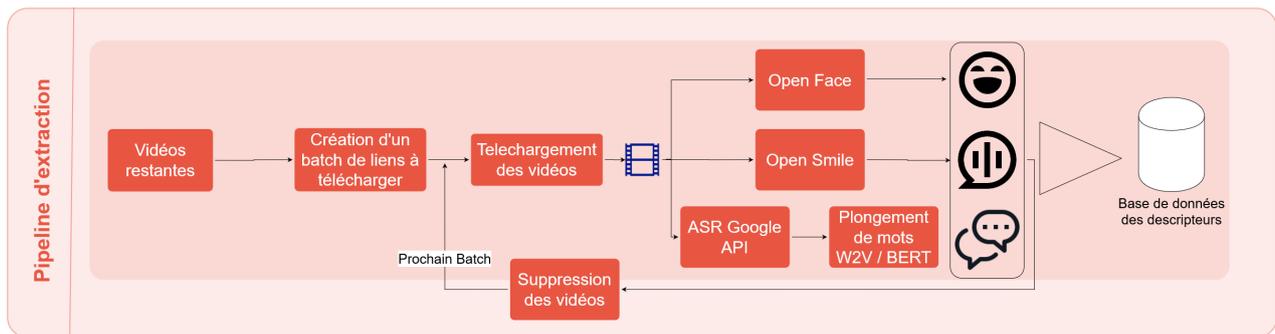


FIGURE 4.10 – Pipeline d'extraction des descripteurs

Nous nous intéressons maintenant à l'extraction des descripteurs issus des vidéos réponses des candidats. La chaîne de traitement complète est décrite par la figure 4.10

En raison de contraintes légales, il ne nous est pas autorisé de conserver les vidéos. Nous procédons à une extraction de descripteurs et une anonymisation des données recueillies. Ceci se traduit par une chaîne de traitement spécifique. Nous nous intéressons qu'aux seules réponses vidéos sélectionnées à la suite de la recherche de campagne, nous téléchargeons la vidéo si celle-ci est encore disponible dans la base de données, puis nous procédons à l'extraction des descripteurs issus de la vidéo et de l'audio de la vidéo. Nous faisons ensuite appel à un service de reconnaissance automatique de la parole afin d'obtenir la retranscription automatique de la réponse vidéo. Ensuite, pour chaque modalité, nous extrayons des descripteurs par trame. Nous obtenons donc des séquences de descripteurs bas niveaux. Une fois l'extraction effectuée, nous supprimons la vidéo et ne conservons uniquement que les descripteurs. Nous reprenons ci-dessous, par modalité, la liste des descripteurs que nous avons choisi d'extraire. À noter que dans un but de reproductibilité, tous les descripteurs sont extraits à l'aide de boîtes outils libres, accessibles, et largement utilisées par la communauté de l'informatique affective.

4.3.1 Extraction de descripteurs issus de la vidéo

Trois groupes de descripteurs ont été initialement extraits de la vidéo : les descripteurs issus des expressions faciales, les descripteurs de la quantité de mouvement et les descrip-

teurs liés à la détection des mouvements de main. À la suite de plusieurs tests, seuls les descripteurs issus des expressions faciales ont été retenus par la suite. Nous expliquons tout de même le raisonnement qui nous a poussé à éliminer les deux derniers groupes de descripteurs.

Expressions faciales

Nous avons décidé d'extraire les caractéristiques visuelles des expressions faciales avec OpenFace 2.1² [Baltrusaitis et al., 2018], un logiciel d'analyse comportementale visuelle de pointe qui fournit diverses mesures significatives par trame. Nous avons choisi d'extraire, pour chaque image, la position et la rotation de la tête, l'intensité et la présence des unités d'action, et la direction du regard, ce qui donne un vecteur de dimension 52.

Ce programme informatique est un outil libre d'accès pour la recherche académique et dont le temps de traitement était acceptable (deux semaines de traitement pour 40 000 vidéos sur un cluster de 32 coeurs). Ce choix d'outil permettra à la communauté d'utiliser et de comparer notre méthode avec l'une des leurs.

Aussi, l'utilisation d'un tel outil a pour avantage de s'extraire de l'apparence physique qui peut avoir une influence considérable en entretien d'embauche (voir section 2.7). Cependant, un tel outil souffre parfois des conditions dans lesquelles les vidéos ont été enregistrées (mauvaise illumination, visage trop rapproché ou coupé, etc.). En ce sens, l'outil OpenFace présente des problèmes d'extraction de descripteurs pour certaines vidéos. Nous décidons de retirer du jeu de données les vidéos pour lesquelles OpenFace n'a pas réussi à détecter un visage dans plus de 20 % de la vidéo. Enfin, l'outil n'est pas parfait et certaines unités d'action faciales ne sont parfois pas bien détectées. Par exemple, [Baltrusaitis et al., 2018] reporte une corrélation de Pearson de 0.22 pour la détection de l'AU 20 de la tension des lèvres contre une corrélation de 0.85 pour la détection de l'AU 12 l'étirement du coin des lèvres. La richesse de la représentation est donc dépendante du succès de l'extraction de l'outil.

Nous avons choisi de ne pas utiliser de réseaux convolutionnels pré-entraînés pour obtenir une représentation du visage trame par trame comme effectué dans [Escalante et al., 2020]. En effet, une telle représentation pourrait encoder des caractéristiques individuelles du candidat (âge, genre, ethnique). Par la suite, cet encodage pourrait renforcer les biais possiblement présents dans le jeu de données vis-à-vis de caractéristiques statiques non désirables.

Quantité de mouvement

Bien qu'il a été montré que la quantité de mouvement était utile pour l'inférence de

2. <https://github.com/TadasBaltrusaitis/OpenFace>

l'employabilité ou de traits de personnalités comme l'extraversion [Nguyen et al., 2014], nous avons fait le choix de ne pas nous intéresser à la quantité de mouvement de la vidéo. Plusieurs méthodes ont été envisagées notamment détecter le mouvement total d'énergie par soustraction d'arrière plan [Nguyen and Gatica-Perez, 2016], ou par segmentation du buste à l'aide de réseaux d'apprentissage profond [Cao et al., 2019]. Cependant, il était parfois difficile de calculer des descripteurs fiables vis-à-vis des différentes configurations d'enregistrement. Plus particulièrement pour les vidéos réponses enregistrées par smartphone ou tablette, le fait que l'appareil enregistreur ne soit pas fixe engendrait de nombreuses valeurs incohérentes.

Les mouvements de mains

Bien que l'information puisse être très informative [Feiler and Powell, 2015] (self touching, modalité supplémentaire, etc), nous avons décidé de ne pas prendre en compte les mouvements de main. Après la mise en place d'une détection de mains par réseaux de neurones pré entraînés [Dibia, 2017] sur le jeu de données EgoHands [Bambach et al., 2015], nous nous sommes rendu compte qu'une faible partie des vidéos contient effectivement des mains (moins de 5 % du total des vidéos). De ce fait, nous avons préféré retirer cet ensemble de descripteurs.

4.3.2 Extraction de descripteurs issus de la voix

Nos descripteurs audio issus d'une trame de signal sonore sont extraits à l'aide d'OpenSmile [Eyben et al., 2013b]. La configuration que nous utilisons est la même que celle utilisée pour obtenir les descripteurs eGeMAPS [Eyben et al., 2016]. GeMAPS est un célèbre ensemble minimaliste de descripteurs sélectionnés pour leur pertinence en informatique sociale et eGeMAPS en est la version étendue. Nous extrayons les descripteurs avant les opérateurs d'agrégations afin d'obtenir un jeu de descripteurs bas niveaux pour chacune des trames. Nous extrayons un vecteur de dimensions 23 à chaque pas de temps de 0.01s. Comme OpenFace, cette boîte à outil est souvent utilisée en informatique affective, ce qui assure une plus grande reproductibilité de la méthode. Enfin, cet outil d'extraction avait aussi un net avantage dans le temps traitement par rapport à d'autres outils d'extraction similaires (ex. Covarep).

4.3.3 Extraction de descripteurs issus du contenu verbal

Comme dans les études précédentes sur les entretiens [Rasipuram and Jayagopi, 2018, Muralidhar et al., 2018], nous utilisons la reconnaissance vocale automatique³ pour obtenir la transcription du contenu verbal et l’horodatage de chaque mot transcrit. À noter que lors de la constitution de JDD1, l’API de Google ne renvoyait pas l’horodatage de chaque mot transcrit. L’alignement du contenu verbal et des modalités n’étant pas possible, la multimodalité ne pouvait être traitée efficacement sans la constitution d’un second jeu de données. Nous nous assurons de la qualité de la retranscription automatique en enlevant du jeu de données les retranscriptions ayant un taux de confiance moyen inférieur à 85%.

Pour JDD1 et JDD2, nous avons utilisé les modèles états de l’art pour l’acquisition de représentation textuelle. Les plongements de mots pré-entraînés sont souvent utilisés comme unité de base en traitement automatique du langage naturel, nous avons donc fait le choix de cette représentation. Étant donné que l’état de l’art en termes de représentation textuelle évolue très rapidement, et avait donc évolué entre JDD1 et JDD2, les représentations utilisées pour les deux jeux de données diffèrent. Ainsi pour JDD1, le contenu verbal de la transcription est transformé en une séquence de vecteurs de plongements lexicaux (word2vec). Nous avons utilisé des mots représentés par un vecteur de dimension 200 [Fauconnier, 2015] préalablement entraînés sur un corpus français de Wikipédia. Pour JDD2, nous obtenons une représentation de dimension 768 pour chaque mot de la transcription verbale en utilisant « CamemBERT » un modèle RoBERTa pré entraîné en langue française. Le modèle utilisé est issu de huggingFace⁴ et la représentation du mot est ici un plongement de mot contextualisé.

4.4 Résumé des jeux de données utilisés dans cette thèse

Nous reportons dans cette partie l’ensemble des jeux de données utilisés au cours de cette thèse. Un tableau récapitulatif des deux jeux de données collectés au travers de la plateforme EASYRECRUE est disponible 4.1. Pour rappel, il a été nécessaire de constituer un second jeu de données en raison de l’expiration de la date de conservation des vidéos dans la première base de données. En effet, en l’absence de ces vidéos il nous était impossible d’extraire de nouveaux descripteurs états de l’art (voir pour le contenu verbal W2V vs BERT) et d’obtenir l’horodatage des retranscriptions automatiques nécessaires à la fusion bas niveau pour un modèle multimodal. Des données complémentaires aux jeux de données

3. Google speech-to-text API.

4. https://huggingface.co/transformers/model_doc/camembert.html

et aux découpages spécifiques pour les expériences sont disponibles dans les sections jeux de données des chapitres respectifs chapitre 5 et chapitre 6.

En plus de ces deux jeux de données, nous avons utilisé un dernier jeu de données publique nommé ChaLearn [Escalante et al., 2020] (voir tableau 3.1 p. 26). Ce jeu de données a été utilisé pour proposer une méthodologie adversaire afin de contrôler les possibles discriminations du modèle automatique. Dans un tel cadre d'étude, il est nécessaire de pouvoir accéder à des variables sensibles (genre, ethnie, âge, etc.) afin d'évaluer le modèle. Récolter de telles variables dans notre cas d'application s'avère interdit du point de vue de la législation française, nous avons donc choisi d'utiliser cette base de données publique déjà étiquetée en genre et en ethnie. Des données complémentaires sont aussi disponibles dans la section description du jeu de données des chapitre 5, 6 et 8.

Jeu de données	JDD 1	JDD 2	ChaLearn [Escalante et al., 2020]
Disponibilité	01/06/18 - 01/06/19	01/09/19 - 01/09/20	Public
Contexte du jeu de données	EVD	EVD	Vlogs
Postes considérés	Commerciaux	Commerciaux	Aucun
Annotateurs	Décisionnaires	Décisionnaires	AMT
Nombres de candidats	35830 réponses vidéo de 7095 candidats	28056 réponses vidéo de 5148 candidats	10 000 clips issus de 3060 vidéos
Dimensions annotées	Convocabilité	Convocabilité	Convocabilité
Étiquettes dérivées par	Opinions (pouces hauts) et dossier de classification (shortlist)	Opinions (pouces hauts ou bas), Évaluations par critères et dossier de classification (shortlist)	-
Contenu verbal obtenu par	ASR (sans horodatage des mots)	ASR (avec horodatage des mots)	Transcription manuelle
Descripteurs audio	eGeMAPS	eGeMAPS	ComParE
Descripteurs vidéo	OpenFace	OpenFace	OpenFace
Descripteurs textuels	W2V	BERT	BERT
Représentation	HireNet V1	HireNet V2	GRU
Fusion	-	Bas Niveau par GMU	Haut Niveau par GMU
Interprétabilité	Attention locale	Attention locale et analyse étendue des pics d'attention	-
Équitabilité de l'AA	-	-	Méthodes adversaires
Utilisation Chapitre	chapitre 5	chapitre 6 et 7	chapitre 8

TABLE 4.1 – Tableau récapitulatif des jeux de données utilisés au cours de cette thèse.

Deuxième partie

Architecture neuronale pour l'analyse automatique d'entretiens d'embauche asynchrones

Section 5

HireNet : Un modèle hiérarchique pour l'analyse automatique d'entretiens vidéo différés

Nous proposons un nouveau modèle neuronal hiérarchique avec attention appelé HireNet, qui vise à prédire la convocabilité des candidats telle qu'évaluée par les recruteurs. Dans HireNet, un entretien est considéré comme une séquence de questions-réponses contenant elles-mêmes une séquence de signaux sociaux. Dans HireNet, deux sources d'information contextuelles sont modélisées : les mots contenus dans la question et dans le titre du poste. Notre modèle obtient de meilleurs résultats en F1 que les approches précédentes pour chacune des modalités (contenu verbal, audio et vidéo). Les résultats de la fusion multimodale précoce et tardive suggèrent que des schémas de fusion plus sophistiqués sont nécessaires pour améliorer les résultats monomodaux. Enfin, quelques exemples de moments capturés par les mécanismes d'attention suggèrent que notre modèle pourrait potentiellement être utilisé pour aider à trouver les moments clés d'un entretien d'embauche asynchrone.

Publication associée à ce chapitre : [Hemamou et al., 2019b] Hemamou, L. et al. 2019. HireNet : A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. Proceedings of the AAAI Conference on Artificial Intelligence. 33, (juill. 2019), 573-581.

5.1 HireNet et hypothèses sous-jacentes

Nous proposons ici un nouveau modèle neuronal nommé HireNet pour la tâche de prédiction de la convocabilité. Il s'inspire des travaux menés dans les réseaux de neurones

pour le traitement du langage naturel et plus particulièrement du HierNet [Yang et al., 2016], qui vise à modéliser une hiérarchie dans un document. Suivant l'idée qu'un document est composé de phrases et de mots, un entretien d'embauche peut être décomposé en une séquence de questions-réponses et une séquence de descripteurs de bas niveau décrivant chaque réponse. L'architecture du modèle (voir la figure 5.1) repose sur quatre hypothèses. La première hypothèse (**H1**) est l'importance de la séquentialité des comportements intervenant dans l'entretien. Ainsi nous supposons qu'un modèle prenant en compte la dynamique des comportements pourra modéliser plus finement la réponse des candidats. Nous avons donc choisi d'utiliser un modèle séquentiel tel qu'un réseau de neurones récurrent. La deuxième hypothèse (**H2**) concerne l'importance de la structure hiérarchique d'un entretien : l'annotation de la convocabilité doit être prise à la suite d'un entretien complet composé de plusieurs questions-réponses et non pas à la suite de chaque question. En effet, il est fort probable qu'un candidat ait réussi son entretien tout en se trompant à une ou deux questions. Nous avons donc choisi d'introduire différents niveaux de hiérarchie dans HireNet, à savoir le niveau de l'entretien, le niveau des questions-réponses et le niveau des mots (ou des trames audio ou visuelles). La troisième hypothèse (**H3**) concerne l'existence d'informations ou de signaux sociaux saillants dans l'entretien vidéo du candidat : toutes les questions n'ont pas la même importance et tous les moments des réponses n'ont pas une même influence sur la décision du recruteur. Nous avons donc choisi d'introduire des mécanismes d'attention dans HireNet. La dernière hypothèse (**H4**) concerne l'importance des informations contextuelles telles que l'intitulé des questions et des postes. Ces informations contextuelles sont nécessaires à la fois pour juger l'adéquation de la réponse aux questions posées et peuvent largement influencer les stratégies des candidats [Peeters and Lievens, 2006]. Par conséquent, HireNet inclut des vecteurs qui codent cette information contextuelle.

5.2 Formalisation

Nous représentons un entretien vidéo comme un objet composé d'un intitulé de poste J et n paires de questions-réponses $\{\{Q_1, A_1\}, \{Q_2, A_2\}, \dots, \{Q_n, A_n\}\}$. Dans notre modélisation, l'intitulé du poste J est composé d'une séquence de l_J mots $\{w_1^J, w_2^J, \dots, w_{l_J}^J\}$ où l_J dénote la longueur de l'intitulé du poste. De la même manière, la i -ème question Q_i est une séquence de l_{Q_i} mots $\{w_1^i, w_2^i, \dots, w_{l_{Q_i}}^i\}$ où l_{Q_i} dénote le nombre de mots de la question i . A_i dénote la séquence de descripteurs de bas niveau $\{x_1^i, x_2^i, \dots, x_{l_{A_i}}^i\}$ composant la i -ème réponse. Dans cette étude, ces descripteurs peuvent être des "plongements de mots" (ou

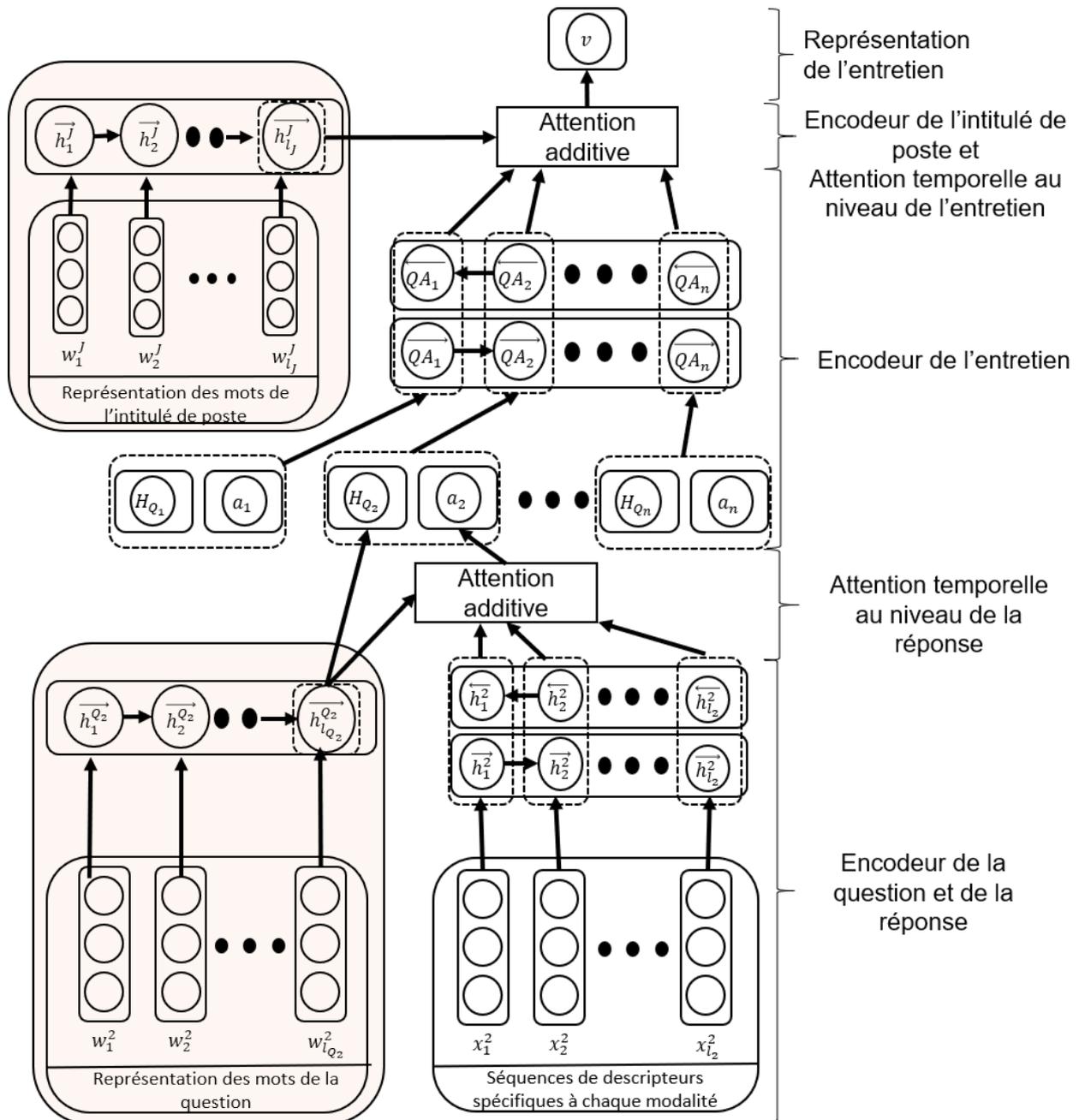


FIGURE 5.1 – HireNet. Les blocs en couleur correspondent aux encodeurs de contexte.

word embedding en anglais), des descripteurs extraits d'une trame du signal sonore, or des descripteurs extraits d'une trame de vidéo. l_{A_i} dénote la longueur de la séquence des descripteurs de bas niveau de la i -ème réponse.

5.2.1 Gated Recurrent Unit Encoder

Nous avons décidé d'utiliser les Gated Recurrent Unit (GRU) [Cho et al., 2014] pour encoder les informations du titre du poste, des questions et des réponses. Un GRU est capable d'encoder des séquences. Il utilise deux mécanismes pour résoudre le problème de la disparition du gradient, à savoir la porte de réinitialisation, contrôlant la quantité d'informations passées nécessaires; et la porte de mise à jour, déterminant la quantité d'informations à conserver du passé et la quantité de nouvelles informations à ajouter. Pour la formalisation, on notera h_t l'état caché de GRU au pas de temps t de la séquence codée.

5.2.2 Encodeur de la réponse

Cette partie du modèle vise à encoder les séquences de descripteurs de bas niveau décrivant une réponse. Comme mentionné précédemment, les séquences peuvent représenter un texte, un flux audio ou un flux vidéo. Un GRU bidirectionnel est utilisé pour obtenir une représentation dans les deux sens pour chaque élément de la séquence X . Elle contient un \overrightarrow{GRU} qui lit la séquence de gauche à droite et un \overleftarrow{GRU} qui lit la séquence de droite à gauche.

$$\overrightarrow{h}_t^i = \overrightarrow{GRU}(x_t^i), t \in [1, l_{A_i}] \quad (5.1)$$

$$\overleftarrow{h}_t^i = \overleftarrow{GRU}(x_t^i), t \in [l_{A_i}, 1] \quad (5.2)$$

Ainsi, un encodage pour un descripteur de bas niveau donné x_t^i est obtenu en concaténant les états cachés du GRU bidirectionnel.

$$h_t^i = [\overrightarrow{h}_t^i, \overleftarrow{h}_t^i] \quad (5.3)$$

L'encodage de séquences de manière bidirectionnelle garantit la même quantité d'informations précédentes pour chacun des éléments de $(A_i)_{1 \leq i \leq n}$. L'utilisation d'un simple encodeur pourrait conduire à des vecteurs d'attention biaisés se concentrant uniquement sur les derniers éléments des réponses.

5.2.3 Encodeur de la question

Dans cette étude, les informations de contexte local correspondent à l'intitulé des questions $(Q_i)_{1 \leq i \leq n}$. Afin d'encoder ces phrases, nous utilisons un simple GRU.

$$\overrightarrow{h}_t^{Q_i} = \overrightarrow{GRU}(w_t^i), t \in [1, l_{Q_i}] \quad (5.4)$$

La représentation finale d'une question est l'état caché du dernier mot de la question.

$$H_{Q_i} = \overrightarrow{h}_{l_{Q_i}}^{Q_i} \quad (5.5)$$

5.2.4 Attention temporelle au niveau de la réponse

Afin d'obtenir une meilleure représentation de la réponse du candidat, nous cherchons à détecter dans la séquence les éléments qui ont joué un rôle déterminant dans la tâche de classification. De plus, nous émettons l'hypothèse que le contexte local est très important. Différents signaux comportementaux ou stratégies d'impression (autopromotion ou autres stratégies ciblées, voir section 2.6) peuvent survenir en fonction du type de question (questions situationnelles ou comportementales) [Levashina et al., 2014] et peuvent influencer sur la façon dont les recruteurs évaluent leurs candidats [Roulin et al., 2015]. Un mécanisme d'attention additif est proposé afin d'extraire l'importance de chaque moment de la séquence représentant la réponse.

$$u_t^i = \tanh(W_A h_t^i + W_Q H_{Q_i} + b_Q) \quad (5.6)$$

$$\alpha_t^i = \frac{\exp(u_p^\top u_t^i)}{\sum_{t'} \exp(u_p^\top u_{t'}^i)} \quad (5.7)$$

$$a_i = \sum_t \alpha_t^i h_t^i \quad (5.8)$$

où W_A et W_Q sont des matrices de poids, u_p et b sont des vecteurs de poids et u_p^\top dénote pour la transposée de u_p .

5.2.5 Encodeur de l'entretien

Afin d'avoir le maximum d'informations, nous concaténons au deuxième niveau, la représentation du contexte local (représentation de la question) et la représentation de la réponse. Cette concaténation permet au réseau de prendre en compte la question posée et donc l'adéquation de la réponse à cette question. De plus, nous pensons que, vu le

fonctionnement des entretiens vidéo, plus le candidat répond à des questions au cours de l'entretien, plus il s'adapte et se met à l'aise. Au vu de ces éléments, nous avons décidé d'encoder les paires de questions-réponses sous forme de séquence. Étant donné $\{[H_{Q_1}, a_1], [H_{Q_2}, a_2], \dots, [H_{Q_n}, a_n]\}$, nous pouvons utiliser le même schéma de représentation que celui de l'encodeur de bas niveau :

$$\overrightarrow{QA}_i = \overrightarrow{GRU}([H_{Q_i}, a_i]), i \in [1, n] \quad (5.9)$$

$$\overleftarrow{QA}_i = \overleftarrow{GRU}([H_{Q_i}, a_i]), i \in [n, 1] \quad (5.10)$$

Nous concaténons aussi les représentations du bidirectionnel GRU.

$$QA_i = [\overrightarrow{QA}_i, \overleftarrow{QA}_i] \quad (5.11)$$

5.2.6 Encodeur de l'intitulé de poste

Nous encodons le titre du poste de la même manière que nous encodons les questions :

$$\overrightarrow{h}_t^J = \overrightarrow{GRU}(w_t^J), t \in [1, l_J] \quad (5.12)$$

Comme pour la représentation de la question, la représentation finale du titre du poste est l'état caché du dernier mot de J (*i.e.* $h_{l_J}^J$).

$$H_J = \overrightarrow{h}_{l_J}^J \quad (5.13)$$

5.2.7 Attention temporelle au niveau de l'entretien

L'importance d'une question dépend du contexte de l'entretien et plus particulièrement du type d'emploi pour lequel le candidat postule. Par exemple, un entretien avec un poste de vente junior pourrait accorder plus d'importance aux compétences sociales, alors qu'un entretien avec un poste de direction pourrait être plus difficile du point de vue technique. Comme l'attention de bas niveau, l'attention de haut niveau est composée d'un mécanisme d'attention additive :

$$u_i = \tanh(W_{QA}QA_i + W_JH_J + b_J) \quad (5.14)$$

$$\alpha_i = \frac{\exp(u_J^\top u_i)}{\sum_{i'} \exp(u_J^\top u_{i'})} \quad (5.15)$$

$$v = \sum_i \alpha_i Q A_i \quad (5.16)$$

où W_{QA} , W_J sont des matrices de poids, u_J et b_J sont des vecteurs de poids et u_J^\top dénote la transposée de u_J . Finalement v résume toutes les informations de l'entretien d'embauche.

5.2.8 Classification du candidat

Une fois v obtenu, nous utilisons sa représentation pour classifier les candidats :

$$\tilde{y} = \sigma(W_v v + b_v) \quad (5.17)$$

où W_v est une matrice de poids et b_v un vecteur de poids.

Comme le problème auquel nous sommes confrontés est celui d'une classification binaire, nous avons choisi de minimiser l'entropie croisée binaire calculée entre \tilde{y} et les vraies étiquettes y des candidats.

5.3 Expérimentations

5.3.1 Jeu de données

Nous utilisons la première base de données constituée (voir section 4.4 p. 59) pour effectuer les différentes expérimentations. Pour rappel, la base de données est spécifique au poste de commercial, obtenue dans des conditions hors laboratoires, et dont les annotations sont de réelles décisions de convocabilité. Étant donné les différentes natures des chaînes de traitement pour l'extraction des descripteurs et leurs possibles problèmes associés (piste sonore inaudible, mauvais éclairage, etc.), nous obtenons un nombre différent de candidats pour chacune des modalités. Nous avons décidé d'utiliser tous les échantillons disponibles dans chaque modalité séparément enfin de juger au mieux de la validité du modèle. Des statistiques sur le jeu de données sont disponibles dans le tableau 5.1.

5.3.2 Protocoles expérimentaux

Les métriques d'évaluation choisies sont la précision, le rappel et le score F1 de la classe *convocable*. Elles sont bien adaptées à la classification binaire et utilisées dans des études précédentes [Chen et al., 2017]. Nous avons divisé l'ensemble de données en un ensemble d'apprentissage, un ensemble de validation pour la sélection d'hyper paramètres basée sur le score F1 et un ensemble de test pour l'évaluation finale de chaque modèle. Chaque ensemble constitue respectivement 80 %, 10 % et 10 % de l'ensemble de données complet.

Modalité	Contenu verbal	Audio	Vidéo
Ensemble d'entraînement	6350	6034	5706
Ensemble de validation	794	754	687
Ensemble de test	794	755	702
Questions par entretien (moyenne)	5.05	5.10	5.01
Longueur totale	3.82 M mots	557.7 h	508.8 h
longueur par questions (moyenne)	95.2 mots	52.19 s	51.54 s
Proportion de l'étiquette <i>Convocable</i>	45.0 %	45.5 %	45.4 %

TABLE 5.1 – Tableau descriptif du jeu de données : nombre de candidats dans chaque ensemble et statistiques globales de l'ensemble du jeu de données.

Model	Texte			Audio			Video		
	<i>Precision</i>	<i>Rappel</i>	F1	<i>Precision</i>	<i>Rappel</i>	F1	<i>Precision</i>	<i>Rappel</i>	F1
Non-sequentiel	0.553	0.285	0.376	0.590	0.463	0.519	0.507	0.519	0.507
Bo*W	0.656	0.403	0.499	0.532	0.402	0.532	0.488	0.447	0.467
Bidirectionnel GRU	0.624	0.510	0.561	0.539	0.596	0.566	0.559	0.500	0.528
HN_AVG	0.502	0.800	0.617	0.538	0.672	0.598	0.507	0.550	0.528
HN_SATT	0.512	0.803	0.625	0.527	0.736	0.614	0.490	0.559	0.522
HireNet	0.539	0.797	0.643	0.576	0.724	0.642	0.562	0.655	0.605

TABLE 5.2 – Résultats pour les modèles monomodaux

5.3.3 Extraction de descripteurs sociaux multimodaux

Pour chaque modalité, nous avons extraits comme descripteurs des séquences de bas niveau.

Word2vec : Des plongements de mots pré entraînés sont utilisés pour le BoTW (Sac de mots textuels, présenté plus loin dans cette section) et les réseaux de neurones. Nous avons utilisé des mots représentés par un vecteur de dimension 200 [Fauconnier, 2015] préalablement entraînés sur un corpus français de Wikipédia.

OpenFace : Nous extrayons les descripteurs visuels image par image avec OpenFace [Baltrusaitis et al., 2018]. Nous avons choisi d'extraire la position et la rotation de la tête,

l'intensité et la présence d'unités d'action faciales et la direction du regard résultant en un vecteur de taille 52. Comme différentes vidéos ont différentes fréquences d'images, nous avons décidé de lisser les valeurs avec une fenêtre temporelle de 0,5s et un chevauchement de 0,25s. La durée de 0.5s est fréquemment utilisée dans la littérature de l'informatique sociale [Varni et al., 2018] et a été estimée dans notre corpus en tant que taille de fenêtre temporelle appropriée en annotant des segments de signaux sociaux d'une quinzaine de vidéos.

eGeMAPS : nos descripteurs audio d'une trame de signal sonore sont extraits à l'aide d'OpenSmile [Eyben et al., 2013b]. La configuration que nous utilisons est la même que celle utilisée pour obtenir les descripteurs eGeMAPS [Eyben et al., 2016]. Nous extrayons les descripteurs trame par trame à une fréquence de 100Hz avant les agrégations effectuées pour obtenir la représentation eGeMAPS résultante en un vecteur de taille 23. De la même manière que les descripteurs visuels, nous lissons les valeurs avec une fenêtre temporelle de 0,5s et un chevauchement de 0,25s.

5.4 Comparaison de modèles

Tout d'abord, nous comparons notre modèle à plusieurs méthodes naïves de référence basées sur le vote :

- i) Vote aléatoire* (cette méthode consiste seulement à attribuer aléatoirement une étiquette lors de la prédiction. Un millier de tirages aléatoires respectant l'équilibre des étiquettes du jeu de données d'entraînement a été effectués. Le score F1 est ensuite moyenné.) ;
- ii) Vote par majorité* (cette méthode consiste simplement à attribuer l'étiquette majoritaire du poste concerné. Puisque notre modèle pourrait simplement apprendre uniquement l'étiquette majoritaire de la position pour lequel le candidat postule, nous avons décidé d'inclure ce modèle pour montrer que HireNet dépasse ces indices).

Deuxièmement, nous comparons notre modèle avec des modèles non séquentiels :

- i) -a Texte non séquentiel* (nous entraînons une représentation Doc2vec [Le and Mikolov, 2014] sur notre corpus, et nous l'utilisons comme représentation de nos entrées textuelles) ;
- i) -b Audio non séquentiel* (nous prenons la représentation audio eGeMAPS telle que décrite dans [Eyben et al., 2016]. Cette représentation est obtenue en appliquant aux descripteurs ci-dessus des fonctions statistiques classiques afin d'aggréger la dimension temporelle. Une dernière raison pour laquelle nous avons choisi les descripteurs eGeMAPS de-

meure dans sa facilité à être réutilisé et pouvoir être comparé à des travaux futurs dans le domaine de l'informatique social ;

i) -c Vidéo non séquentiel (nos descripteurs vidéo de bas niveau issus d'OpenFace incluent des descripteurs binaires et des descripteurs continus. L'aggrégation temporelle grâce aux fonctions moyenne, écart type, minimum, maximum, somme des gradients positifs et somme des gradients négatifs ont été utilisés avec succès pour une classification comportementale du contenu multimédia dans [Ryoo et al., 2015]. Nous avons suivi ce schéma de représentation pour nos descripteurs continus. Pour ce qui est de nos descripteurs discrets, nous avons choisi d'extraire le nombre de segments actifs ; la moyenne et l'écart type de la durée des segments actifs)

*ii) Bag of * words* (nous avons également choisi de comparer notre modèle au sac de mots audio et vidéo de [Chen et al., 2017] : nous utilisons un algorithme K-means sur toutes les trames (représentées par les descripteurs bas niveaux) pour essayer de détecter des groupes similaires de moments. Les classes obtenues représentent ensuite notre dictionnaire qui lie les différents moments de la réponse à des "mots" spécifiques représentés par l'identité du cluster. De cette manière, nous transformons la séquence des descripteurs bas niveaux en document où chacune des trames est convertie en "mot" grâce au dictionnaire construit de façon non supervisée. Puis nous utilisons une représentation "Fréquence du terme - Fréquence inverse de document" (TF-IDF) pour modéliser chaque réponse).

Pour chaque modalité, nous utilisons les représentations non séquentielles mentionnées ci-dessus de manière monomodale comme entrée dans trois algorithmes d'apprentissage classiques (à savoir SVM, régression Ridge et forêt aléatoire) avec une recherche d'hyperparamètres. Le meilleur des trois algorithmes est sélectionné. Comme ces modèles n'ont pas de structure hiérarchique, nous les entraînons pour prédire des étiquettes en fonction des réponses (par opposition à l'étiquetage des candidats effectué par notre modèle hiérarchique). Au moment du test, nous faisons la moyenne de toutes les prédictions produites pour chacune des questions d'un candidat comme valeur de sortie de convocabilité pour le candidat.

Troisièmement, les modèles séquentiels proposés visent à vérifier les quatre hypothèses décrites ci-dessus : *i)*, la comparaison du modèle **bidirectionnel-GRU** avec les approches non séquentielles décrites précédemment visent à valider **H1** sur la contribution de la séquentialité ; *ii)* le modèle **HN_AVG** (Hierarchical Averaged Network) ajoute la hiérarchie dans le modèle afin de vérifier **H2** et **H3** (nous remplaçons le mécanisme d'attention en moyennant les sorties GRU bidirectionnelles non nulles) ; *iii)* le réseau hiérarchique auto attentif (ou *self attention* en anglais) (**HN_SATT**) est une version de HireNet avec une auto attention qui vise à voir l'effet réel des informations de contexte ajoutées (**H4**).

5.5 Modèles multimodaux

Compte tenu des versions textuelles, audio et vidéo de notre HireNet, nous proposons deux modèles de base effectuant une inférence multimodale, à savoir une approche de fusion précoce et une approche de fusion tardive. Pour la fusion précoce, nous concaténons la dernière couche v de chaque modalité en tant que représentation, et procédons en appliquant la même procédure de test que nos modèles non séquentiels. Pour notre approche de fusion tardive, la décision du candidat est prise en effectuant la moyenne des scores \tilde{y} de chacune des modalités.

5.6 Résultats et analyses

Modèle	<i>Precision</i>	<i>Rappel</i>	F1
Vote aléatoire	0.459	0.452	0.456
Vote majoritaire	0.567	0.576	0.571
Fusion précoce	0.587	0.705	0.640
Fusion tardive	0.567	0.748	0.645

TABLE 5.3 – Résultats pour les modèles multimodaux naïfs et les modèles basés sur le vote

Tout d’abord, les tableaux 5.2 et 5.3 montrent que la plupart de nos modèles neuronaux dépassent suffisamment les modèles basés sur les votes aléatoires et majoritaires.

Dans la table 5.2, le score F1 est plus élevé, par rapport aux modèles non séquentiels, en utilisant des Bidirectionnel-GRU pour toutes les modalités, ce qui soutient l’hypothèse **H1**. Nous pouvons également constater que HN_AVG est supérieur aux modèles bidirectionnels-GRU pour les modalités audio et textuelles validant **H2** pour ces deux modalités. Ceci suggère que la séquentialité et la hiérarchie sont des biais inductifs adéquats pour un algorithme d’apprentissage automatique d’évaluation d’entretien d’embauche. En ce qui concerne **H3**, HN_SATT a affiché de meilleurs résultats que HN_AVG, pour le texte et l’audio. Au final, notre modèle HireNet dépasse HN_AVG et HN_SATT pour chaque modalité. Par conséquent, un nombre important d’informations utiles est présent dans le cadre contextuel d’une interview et peut être exploité à l’aide de notre modèle, comme indiqué dans **H4**. Les modèles monomodaux audio et texte affichent de meilleures performances que les modèles vidéo. Les mêmes résultats ont été obtenus dans [Chen et al., 2017].

Nos tentatives de fusion des informations multimodales synthétisées dans la dernière couche

de chaque modèle HireNet n'ont apporté qu'un gain de performances minimales par rapport aux modèles à modalité unique.

5.7 Visualisation de l'attention

Les mécanismes d'attention nous fournissent une méthode d'interprétabilité locale (exemple par exemple) [Yang et al., 2019b]. Cette méthode est complémentaire aux méthodes d'interprétations plus globales employées dans les précédentes études [Rasipuram and Jayagopi, 2018, Nguyen and Gatica-Perez, 2016, Naim et al., 2018]. Ainsi, nous pouvons explorer qualitativement et visualiser l'importance des différents moments et des questions-réponses pour chacune des modalités.

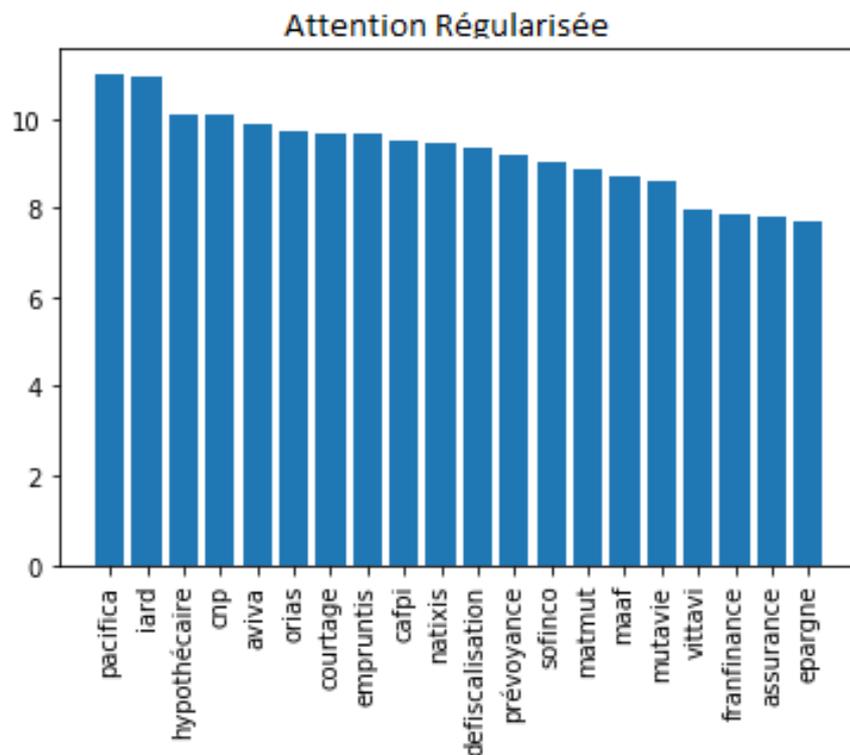


FIGURE 5.2 – Valeur d'attention régularisée des 20 mots identifiés comme les plus importants. Ces résultats suggèrent, pour la modalité du langage, une importance pour les termes techniques plus que pour les termes non techniques.

Texte Afin de visualiser les différents mots sur lesquels les valeurs d'attention étaient élevées, nous avons calculé les nouvelles valeurs d'intérêt comme cela a été fait dans [Yu et al., 2017]. Comme la longueur de la phrase varie possiblement entre les réponses, nous multiplions la valeur de l'attention de chaque mot (α_i^j) par le nombre de mots de la réponse,

résultant en une attention relative du mot par rapport à la phrase. De la même manière, nous multiplions l'attention de chaque question par le nombre de questions, ce qui entraîne une attention relative de la question par rapport à l'entretien d'embauche. Ensuite, d'une manière similaire à [Yang et al., 2016], nous calculons $\sqrt{p_q p_w}$ où p_w et p_q sont respectivement les valeurs d'intérêt pour le mot w et la question q . La liste des 20 mots les plus importants contient de nombreux noms de banques et d'assurances (Natixis, Aviva, CNP, etc.) et du vocabulaire relatif aux connaissances professionnelles (hypothèque, courtage, exonération fiscale, etc.), ce qui signifie que leur apparition joue un rôle important dans la prédiction de la convocabilité.

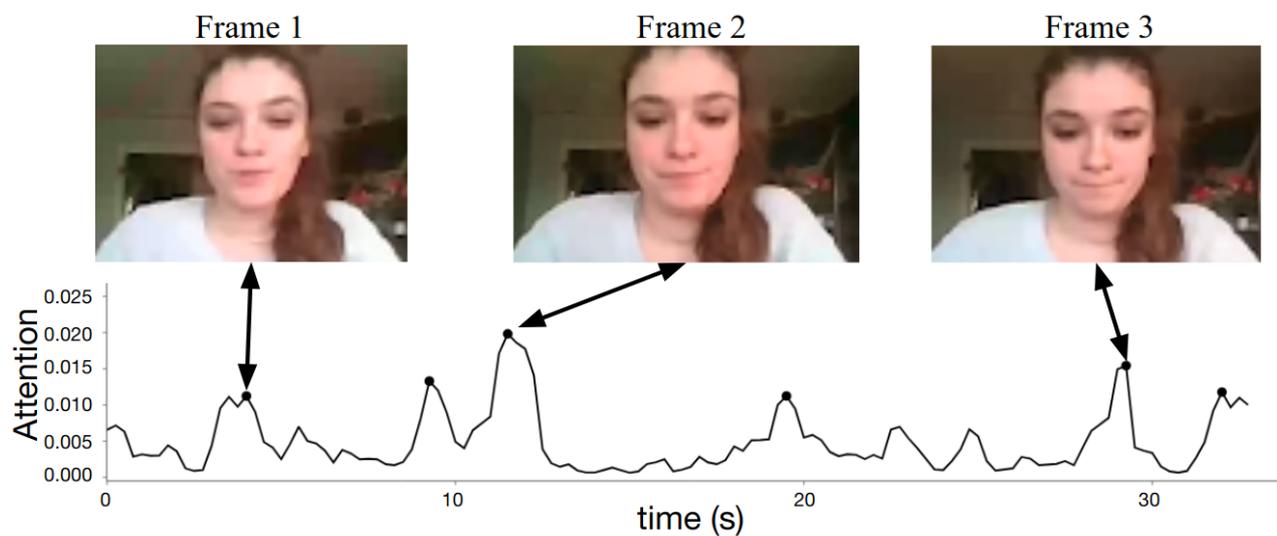


FIGURE 5.3 – Exemple de moments saillants détectés grâce aux pics d'attentions pour la modalité vidéo

Vidéo Pour visualiser les moments mis en évidence par les mécanismes d'attention dans une vidéo, nous présentons un exemple des valeurs d'attention pour une réponse dans la figure 5.3. Dans cette figure, plus la valeur d'attention est élevée, plus les moments correspondants sont considérés comme pertinents pour la tâche de classification par le mécanisme d'attention. Comme on peut le constater, certains pics sont présents. Trois moments avec une forte valeur d'attention sont présentés. Certains signaux sociaux importants lors d'un entretien d'emploi sont identifiés. Nous émettons l'hypothèse que le sourire détecté dans l'image 1 pourrait faire partie d'une stratégie d'impressions [Schneider et al., 2015]. De plus, les images 2 et 3 sont représentatives des signaux de stress du candidat. En fait, il a été suggéré que la succion des lèvres était liée à l'anxiété dans [Feiler and Powell, 2016].

Audio La même procédure de visualisation que celle utilisée pour la vidéo a été étudiée pour l'audio. Le signal audio étant plus difficile à visualiser, nous avons décidé de décrire

le schéma général des pondérations d'attention audio. Dans la plupart des cas, lorsque la prosodie est homogène dans la réponse, les pondérations de l'attention sont uniformément réparties et ne présentent aucun pic, contrairement à ce qui a été observé pour la vidéo. Cependant, des moments marquants peuvent apparaître, en particulier lorsque les candidats produisent des disfluences successives. Ainsi, nous avons identifié que les pics d'attention sont indicatifs des moments où se produisent de faux départs, de mots remplisseurs ou de répétitions.

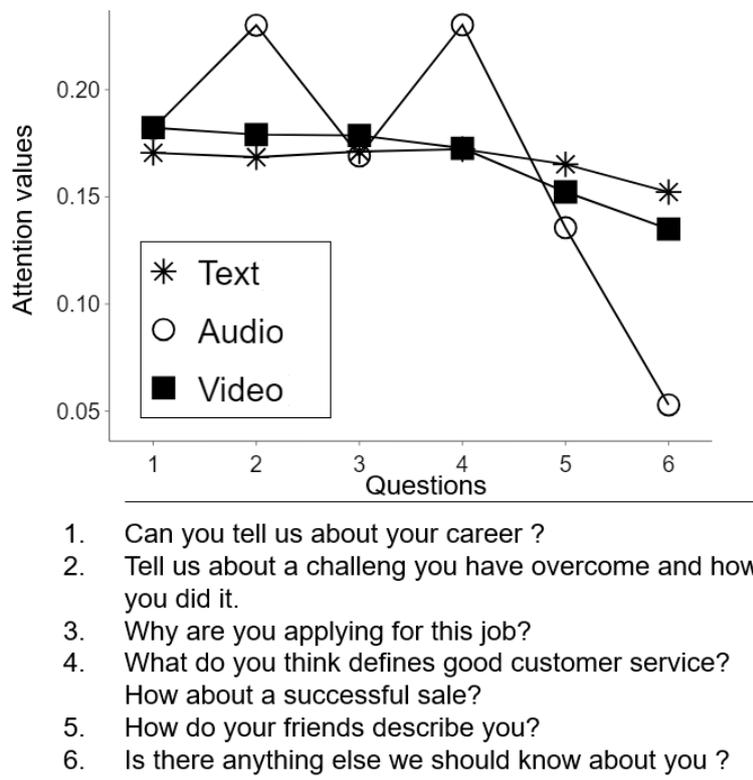


FIGURE 5.4 – Questions de l'entretien d'embauche pour la position aléatoire choisie et leurs valeurs d'attentions respectives selon les différentes modalités.

Questions Notre objectif est d'explorer l'attention portée aux différentes questions lors d'un même entretien. À cette fin, nous avons choisi au hasard un poste ouvert dans le jeu de données de test comprenant 40 candidats. Les questions décrivant l'entretien et les pondérations d'attention moyennes correspondantes sont affichées dans la Figure 5.4. Premièrement, il semble que la variabilité du poids de l'attention entre les questions soit plus élevée pour la modalité audio que pour les modalités texte et vidéo. Deuxièmement, la diminution de l'attention accordée aux questions 5 et 6 (*Comment vos amis vous décrivent ?* et *Y a-t-il quelque chose d'autre que nous devrions savoir à propos de vous ?*) pourrait s'expliquer par le fait que ces questions soient moins importantes pour

évaluer les connaissances et compétences clés du candidat.

Troisièmement, le fait que des pics d'attention pour la modalité audio occurrent lors des questions 2 et 4 peut être provoqué par le fait que ces questions soient conçues pour évaluer des compétences et connaissances spécifiques à la position. En effet, il est possible que les disfluences tendent à apparaître davantage dans les questions de connaissances ou situationnelles.

5.8 Conclusion

Les acteurs du secteur des ressources humaines proposent aujourd'hui des outils permettant d'évaluer automatiquement les candidats subissant des entretiens vidéo asynchrones. Cependant, aucune étude n'a été publiée concernant ces outils et leur validité prédictive. La contribution de ce travail est double. Premièrement, nous évaluons la validité des approches précédentes dans des conditions réelles (vidéos "in-the-wild", réelles candidatures, réelles évaluations, etc.). Deuxièmement, nous avons utilisé des méthodes d'apprentissage profond afin de modéliser fidèlement la structure des entretiens vidéo asynchrones. En ce sens, nous avons proposé une nouvelle version du "Hierarchical Attention Networks", qui prend en compte les éléments contextuels des entretiens (questions et intitulé du poste) dénommés HireNet, qui a donné de meilleurs résultats que les approches précédentes. Les premières expériences de base sur la fusion multimodale ont également été réalisées (fusion précoce et tardive).

Nous avons exploré qualitativement l'apport des mécanismes d'attention comme outil d'interprétabilité locale. Cette interprétabilité demeure très importante pour que les recruteurs puissent faire confiance à une évaluation automatique. De plus, elle est complémentaire aux approches plus globales initialement utilisées dans les travaux états de l'art section 3.5.1 dans le sens où elles fournissent une interprétabilité locale par rapport à des événements ponctuels de l'entretien d'embauche. Néanmoins, une étude supplémentaire est nécessaire pour confirmer la pertinence des moments sélectionnés que nous effectuerons en chapitre 7.

Enfin, il aurait été intéressant d'étudier des différences interindividuelles entre candidats. Ainsi, il aurait été intéressant de comprendre si certains candidats se démarquent dans une modalité particulière plutôt qu'une autre.

Section 6

Attention et Multimodalité

Dans ce chapitre, nous nous focalisons sur deux aspects importants de l'entretien d'embauche nommément la prise en compte du contexte grâce à des mécanismes d'attention (de la section 6.1 à la section 6.3) et la prise en compte de la multimodalité au sein des réponses des candidats (de la section 6.4 à la section 6.6).

De la même façon que les recruteurs évaluent la pertinence de la réponse au regard des questions posées, un système automatique se doit de prendre en compte au mieux le contenu de l'entretien et son interaction avec l'entretien des candidats. Une première approche a été proposée dans le chapitre précédent avec l'utilisation d'une fonction d'attention temporelle prenant en compte le contexte (voir section 7.3). Cependant, nous montrons dans ce chapitre que cette fonction d'attention est insuffisante pour modéliser l'interaction entre le contexte et les séquences de réponse. Or il paraît logique, afin d'évaluer une réponse à une question, de comprendre les éléments clés au regard de la question posée. De même, il fait sens de comprendre quelles sont les questions-réponses clés au regard du poste auquel le candidat postule. Pour pallier cette limite, nous proposons une nouvelle fonction d'attention contextuelle permettant d'encoder l'importance intrinsèque de chaque pas de temps, et l'importance contextuelle de chaque pas de temps au regard de l'intitulé de la question ou du poste.

En première partie de ce chapitre, nous proposons l'ajout d'une fonction temporelle contextuelle au sein de HireNet. Dans une moindre mesure, nous remplaçons la façon dont l'intitulé des questions et du poste est encodé par une méthode état de l'art. Nous évaluons ces modifications sur un nouveau jeu de données. Nous montrons que la mise à jour dont le contexte est encodé améliore les performances du système pour les modalités audio et langage. De plus, nous montrons que l'attention temporelle contextuelle améliore les performances pour le modèle utilisant le langage.

La prise en compte de la multimodalité est importante d'autant plus que l'entretien

d'embauche est un terrain propice aux comportements où la multimodalité joue un rôle important tels que l'emphase ou la tromperie [Schneider et al., 2015, Buehl et al., 2019]. Néanmoins, les mécanismes de fusion n'ont été que très peu étudiés dans le contexte de l'analyse automatique des entretiens vidéo différés (section 3.3.3 p. 33 et section 3.4.3 p. 37).

Au contraire, dans la littérature de l'analyse automatique des émotions et de leurs expressions, l'étude de la fusion multimodale a fait l'objet d'une grande attention. De multiples méthodes ont été proposées, ce qui a permis d'améliorer grandement les performances de détection des émotions (section 3.4.3 p. 37). Ces méthodes, initialement axées sur la fusion au niveau de la décision ou de l'énoncé (fusion tardive), fusionnent maintenant au niveau des mots, mettant en évidence une unité de bas niveau efficace pour une représentation efficiente des modalités [Zadeh et al., 2018a, Zhang et al., 2019, Wang et al., 2019b, Gu et al., 2018, Liang et al., 2018]. Cette fusion fin grain est très importante car cette interaction locale entre les modalités pourrait aider à résoudre l'ambiguïté entre modalités ou à souligner des moments multimodaux particuliers.

Cependant, ces méthodes reposent sur l'utilisation d'une retranscription manuelle (et donc sans bruit) du contenu verbal. Le fait que cette modalité soit potentiellement bruitée pourrait remettre en cause le choix de l'unité de base (le mot) pour la fusion. Notamment, l'obtention d'une transcription manuelle n'est pas toujours possible, et un système réel reposera probablement sur l'utilisation de la reconnaissance automatique de la parole (ASR). Cette retranscription automatique, beaucoup plus bruitée que la retranscription manuelle, pourrait potentiellement entraîner une baisse des performances des méthodes de fusion utilisant le mot comme unité de fusion.

Un autre aspect manquant des systèmes multimodaux proposés est le manque de transparence tant dans la manière dont la fusion est effectuée que dans les connaissances qui peuvent être tirées de ces modèles entraînés. Dans le cadre des ressources humaines et dans un contexte de nouvelles contraintes législatives (règlement général sur la protection des données), l'interprétabilité est cruciale. Une façon possible d'augmenter la capacité d'interprétation serait de comprendre quelles sont les modalités les plus influentes en entretien d'embauche.

Dans la deuxième partie de ce chapitre, nous proposons un mécanisme de fusion à portes à pas de temps réguliers. Ce mécanisme a pour avantage de s'affranchir de l'unité de base : le mot. Nous montrons que dans le cas de l'utilisation d'une retranscription automatique, la fusion à intervalle régulier montre de meilleurs résultats que la fusion au niveau du mot. De plus, notre mécanisme vise à évaluer automatiquement la contribution de chaque modalité à chaque pas de temps, fournissant ainsi une interprétabilité locale aux utilisateurs.

Nous montrons que ce mécanisme de fusion, intégré à HireNet, améliore les performances par rapport à son homologue monomodal. Nous comparons aussi notre modèle multimodal aux modèles multimodaux état de l'art de la littérature de l'analyse automatique de l'entretien vidéo différé.

Publication associée à ce chapitre : En soumission, Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Multimodal Hierarchical Attention Neural Network : Looking for Candidates Behaviour which Impact Recruiter's Decision

6.1 HireNet monomodal, et mécanisme d'attention contextuelle

Nous introduisons un mécanisme d'attention contextuelle qui vise à mieux prendre en compte l'interaction entre le contexte (intitulé du poste et des questions) et l'élément de la séquence de réponse à pondérer. La fonction d'attention originale proposée dans HireNet était strictement additive. Cette forme d'attention a potentiellement conduit à une modélisation sous-efficace de l'interaction entre le contexte et les éléments de la séquence [Pascanu et al., 2020]. Dans ce sens, nous suggérons une nouvelle fonction d'attention à portes. Cette fonction pondère l'adéquation de chaque instant, en fonction du contexte et de sa valeur intrinsèque répondant à la problématique d'une meilleure prise en compte du contexte.

Dans une moindre mesure, nous mettons à jour la représentation textuelle du contenu verbal de la réponse, du titre du poste et des titres des questions par un plongement de mots contextualisés à l'état de l'art, extraite grâce à un modèle français pré entraîné BERT. Le plongement contextuel des mots a permis d'améliorer de nombreuses tâches, comme l'inférence en langage naturel ou l'analyse des sentiments [Peters et al., 2018]. L'utilisation de représentations de mots pré entraînés pourrait être très bénéfique pour l'inférence de l'employabilité.

De plus, une meilleure modélisation du contexte (intitulé des questions et des postes) pourrait permettre un gain de performances. En ce sens, nous avons apporté des modifications à la manière dont le contenu verbal des réponses, des questions et des titres de poste est encodé (Section 6.1.2). Nous utilisons par la suite un modèle français RoBERTa pré entraîné publié par huggingFace [Wolf et al., 2019] pour encoder les mots des réponses candidats et l'intitulé des questions et du poste « CamemBERT » .

6.1.1 Limites de l'attention additive

Dans le chapitre 5 p. 63, nous avons précédemment proposé un mécanisme d'attention pour pondérer les pas de temps dans chaque réponse en fonction du contexte. La fonction d'attention est décrite par les équations 6.1, 6.2 et 6.3.

$$u_t = \tanh(W_k h_t + W_Q H_Q + b_Q) \quad (6.1)$$

$$\alpha_t = \frac{\exp(u_p^\top u_t)}{\sum_{t'} \exp(u_p^\top u_{t'})} \quad (6.2)$$

$$a = \sum_t \alpha_t h_t \quad (6.3)$$

où W_Q , W_k sont des matrices de poids, u_p et b_Q sont des vecteurs de poids et u_p^\top désigne la transposition de u_p . h , et H_Q désignent respectivement la représentation de la séquence des réponses encodées et la représentation de la question encodée. α_t indique la valeur d'attention calculée pour le pas de temps t .

Comme on peut le remarquer, le terme $W_Q H_Q$ est une constante parmi les pas de temps (il ne dépend pas de t) résultant en une modélisation sous-optimale de l'interaction avec le contexte.

Dans ce sens, nous avons construit des expérimentations avec des données simulées disponible en appendices B pour évaluer les limites de l'attention additive. Nous montrons à l'aide de deux scénarios que l'attention telle que définie n'est pas suffisante pour prendre en compte l'interaction avec le contexte et nous proposons une fonction d'attention contextuelle pouvant résoudre notre problématique.

Un schéma des principales modifications est disponible en figure 6.1. Cette nouvelle architecture modifie l'architecture précédemment présentée en chapitre 5, dans deux directions : 1) la modification des fonctions d'attention ; 2) la modification de l'encodage des informations contextuelles (titres de postes et de questions).

6.1.2 Modification des fonctions d'attention

Suite aux résultats concernant les données simulées (voir appendice B), nous nous intéressons à l'utilisation de la fonction d'attention contextuelle dans le cadre de HireNet. Plus précisément, nous utilisons la fonction d'attention décrite ci-dessous pour l'attention temporelle au niveau de la réponse.

$$u_t = \tanh(\lambda_t(W_Q H_Q (W_h h_t^\top) + (1 - \lambda_t)(b^\top (W_k h_t))) \quad (6.4)$$

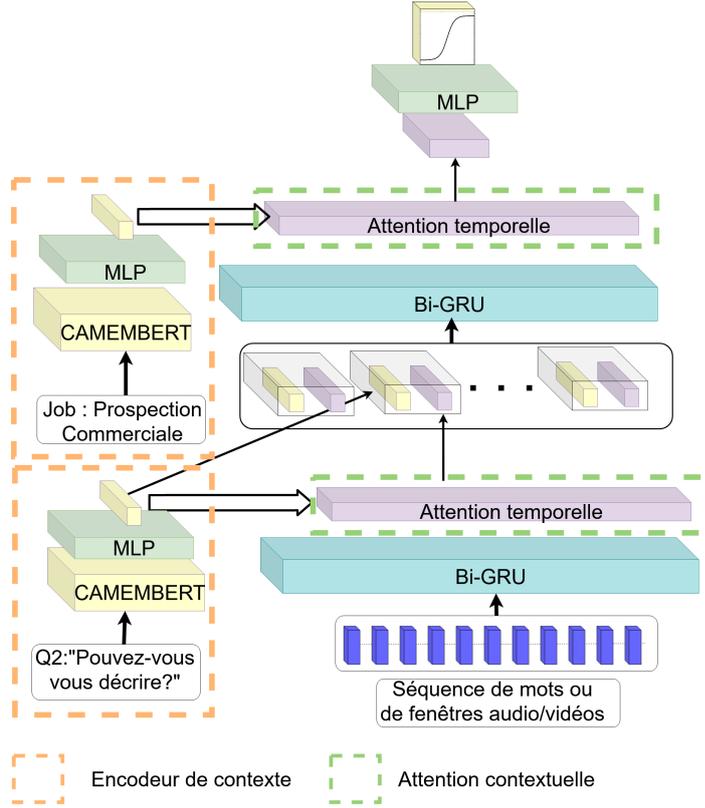


FIGURE 6.1 – Architecture neuronale HireNet modifiée.

Les modifications par rapport à HireNet original sont mises en avant grâce aux encadrés pointillés.

$$\lambda_t = \sigma([W_Q H_Q * W_z h_t; b * W_k h_t]) \quad (6.5)$$

$$\alpha_t = \frac{\exp(u_t)}{\sum_{t'} \exp(u_{t'})} \quad (6.6)$$

$$H_{rponse} = \sum_t \alpha_t h_t \quad (6.7)$$

où W_Q , W_z , W_k sont des matrices de poids, b est un vecteur de poids et b^\top désigne la transposition de b , σ désigne la fonction sigmoïde, $;$ désigne l'opération de concaténation et $*$ désigne le terme opération de produit. H_Q et h_t désignent respectivement la représentation de la question et l'état caché du pas de temps t de la séquence de réponse prosodique, visuelle ou textuelle.

Pour rappel, cette fonction a été construite pour modéliser à la fois l'importance intrinsèque des pas de temps au travers du terme $b^\top (W_k h_t)$ et de l'importance du pas de temps avec le contexte au travers du terme $W_Q H_Q (W_h h_t^\top)$.

Tout comme l'attention temporelle au niveau des réponses, l'attention temporelle au niveau de l'entretien est composée du nouveau mécanisme d'attention proposé :

$$u_i = \tanh(\lambda_i(W_j H_j (W_{qa} Q A_i)^\top) + (1 - \lambda_i)(b^\top (W_l Q A_i))) \quad (6.8)$$

$$\lambda_i = \sigma([W_j H_j * W_{qa} Q A_i; b * W_l Q A_i]) \quad (6.9)$$

$$\alpha_i = \frac{\exp(u_i)}{\sum_{i'} \exp(u_{i'})} \quad (6.10)$$

$$h_{interview} = \sum_i \alpha_i Q A_i \quad (6.11)$$

où W_j , W_{qa} , W_l sont des matrices de poids, b est un vecteur pouvant être entraîné et H_j , QA désignent respectivement la représentation de l'intitulé du poste et la représentation de la séquence de questions-réponses encodée.

Modification des encodeurs Dans une moindre mesure, nous mettons à jour la façon dont les informations contextuelles (questions et titre de poste) sont encodées. Nous les représentons par l'utilisation du modèle CamemBERT pré entraîné [Martin et al., 2020]. Nous utilisons la représentation au niveau de la phrase, ce qui nous fournit un vecteur de dimension 768.

Ensuite, nous utilisons un perceptron à une couche afin de compresser la représentation originale et de réduire le nombre de paramètres. Notre objectif est qu'ainsi des contextes similaires puissent avoir des représentations similaires.

Nous encodons les questions de la façon suivante :

$$H_Q = \tanh(W_{QE} M_Q) \quad (6.12)$$

où W_{QE} est une matrice de poids entraînable, M_Q est la représentation BERT originale de la question et H_Q est la représentation compressée résultante.

Nous encodons le titre de poste de la façon suivante :

$$H_J = \tanh(W_{JE} M_J) \quad (6.13)$$

où W_{JE} est une matrice de poids pouvant être formée, M_J est la représentation BERT originale du titre du poste et H_J est la représentation compressée résultante.

6.2 Expériences monomodales

Nous effectuons différentes expériences afin de valider ou d’invalider les hypothèses suivantes :

- H1 : Le modèle HireNet obtient de meilleures performances que les modèles état de l’art sur un second jeu de données
- H2 : L’utilisation de la représentation de mots contextualisés et la modification des encodeurs de contextes (intitulé des questions et du poste) améliorent la performance de HireNet décrit en chapitre 5.
- H3 : L’attention contextuelle permet une meilleure interaction avec le contexte induisant de meilleures performances que l’attention additive.

Nous décrivons dans la suite de cette section les métriques d’évaluation, le jeu de données et les modèles comparés afin de valider ou d’invalider les précédentes hypothèses.

6.2.1 Jeu de données et métriques d’évaluations

Nous utilisons le second jeu de données (JDD2) précédemment présenté dans la section 4.4 p. 59. Pour rappel, la plupart des vidéos du premier jeu de données était arrivé à expiration, en terme de durée de conservation, d’où la nécessité de constituer un second jeu de données. Des statistiques descriptives sont disponibles en tableau 6.1.

Les métriques d’évaluation choisies sont l’aire sous la courbe ROC et la moyenne du score F1 de la classe *convocable* et *non-convocable*. Elles sont bien adaptées à la classification binaire et utilisées dans des études antérieures [Naim et al., 2018, Chen et al., 2017]. Ce choix permet d’aller au-delà de l’évaluation de la performance sur la seule classe positive initialement choisie en chapitre 5.

Nous avons divisé l’ensemble de données en un ensemble d’entraînement, un ensemble de validation pour la sélection des hyperparamètres en fonction de l’AUC, et un ensemble de test pour l’évaluation finale de chaque modèle. Ces sous-ensembles constituent respectivement 70 %, 15 % et 15 % de l’ensemble des données.

6.2.2 Modèles de références naïfs

Nous comparons notre modèle avec plusieurs modèles de référence naïfs :

1. **Vote aléatoire.** Un millier de tirages au sort respectant la balance du jeu de données d’entraînement ont été effectués. La moyenne des résultats est ensuite calculée sur l’ensemble des tirages.

2. **Vote majoritaire.** Ce modèle de référence consiste à attribuer le label majoritaire.
3. **Vote majoritaire par poste.** Ce modèle de référence consiste à attribuer le label majoritaire par campagne de recrutement afin de vérifier que le modèle n'apprend pas seulement à prédire l'origine de l'entretien d'embauche.

6.2.3 Modèles de références état de l'art

Nous comparons notre modèle avec des modèles de références issus de la littérature sur l'analyse automatique d'entretiens d'embauche. Ces approches sont comparables à celle du chapitre 5.

1. **Modèles avec descripteurs agrégés** Nous appliquons des fonctions statistiques classiques pour réduire la dimension temporelle et obtenir un vecteur fixe pour chaque modalité. La moyenne, l'écart-type, le minimum, le maximum, la somme des gradients positifs et la somme des gradients négatifs sont appliqués aux séquences audio et vidéo. La moyenne est utilisée pour la modalité linguistique. Cette approche est l'une des plus utilisées parmi les travaux précédents [Nguyen and Gatica-Perez, 2015, Naim et al., 2018, Rasipuram and Jayagopi, 2018].
2. **Modèles avec descripteurs représentés par Sac de Mots Audio/ Video / Languages** Nous avons choisi de comparer notre modèle à la représentation Sac de Mots Audio et Vidéos de [Chen et al., 2017] : nous exécutons un algorithme K-means sur toutes les trames de bas niveau de notre ensemble de données. Nous prenons ensuite nos échantillons comme documents, et les classes prédites de nos trames comme mots, et utilisons une représentation "Term Frequency-inverse Document Frequency" (TF-IDF) pour modéliser chaque entretien.

6.2.4 Modèles de références HireNet

Nous comparons différentes variantes d'HireNet afin de mesurer l'efficacité des modifications apportées par rapport à la première version du modèle décrite dans le chapitre 5. Dans ce sens, nous comparons les modèles suivants :

1. HireNet comme décrit dans le chapitre 5, les mots sont encodés grâce à word2Vec pour la modalité du langage.
2. HireNet avec les modifications des encodeurs et des descripteurs textuels. La fonction d'attention est toujours celle de HireNet original c'ad additive.
3. HireNet avec les modifications des encodeurs et des descripteurs textuels. La fonction d'attention est la fonction contextuelle proposée en sous-section. 6.1.2

Jeu de données	Entraînement	Validation	Test
Nombre de candidats	3581	784	783
Nombre de campagnes de recrutement	455	225	219
Moyenne de questions par entretien	5.43	5.46	5.41
Temps moyen par réponse	54.7 s	53.4 s	53.9 s
Durée totale des entretiens vidéos	285.7h	61.6h	61.5h
Proportion de l'étiquette convocable	55 %	55 %	54 %

TABLE 6.1 – Tableau descriptif du jeu de données 2 : nombres de candidats et statistiques globales.

A noter que les séquences d'entrée pour l'audio et la vidéo ont été construites de la même façon que lors du chapitre 5.

6.3 Résultats pour les expériences monomodales

Le tableau 6.2 résume les résultats des expériences monomodales.

Premièrement, nous pouvons constater que l'hypothèse H1 est validée : de nouveau, les modèles de référence HireNet surpassent en performance tous les modèles de référence de l'état de l'art sur ce second jeu de données. Cette première constatation renforce les premiers résultats obtenus dans le chapitre 5, le modèle proposé modélise bien la structure de l'entretien vidéo différé.

Deuxièmement, l'hypothèse H2 est validée pour les modalités langage et audio : notre expérience montre que la modification des méthodes d'encodage du contexte et de la représentation textuelle contribue à l'obtention de meilleures performances pour les modalités audio et linguistiques. Cependant, une dégradation des performances pour la modalité vidéo est visible. Ce résultat pourrait être expliqué par le fait que l'attention au niveau de l'entretien se focalise plus sur l'intitulé des questions qu'auparavant et que les comportements non verbaux contenus dans les réponses associées ne soient pas les plus informatifs.

Troisièmement, l'hypothèse H3 est mitigée : l'attention proposée montre de meilleurs résultats pour la modalité linguistique, mais de moins bons résultats pour la modalité audio. Il n'y a pas d'effet pour la modalité vidéo. Ce résultat pourrait être expliqué par l'hypothèse que l'interaction entre le contexte (intitulé des questions et du poste) et le contenu de l'entretien soit importante pour la modalité linguistique, mais peu importante pour les comportements non verbaux.

Dans l'ensemble, l'amélioration de la prise en compte du contexte est bénéfique. La

modalité du langage est la modalité profitant le plus de la meilleure prise en compte du contexte. Ce phénomène n'est pas étonnant dans le sens où le fond de la réponse est largement conditionné par les questions posées. Enfin, il semble que l'influence du contexte soit moins importante pour les autres modalités et notamment pour la modalité vidéo qui comporte essentiellement des expressions faciales.

Modèles naïfs	AUC		F1			
Vote aléatoire	0.50		0.50			
Vote majoritaire	0.50		0.354			
Vote majoritaire par position	0.630		0.627			
Modèles de référence SOA	Language		Audio		Video	
	AUC	F1	AUC	F1	AUC	F1
Sac de mots	0.574	0.512	0.589	0.551	0.552	0.532
Descripteurs agrégés	0.613	0.570	0.624	0.584	0.579	0.549
Modèles de références HireNet	Language		Audio		Video	
	AUC	F1	AUC	F1	AUC	F1
HireNet	0.693 ± 0.002	0.589 ± 0.042	0.706 ± 0.005	0.628 ± 0.026	0.692 ± 0.005	0.617 ± 0.009
HireNet + BERT	0.706 ± 0.011	0.643 ± 0.011	0.736 ± 0.005	0.668 ± 0.008	0.682 ± 0.012	0.615 ± 0.008
HireNet + BERT + Attention contextuelle	0.717 ± 0.007	0.644 ± 0.013	0.727 ± 0.006	0.656 ± 0.009	0.683 ± 0.010	0.620 ± 0.005

TABLE 6.2 – Résultats des expériences évaluant HireNet sur le second jeu de données.

6.4 Multimodal Hirenet, la multimodalité en pratique

Notre modèle précédent était conçu pour être uniquement utilisé dans un cadre monomodal. Pour combler cette lacune, nous proposons un nouveau modèle appelé *Multimodal HireNet* pour la prédiction de la convocabilité qui étend HireNet en intégrant un module supplémentaire nommé "l'encodeur multimodal" (voir figure 6.2). Deux composantes principales constituent cet encodeur (voir figure 6.3).

Premièrement, nous proposons une nouvelle méthode suprasegmentale pour obtenir une représentation multimodale. Notre méthode consiste à extraire la dynamique intramodale en respectant le taux d'échantillonnage spécifique à chaque modalité (Section 6.4.3) puis à sous-échantillonner parmi les états cachés de chacune des modalités (Section 6.4.4), à un intervalle de temps régulier pour obtenir la représentation multimodale.

Cette méthode a deux avantages par rapport à la fusion au niveau du mot :

- Cette méthode permet de se soustraire à la nécessité d'un contenu verbal. Cet aspect est d'autant plus important que la qualité des vidéos est très variable et qu'il est courant de faire face à des erreurs de reconnaissance vocale.
- Cette méthode permet de prendre en compte la structure fin grain des comportements

non verbaux au niveau des sous-mots, structure qui n'est pas prise en compte lorsque l'on moyenne les valeurs audio et vidéos lors d'une fusion au niveau du mot [Wang et al., 2019b].

Deuxièmement, nous étudions les avantages des cellules neuronales de fusion multimodale interprétables, à savoir les Gated Multimodal Units (GMU) [Arevalo et al., 2020]. Cette cellule projette chacune des modalités dans un espace joint et pondère l'importance de chacune des modalités par un scalaire. Nous décrivons par la suite cette cellule. Nous avons choisi cette méthode de fusion car elle est interprétable (en permettant d'examiner la contribution de chacune des modalités), compétitive (par rapport à une simple concaténation des modalités) et reste plus facile à mettre en œuvre que les autres méthodes de fusion section 3.4.3 p. 37. En outre, comme les modalités pourraient être très bruitées, nous espérons que le mécanisme de porte pourrait répondre à ce problème en diminuant l'influence des modalités bruitées.

Dans les sous-sections suivantes, nous décrivons le nouveau composant "Encodeur multimodal" qui encode chaque vidéo réponse multimodale du candidat.

6.4.1 Formalisation

Nous indiquons avec $X_{\{L,V,A\}}^i \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$ la séquence de descripteurs de bas niveaux décrivant la i -ième réponse respectivement des trois modalités : langage (L), vidéo (V) et audio (A). T_m et d_m représentent la longueur de la séquence et la dimension de la représentation de la modalité $m \in \{L, V, A\}$ car chaque modalité a une fréquence d'échantillonnage et une dimension de représentation différente.

6.4.2 Représentation des séquences d'entrée multimodales

Comme nous voulons fusionner les modalités à un intervalle de temps régulier, nous devons trouver une représentation efficace des séquences d'entrée. Pour chaque modalité de la réponse candidate, nous construisons une représentation par trame. Notre modèle peut gérer les différentes fréquences de chaque modalité. Ces séquences de descripteurs sont représentées dans le bas de la figure 6.2. Ainsi, nous utilisons les mêmes descripteurs que la section 6.2.4 à la seule différence des fréquences d'échantillonnage choisies.

Expression faciale :

Nous extrayons les caractéristiques visuelles au niveau de la trame avec OpenFace [Baltrusaitis et al., 2018] comme ce qui a été fait dans la section précédente et le chapitre 5. Nous avons choisi d'extraire, pour chaque image, la position et la rotation de la tête, l'in-

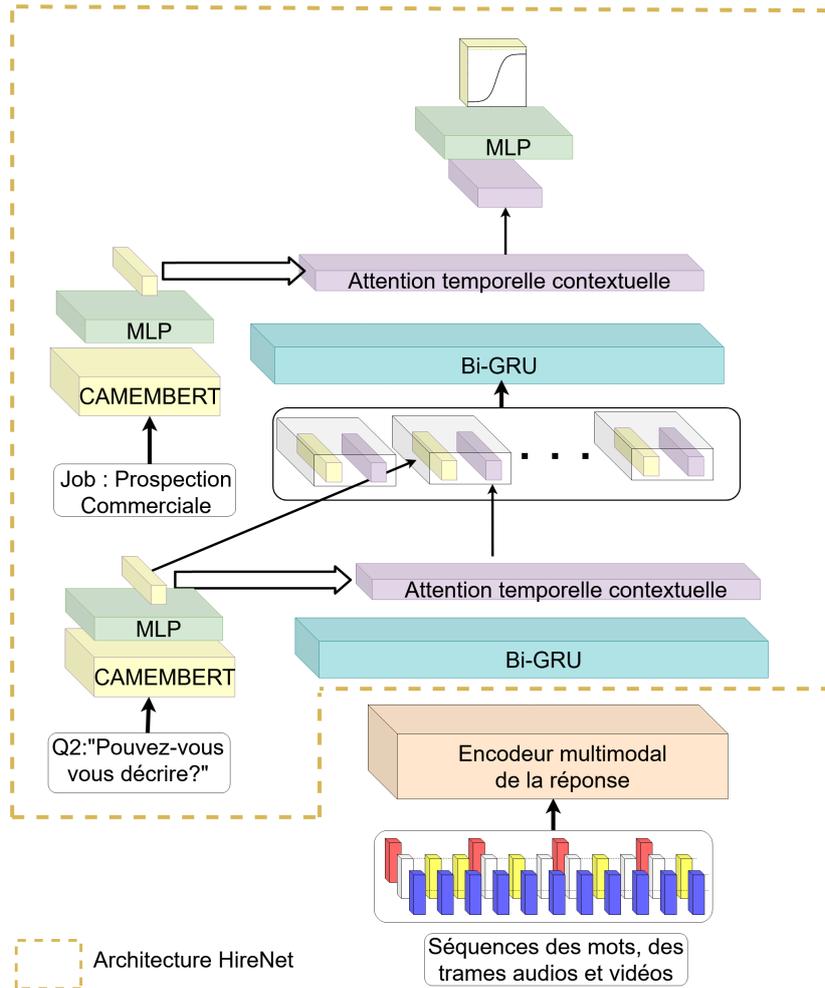


FIGURE 6.2 – Architecture du Multimodal HireNet

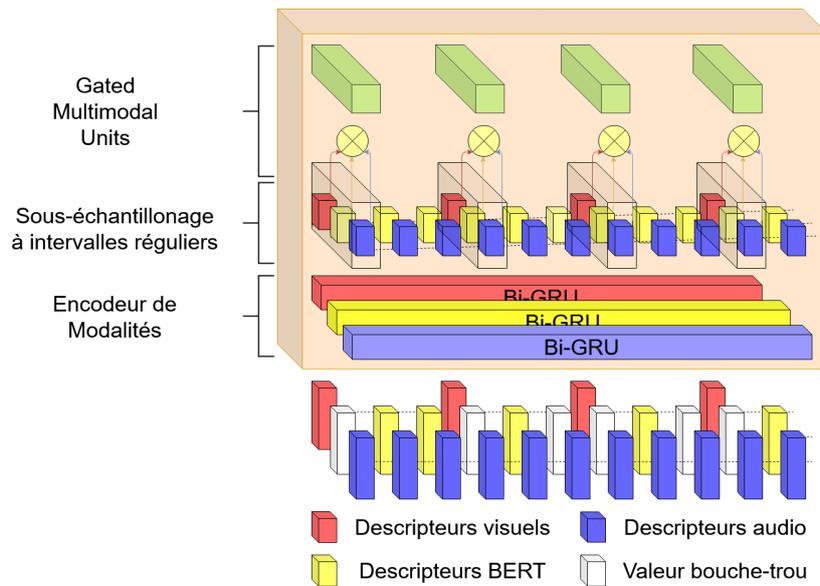


FIGURE 6.3 – Encodeur multimodal de la réponse

Figures décrivant le réseau Multimodal HireNet.

tensité et la présence des unités d'action, et la direction du regard, ce qui donne un vecteur de dimension 52. A la différence des précédentes expériences, nous avons décidé de lisser les valeurs avec une fenêtre temporelle de 0,3 s sans chevauchement.

Indices prosodiques : Nous utilisons l'ensemble de descripteurs bas niveaux eGeMAPS affective [Eyben et al., 2016] comme ce qui a été fait dans la section précédente et le chapitre 5. Grâce à OpenSmile [Eyben et al., 2013b], nous extrayons un vecteur de dimensions 23 à la fréquence de 100Hz. A la différence des précédentes expériences, nous lissons ces valeurs avec une fenêtre temporelle de 0,1 s.

Contenu verbal : Nous utilisons un outil de reconnaissance automatique de la parole¹ pour acquérir la transcription du contenu verbal et l'horodatage de chaque mot transcrit.

Le contenu verbal de la transcription est ensuite transformé en une séquence de vecteurs de plongement de mots contextualisés comme ce qui a été fait en section 6.2.4. En utilisant « CamemBERT » un modèle français RoBERTa pré entraîné publié par huggingFace [Wolf et al., 2019], nous obtenons une représentation de dimension 768 pour chaque mot de la transcription verbale.

Comme nous voulons aligner temporellement le contenu verbal, nous modifions la séquence du langage. Plus précisément, nous construisons la séquence de la modalité linguistique comme une séquence de valeurs "bouche-trou", sauf aux étapes temporelles où un mot a fini d'être prononcé (voir la séquence linguistique de la figure 6.3). À ce pas de temps spécifique, nous remplaçons le vecteur "bouche-trou" par la représentation BERT du mot prononcé. En outre, Bi-GRU ignorera les valeurs "bouche-trou", par un mécanisme de masquage, pendant le traitement de la séquence. Comme la reconnaissance vocale automatique nous donne l'horodatage de chacun des mots avec une précision de 0,1 s, nous choisissons cette valeur comme pas de temps pour la séquence linguistique.

6.4.3 Encodeur de modalité

L'encodeur de modalité est le premier module constituant l'encodeur multimodal visible en figure 6.3. Cette partie du modèle vise à encoder les séquences des descripteurs de bas niveau pour chaque modalité. Chaque modalité a son encodeur unimodal, qui assure une meilleure représentation de la dynamique intramodale [Poria et al., 2017]. Un GRU bidirectionnel est utilisé pour obtenir les représentations dans les deux sens pour chaque élément de la séquence X . L'encodage des séquences de manière bidirectionnelle garantit la

1. Google speech-to-text API.

même quantité d'informations préalables pour chaque élément de $(x_t)_{1 \leq t \leq T_m}$.

$$\vec{z}_t^m = BiGRU(x_t), t \in [1, T_m]. \quad (6.14)$$

6.4.4 Sous-échantillonnage à pas de temps réguliers

La sortie de l'encodeur de modalités a une fréquence d'échantillonnage différente selon la modalité (*c.-à-d.* pas de temps de 0,3 s pour la modalité vidéo et 0,1 s pour les modalités audio et linguistiques). Nous effectuons un sous-échantillonnage parmi les états cachés de l'encodeur de modalités pour fusionner les modalités à intervalles de temps réguliers. Plus précisément, pour une fréquence choisie f et une séquence à sous-échantillonner Z_m avec une fréquence f_m d'une longueur T_m , la durée d'un pas de temps η à échantillonner est :

$$\eta^m = \frac{f}{f_m} \quad (6.15)$$

$$f_\eta^m = \frac{1}{\eta^m} \quad (6.16)$$

$$\widetilde{Z}^m \leftarrow (z_{i * f_\eta^m}^m)_{1 \leq i \leq T_m * f_\eta^m} \quad (6.17)$$

Ainsi, grâce à l'utilisation d'un encodeur et d'un sous-échantillonnage propre à chaque modalité, l'alignement tient compte de la dynamique intramodale et de leur fréquence respective. La fréquence f pourrait être n'importe quelle valeur tant que η^m est un entier. Dans notre cas, nous avons choisi un pas de temps égal à 0,3 s.

6.4.5 Gated Multimodal Unit

Le Gated Multimodal Unit (GMU) est un composant neuronal qui vise à fusionner différentes modalités [Arevalo et al., 2020]. Plus précisément, le GMU projette chaque modalité dans un espace commun et apprend à contrôler la contribution de chaque modalité par le biais d'un mécanisme à portes.

Une fois qu'une représentation de tous les éléments de chaque modalité (langue, audio

et vidéo) est obtenue, une représentation multimodale h est calculée comme suit :

$$\begin{aligned}
h_t^a &= \tanh(W_{Aprojection}\tilde{z}_t^a) \\
h_t^l &= \tanh(W_{Lprojection}\tilde{z}_t^l) \\
h_t^v &= \tanh(W_{Vprojection}\tilde{z}_t^v) \\
\sigma_t^a &= \sigma(W_{Agating}[\tilde{z}_t^a, \tilde{z}_t^l, \tilde{z}_t^v]) \\
\sigma_t^v &= \sigma(W_{Lgating}[\tilde{z}_t^a, \tilde{z}_t^l, \tilde{z}_t^v]) \\
\sigma_t^l &= \sigma(W_{Vgating}[\tilde{z}_t^a, \tilde{z}_t^l, \tilde{z}_t^v]) \\
z_t^{multimodal} &= \sigma_t^a * h_t^a + \sigma_t^l * h_t^l + \sigma_t^v * h_t^v
\end{aligned} \tag{6.18}$$

où $W_{Aprojection}$, $W_{Lprojection}$, $W_{Vprojection}$, $W_{Agating}$, $W_{Lgating}$, $W_{Vgating}$ sont des matrices de poids, et \tilde{z}_a , \tilde{z}_v , \tilde{z}_l désignent respectivement la représentation de la modalité audio, vidéo et linguistique. Enfin, $z_t^{multimodal}$ résume les informations de la représentation multimodale au niveau du pas de temps t .

Une fois la réponse multimodale encodée, on peut procéder de la même manière que pour HireNet : la séquence multimodale remplace la séquence monomodale originale. A noter que la dynamique intermodale est alors encodée par l'encodeur de réponse Bi-GRU. A noter, que nous utilisons l'attention temporelle contextuelle décrite en section section 6.1.2, de même, nous utilisons les encodeurs de contexte précédemment décrits dans la même section.

6.5 Expériences Multimodales

6.5.1 Jeu de données et métriques d'évaluations

Nous utilisons le deuxième jeu de données, identique à celui utilisé dans la section précédente. Les métriques d'évaluation choisies sont l'aire sous la courbe ROC (AUC) et la moyenne du score F1 des classes *convocable* et *non-convocable*. Ces métriques sont les mêmes que précédemment, ce qui nous permettra de comparer HireNet multimodal à son homologue monomodal. Nous avons divisé l'ensemble de données en un ensemble de formation, un ensemble de validation pour la sélection des hyperparamètres en fonction de la AUC, et un ensemble de tests pour l'évaluation finale de chaque modèle. Ces sous-ensembles constituent respectivement 70 %, 15 % et 15 % de l'ensemble des données et sont identiques à ceux utilisés dans la précédente section.

6.5.2 Modèles de référence de l'état de l'art en entretien d'embauche

Nous comparons *Multimodal HireNet* avec les précédents modèles monomodaux évalués dans la section précédente. De plus, nous comparons notre modèle avec différentes bases de référence multimodales correspondant à l'état de l'art des entretiens d'embauche :

1. **Fusion tardive** Moyenne de la sortie des modèles monomodaux (descripteurs agrégés temporellement) [Rasipuram and Jayagopi, 2018].
2. **Fusion précoce** Nous concaténons les descripteurs agrégés temporellement des trois modalités comme cela a été fait dans [Naim et al., 2018, Rasipuram and Jayagopi, 2018].
3. **Sac de Mots Multimodaux** Similaire au sac de mots monomodaux sauf que nous prenons comme entrée la concaténation des vecteurs audio, vidéo et mots. Afin d'aligner les modalités, nous faisons la moyenne des valeurs des descripteurs vidéo et audio au niveau des mots.

6.5.3 Modèles de références HireNet

Le réseau multimodal HireNet proposé dans la section 6.4 repose sur quatre hypothèses.

- (H1) Un modèle multimodal fournira un modèle plus efficace qu'un modèle monomodal.
- (H2) La représentation au niveau du mot détériorera les performances dans le cas de l'utilisation d'un contenu verbal retranscrit automatiquement.
- (H3) Une fusion à intervalles réguliers peut permettre de surmonter cette dépendance à l'égard de la qualité de la reconnaissance automatique de la parole.
- (H4) Une méthode de fusion interprétable utilisée conjointement avec cette fusion à intervalles réguliers pourrait améliorer les performances en contrôlant l'importance de chaque modalité.

Afin de vérifier ces hypothèses, nous comparons le Multimodal HireNet avec trois autres bases de référence. Nous fournissons un schéma de l'architecture et une description pour chacune de ces bases de références :

1. **Fusion intermédiaire au niveau de la question-réponse.** Un schéma de cette base de référence est disponible en figure 6.4. Ce modèle consiste à concaténer les encodeurs de réponse de trois HireNet monomodaux au niveau de la question-réponse. Nous faisons l'hypothèse que ce modèle sera plus performant que les modèles de référence monomodaux (H1). Cependant, comme nous fusionnons les modalités au niveau des

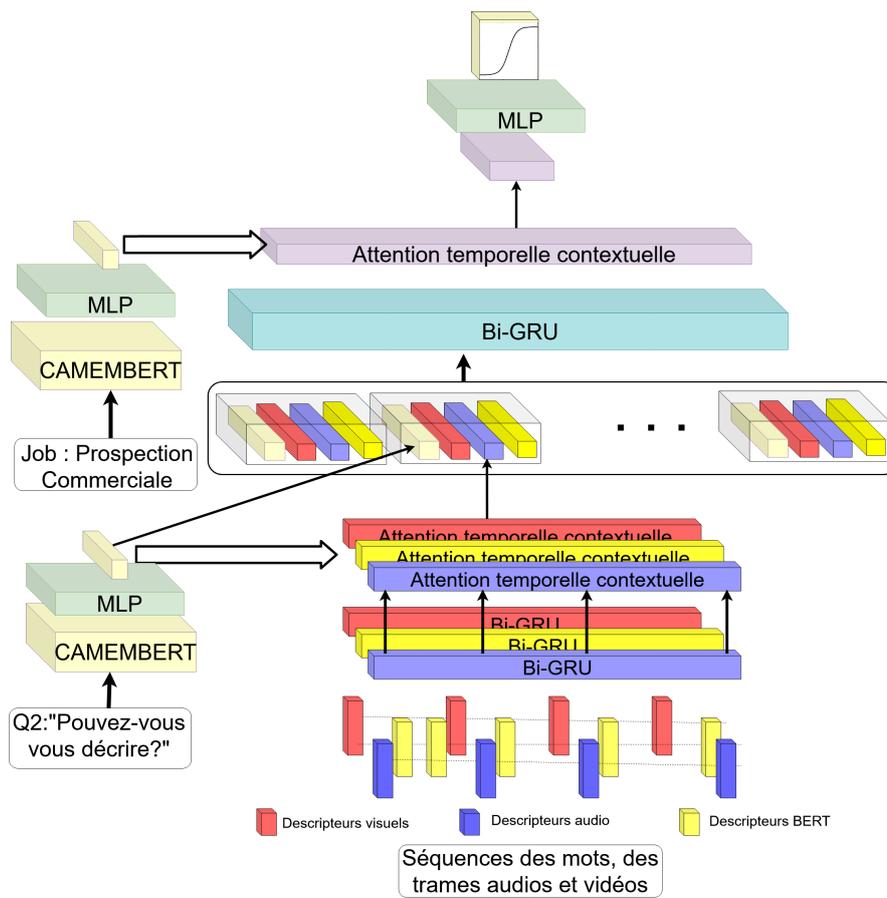
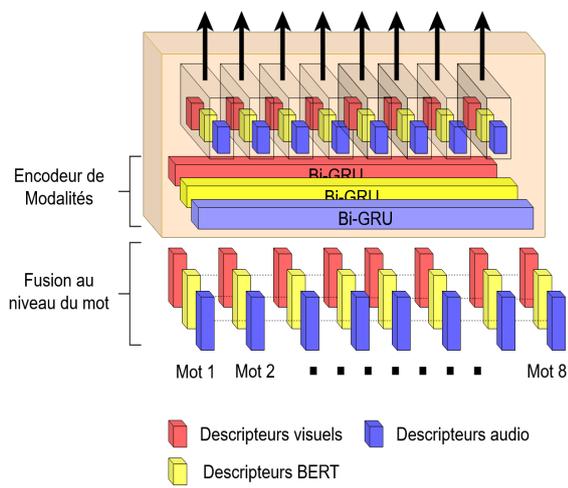
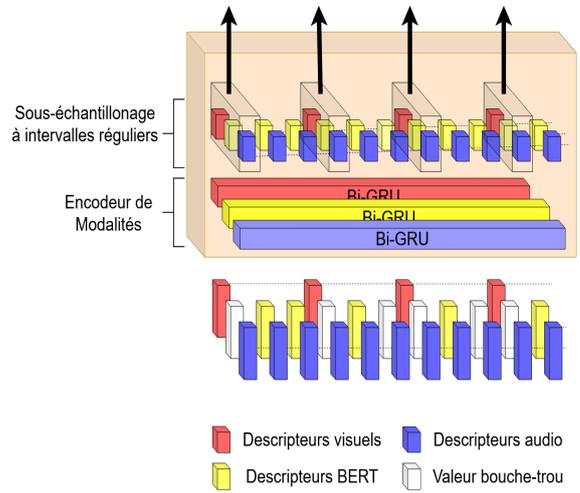


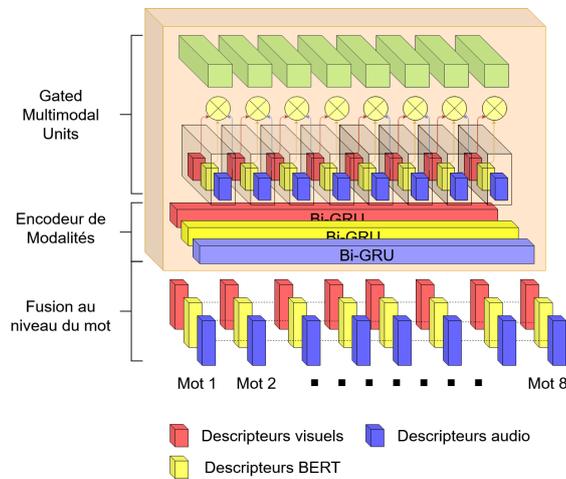
FIGURE 6.4 – Modèle de référence de la fusion intermédiaire au niveau de la question-réponse



(a) Modèle de référence de la fusion au niveau du mot



(b) Modèle de référence de la fusion à intervalle de temps régulier



(c) Modèle de référence de la fusion au niveau de mot par GMU

FIGURE 6.5 – Figures décrivant les modèles de références pour l’encodeur multimodal.

questions-réponses, le modèle ne prend pas en compte l'interaction de bas niveau entre les modalités telles que HireNet Multimodal.

2. **Fusion au niveau du mot.** Ce modèle modifie l'encodeur multimodal, un schéma de référence est disponible en figure 6.5a. Ce modèle tire profit des informations d'horodatage fournies par la sortie de la reconnaissance vocale automatique des mots prononcés. Nous suivons la plupart des pipelines multimodales et les valeurs des caractéristiques vidéo et audio sont moyennées pendant la durée du mot (voir les sections 3.4.3). Chaque caractéristique unimodale contextuelle est apprise par les encodeurs de modalités des réponses et les valeurs cachées sont concaténées à chaque étape. Ainsi, nous retirons du Multimodal HireNet la composante GMU et la représentation est fusionnée au niveau du mot. Un tel réglage est très comparable au modèle état de l'art [Poria et al., 2017]. Nous supposons qu'un tel modèle sera moins performant que la fusion intermédiaire au niveau de la question-réponse, même si la granularité de la fusion est plus fine (**H2**).
3. **Fusion à intervalle de temps régulier.** Ce modèle modifie l'encodeur multimodal, un schéma de référence est disponible en figure 6.5b. Contrairement à la fusion au niveau du mot, ce modèle fusionne les modalités à un intervalle de temps régulier. La seule différence avec Multimodal HireNet réside dans l'absence du composant GMU. Nous supposons qu'une telle fusion sera moins sujette aux erreurs de reconnaissance vocale automatique et, par conséquent, sera plus performante que le modèle de fusion au niveau du mot (**H3**). En outre, ce modèle nous permettra de comprendre les avantages de l'utilisation de l'unité GMU en comparant ses performances avec celles de Multimodal HireNet (**H4**).
4. **HireNet Fusion au niveau du mot par GMU.** Ce modèle modifie l'encodeur multimodal, un schéma de référence est disponible en figure 6.5c. Ce modèle est construit sur le modèle *HireNet Fusion au niveau du mot*. Ce modèle est évalué pour confirmer ou infirmer que c'est la combinaison de l'utilisation des GMU et d'un alignement suprasegmental qui sont responsables du biais inductif. Ainsi, la comparaison de ce modèle avec *HireNet Fusion au niveau du mot*, *HireNet Fusion à intervalle de temps régulier* et *Multimodal HireNet* mettra en perspective les modifications proposées (**H4**).

Détails d'entraînements Comme le temps d'entraînement et la combinaison des hyperparamètres dans le cadre multimodal explosent, nous décidons de limiter notre recherche d'hyperparamètres au seul niveau de l'entretien du Multimodal HireNet (c'est-à-dire le

Modèles de référence SOA monomodal	Language		Audio		Video	
	<i>AUC</i>	F1	<i>AUC</i>	F1	<i>AUC</i>	F1
Sac de mots	0.574	0.512	0.589	0.551	0.552	0.532
Descripteurs agrégés	0.613	0.570	0.624	0.584	0.579	0.549
Modèles de référence HireNet monomodal	Language		Audio		Video	
	<i>AUC</i>	F1	<i>AUC</i>	F1	<i>AUC</i>	F1
HireNet Attention Contextuelle	0.717 ± 0.007	0.644 ± 0.013	0.727 ± 0.006	0.656 ± 0.009	0.683 ± 0.010	0.620 ± 0.005
Modèles de référence SOA multimodal	AUC			F1		
	Fusion précoce	0.643			0.601	
Fusion tardive	0.636			0.579		
Sac de mots multimodaux	0.596			0.562		
Modèles de référence HireNet multimodal	AUC			F1		
	HireNet Fusion intermédiaire (<i>H1</i>)	0.741 ± 0.004			0.671 ± 0.005	
HireNet Fusion au niveau du mot sans GMU (<i>H2</i>)	0.705 ± 0.012			0.635 ± 0.007		
HireNet Fusion par intervalle de temps régulier (<i>H3-H4</i>)	0.729 ± 0.001			0.657 ± 0.018		
HireNet Fusion au niveau du mot avec GMU (<i>H4</i>)	0.718 ± 0.010			0.660 ± 0.005		
Multimodal HireNet (<i>H4</i>)	0.749 ± 0.008			0.689 ± 0.015		

TABLE 6.3 – Comparaison des performances des modèles de références monomodales et multimodales par rapport Multimodal HireNet.

nombre de neurones de l’encodeur question-réponse). Nous avons choisi d’initialiser certains modules par ceux préentraînés dans le cadre monomodal nommément : les encodeurs de modalités, l’encodeur de questions et l’encodeur du titre de poste. Nous espérons ainsi disposer d’une architecture efficace tout en accélérant le temps d’entraînement. A noter que ce choix peut conduire à des résultats sous-optimaux pour le système multimodal car le réseau peut être bien dimensionné pour les modèles monomodaux mais pas pour le modèle multimodal. Nous utilisons une recherche par grille pour trouver les hyperparamètres les mieux adaptés. Afin d’évaluer l’intervalle de confiance, nous reproduisons cinq fois toutes les expériences concernant les architectures neuronales, l’initialisation aléatoire pouvant avoir un impact sur le résultat.

6.6 Résultats pour les expériences multimodales

Le tableau 6.3 résume les résultats de la tâche de prédiction de la convocabilité. Les modèles multimodaux à l'état de l'art ont tous de meilleurs résultats que leur homologue monomodal respectif. Ce résultat est conforme à celui de [Naim et al., 2018, Rasipuram and Jayagopi, 2018, Chen et al., 2016a]. Ces résultats confirment l'hypothèse H1 même pour les modèles de références état de l'art : les modèles multimodaux donnent de meilleurs résultats que les modèles monomodaux.

Concernant les modèles de références HireNet, l'AUC et le score F1 ont augmenté, passant du modèle HireNet monomodal à HireNet *Fusion au niveau de la question* qui supporte H1. On peut également constater que la *fusion au niveau du mot* montre une baisse significative de performance par rapport au modèle *Fusion au niveau de la question* validant H2. Cela suggère que le bruit de la reconnaissance vocale automatique diminue les performances lorsqu'une fusion au niveau du mot est choisie. Pour l'hypothèse H3, la fusion à intervalles de temps réguliers montre des résultats nettement meilleurs que la fusion au niveau des mots. En fin de compte, notre modèle HireNet multimodal surpasse tous les modèles précédents, en particulier *Fusion par intervalles de temps réguliers* et *Fusion au niveau des mots avec GMU*. Par conséquent, la combinaison de la fusion à intervalles de temps réguliers et d'un mécanisme de fusion pourrait améliorer et permettre une fusion à une granularité plus fine que la fusion au niveau des questions sans perdre (et même en améliorant légèrement) les performances, supportant H4.

6.7 Conclusion

Ce chapitre répond à deux limites de l'état de l'art de l'analyse automatique des entretiens vidéo différés, à savoir l'absence de la prise en compte effective du contexte et la gestion de la multimodalité.

Premièrement, nous avons proposé une nouvelle fonction d'attention temporelle modélisant de façon plus précise l'interaction avec le contexte. L'utilisation de cette fonction d'attention montre des améliorations de performance pour la modalité du langage mais peu d'améliorations pour la modalité vidéo et audio. Ainsi, l'intitulé de la question semble conditionner l'importance des mots (c.-à-d. selon le titre de la question, certains mots peuvent devenir importants) au contraire des comportements non verbaux qui ne semblent pas dépendre du contexte de la question (c.-à-d. peu importe le titre de la question, les comportements non verbaux importants restent les mêmes). D'une façon intéressante, notre conclusion se rapproche des travaux de [Peeters and Lievens, 2006]. D'une manière similaire,

ces travaux montrent que l'usage des stratégies d'impressions verbales (voir section 2.6) varie selon le type de question posé (situationnelle vs comportementale) contrairement aux stratégies d'impressions non verbales. Enfin, dans une moindre mesure, la mise à jour de la façon dont l'intitulé des questions et du titre de poste est encodé montre une amélioration de performances pour les modalités du contenu verbal et audio.

Deuxièmement, nous avons proposé une méthode de fusion multimodale pour l'encodage des vidéos réponses des candidats. Cette méthode va au-delà de la concaténation des représentations de chacune des modalités en proposant une fusion à un niveau plus fin grain. Elle repose sur plusieurs contraintes pratiques : la multimodalité grâce à un mécanisme de fusion adéquat, le bruit de la sortie de l'ASR et l'interprétabilité. Nos expériences montrent que 1) les systèmes multimodaux donnent de meilleurs résultats que leur homologue monomodal, 2) la méthode de fusion par l'utilisation du GMU conduit à une amélioration des performances, et 3) les approches de fusion au niveau du mot sont moins efficaces que les autres méthodes de fusion que nous étudions, que nous interprétons comme une conséquence d'une sortie d'ASR imparfait. Cela contredit les paramètres habituels des études multimodales qui reposent sur la transcription manuelle du discours.

Enfin, l'unité GMU permet d'analyser la contribution de chaque modalité à chaque pas de temps de la fusion. Cette fusion plus fine permet de prendre en compte les interactions bas niveaux entre modalités et pourrait permettre de mettre en lumière, grâce à l'attention temporelle, des comportements multimodaux que nous étudions par la suite en chapitre 7.

Troisième partie

Vers une meilleure interprétabilité et
équité des modèles de convocabilité

Section 7

Post-analyse des mécanismes d'attention : vers une interprétation des moments clés

Nous avons proposé un modèle unimodal (HireNet) et un modèle multimodal (Multimodal HireNet chapitre 6 p. 78) qui utilisent des mécanismes d'attention pour fournir une interprétabilité locale et ce de trois manières : en informant de l'importance de chaque moment pendant une réponse, en informant de la contribution de chaque modalité en chacun des moments et en informant de l'importance de chaque question-réponse au niveau de l'entretien.

Ces mécanismes d'attention nous fournissent une méthode d'interprétabilité locale intrinsèque en indiquant candidat par candidat les moments les plus importants dans ses réponses, les modalités les plus importantes dans ses réponses et les questions-réponses les plus influentes dans son entretien. Cependant, la validité comme méthode d'interprétabilité de ces mécanismes d'attention a été discutée récemment (voir section 3.5.1). De plus, ces mécanismes nous fournissent uniquement une interprétabilité locale insuffisante pour comprendre en quoi consistent généralement les moments les plus importants de l'entretien d'embauche.

Dans cette section, nous menons une analyse approfondie de ces mécanismes d'attention pour obtenir une interprétabilité plus générale. En ce sens, nous cherchons à :

1. Savoir quelles questions-réponses dans un entretien ont un rôle plus important dans la prédiction (7.2)
2. Comprendre, en étudiant l'attention temporelle au niveau des réponses, ce qui rend certains moments plus importants que d'autres au sein de la réponse (Section 7.3)

à 7.5)

3. Comparer ces moments à des moments aléatoires pour déterminer si ces moments sont effectivement pertinents pour la tâche de convocabilité (Section 7.6).

Publications associées à ce chapitre :

- [Hemamou et al., 2019a] Hemamou Léo, Felhi Ghazi, Martin Jean-claude, Clavel Chloé. Slices of Attention in Asynchronous Video Job Interviews. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)(sept. 2019)
- [Hemamou et al., 2020] Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Attention Slices dans les Entretiens d 'Embauche Vidéo Différés. Workshop sur les "Affects, Compagnons Artificiels et Interactions" (ACAI)(2020)
- En soumission, Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Multimodal Hierarchical Attention Neural Network : Looking for Candidates Behaviour which Impact Recruiter's Decision

7.1 Définition des courbes d'attention utilisées

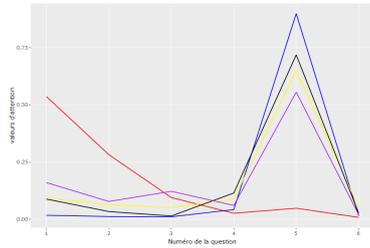
Les courbes d'attention désignent la séquence des valeurs d'attention temporelle α_t et α_i respectivement définie par (équation 6.6 p. 82) et (équation 6.10 p. 83).

Les réseaux neuronaux sont soumis à diverses sources de variabilité, telles que l'initialisation des poids aléatoires, la descente du gradient stochastique ou le dropout. Ainsi, dans ce qui suit, nous calculons la moyenne des courbes d'attention pour chacun des réseaux HireNet monomodaux et multimodaux sur cinq instances entraînées.

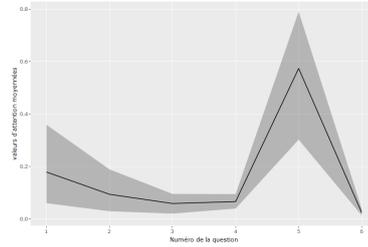
De cette façon, nous visons à saisir le comportement plus général des mécanismes d'attention [Jetley et al., 2018].

Un exemple des courbes originales d'attention et de la courbe moyennée résultante que l'on utilise par la suite est disponible en figure 7.2 pour l'attention au niveau des réponses et figure 7.1 pour l'attention au niveau de l'entretien.

La suite de ce chapitre est organisée sous forme de questions, auxquelles nous nous efforçons de répondre au travers d'expériences.

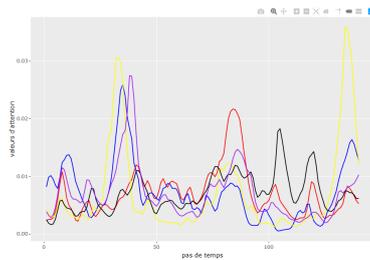


(a) Exemple de courbes d'attention temporelle au niveau de l'entretien pour 5 instances de HireNet.

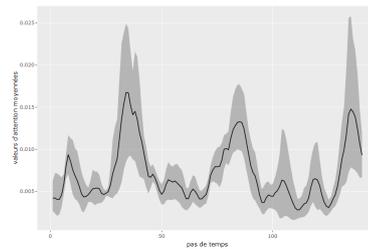


(b) Courbe d'attention moyennée résultante. La partie grisée correspond à deux fois l'écart type.

FIGURE 7.1 – Exemple de courbes d'attention temporelle au niveau de l'entretien et la courbe d'attention moyennée résultante.



(a) Exemple de courbes d'attention temporelle au niveau de la réponse pour 5 instances de HireNet.



(b) Courbe d'attention moyennée résultante. La partie grisée correspond à deux fois l'écart type.

FIGURE 7.2 – Exemple de courbes d'attention temporelle au niveau de la réponse et la courbe d'attention moyennée résultante.

7.2 Comment les valeurs d'attention sont-elles distribuées au niveau de l'entretien ?

Notre objectif est d'explorer l'attention attribuée aux différentes questions-réponses au cours d'un même entretien. Une question spécifique demeure s'il existe un schéma temporel général d'importance parmi les questions (*e.g.* est ce que la première ou la dernière question sont plus importantes?) [Swider et al., 2016, Naim et al., 2018, Nguyen and Gatica-Perez, 2015].

Méthode : Statistiques sur l'attention temporelle au niveau de l'entretien À cet égard, nous examinons les valeurs de l'attention temporelle au niveau de l'entretien α_i définies dans l'équation 6.10. Lorsque le nombre de questions change d'un poste à l'autre, nous multiplions la valeur d'attention de la i -ième question par le nombre total de questions de l'entretien (*i.e.* $\alpha_i * n$) afin d'obtenir l'attention relative de la question i par rapport à la durée de l'entretien. Ensuite, nous regroupons les différentes questions en fonction de leur ordre au cours de l'entretien d'embauche (*i.e.* $1/5, 2/5 \dots 5/5$ parties de l'entretien). De cette façon nous pouvons investiguer si un effet sur les valeurs d'attention est visible par rapport à la temporalité de l'entretien.

Résultats et discussion La moyenne des valeurs d'attention associées à chaque partie de l'entretien est affichée sur la figure 7.3 pour chaque modalité. Comme nous pouvons l'observer, la première partie de l'entretien est plus importante que les autres parties de l'entretien pour les modalités Langage, Vidéo et Multimodal. Cette constatation est cohérente avec [Naim et al., 2018]. Elle pourrait s'expliquer par l'effet de premières impressions [Swider et al., 2016], la présence des questions les plus importantes au début de l'entretien d'embauche (*e.g.* "Pouvez-vous vous décrire?") ou le fait que les recruteurs ne regardent potentiellement que les premières réponses [Torres and Mejia, 2017]. La modalité audio ne suit pas ce schéma, ce qui pourrait s'expliquer par une prosodie constante des candidats dans toutes les parties de l'entretien.

7.3 A quoi correspondent les pics d'attention temporelle au sein d'une réponse ?

Les courbes d'attention temporelle au niveau de la réponse consistent principalement en des séries temporelles bruitées avec des pics de valeur élevée [Yu et al., 2017]. Un exemple typique est présenté en figure 7.4. Nous cherchons à comprendre ce qui se passe pendant ces pics et à les caractériser. Nous commençons par extraire pour chaque question, et chaque

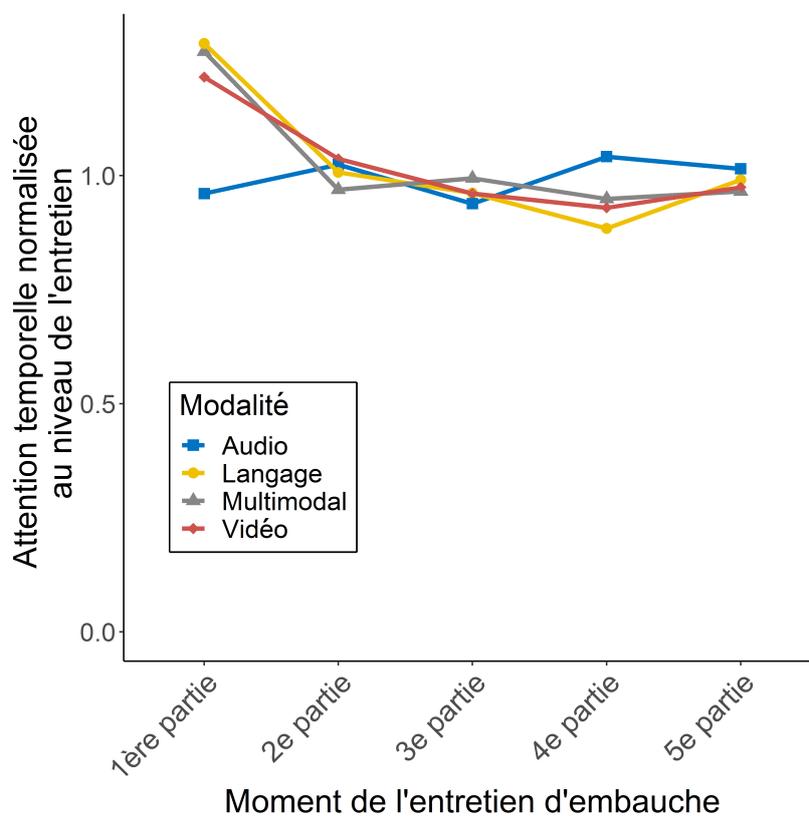


FIGURE 7.3 – Moyenne des scores d'attention temporelle normalisés au niveau de l'entretien, regroupés selon l'ordre des questions.

7.3. A QUOI CORRESPONDENT LES PICS D'ATTENTION TEMPORELLE AU SEIN D'UNE RÉPONSE ?

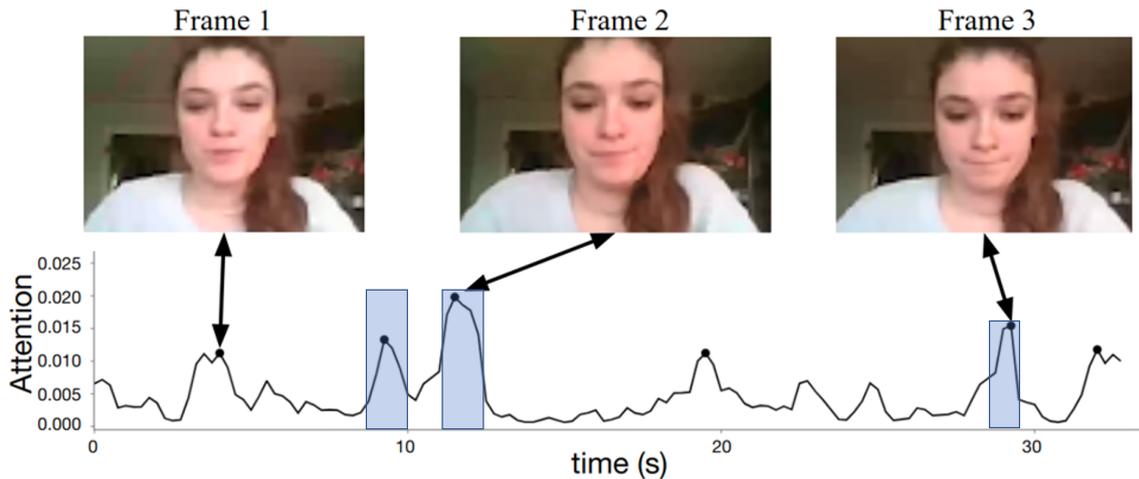


FIGURE 7.4 – Un exemple de courbe d’attention temporelle au niveau de la réponse et quelques moments saillants. Les images 2 et 3 sont issues des tranches d’attention. Les indices comportementaux de l’anxiété (étirement et resserrement des lèvres) semblent être considérés comme importants par le mécanisme de l’attention.

modalité, ces pics d’attention dans la séquence de α_t définie dans l’équation (6.6). Pour le reste de l’article, nous définissons une *tranche d’attention* comme un moment de la vidéo comprenant un pic d’attention.

Méthodologie : Extraction de tranches d’attention en détectant les valeurs aberrantes de manière non supervisée. Nous utilisons l’algorithme DBSCAN [Ester et al., 1996], un algorithme de partitionnement par densité qui nous permet de sélectionner les régions d’attention maximale, illustrées par les cases bleues de la figure 7.4. Cette méthode s’est avérée efficace, car elle gère les valeurs bruitées des courbes d’attention et permet de ne spécifier ni le nombre ni l’étendue des régions (la durée dans notre cas particulier de la série temporelle) à extraire. Nous appliquons DBSCAN à la série temporelle en pondérant chacun des points avec leur valeur d’attention. Si certaines courbes d’attention ne présentent pas de pics d’attention, aucune tranche d’attention n’est extraite de celles-ci.

Résultats et statistiques descriptives à propos des pics extraits. Le tableau 7.2 fournit un résumé des données utilisées comme base de notre étude en termes de pics d’attention détectés et quelques statistiques descriptives. Tout d’abord, il est intéressant de noter que la durée des tranches d’attention extraites par le mécanisme d’attention pour les modalités audio et vidéo se situe principalement entre 0,5 s et 2,5 s. Cette durée est typique de la durée des expressions faciales ou de phénomènes de disfluences [Varni et al., 2018], alors que les tranches d’attention multimodales sont généralement plus longues, avec des durées comprises entre 0,9 s et 3,3 s.

La figure 7.5 montre la répartition des tranches d'attention en fonction de la longueur des réponses. Les tranches d'attention pour les modalités audio et vidéo apparaissent plus souvent au début et à la fin d'une réponse. Cela pourrait indiquer que les comportements non verbaux se produisant au début et à la fin de la réponse ont un impact significatif sur la décision de convocabilité. D'une façon similaire, durant les interactions sociales face-à-face certains comportements non verbaux sont des marqueurs singuliers du début et de la fin d'un tour de parole [Cassell et al., 2001, Goodrich, 1979]. Même si l'effet est moins prononcé, les tranches d'attention multimodales suivent le même schéma que les tranches d'attention audio et vidéo.

Nous cherchons maintenant à savoir si les tranches d'attention multimodales coïncident ou non avec des tranches monomodales extraites précédemment. Dans ce sens, nous calculons le coefficient de similarité Jaccard (défini pour deux ensembles A et B comme le rapport $\frac{|A \cap B|}{|A \cup B|}$) entre les tranches d'attention des différentes modalités (Tableau 7.1). Tout d'abord, on peut observer qu'entre les tranches monomodales, peu de tranches sont partagées, comme le montre la faible valeur du plus grand coefficient (0,07 entre l'audio et la vidéo). La plupart des tranches multimodales ne se chevauchent pas avec les tranches monomodales, ce qui suggère la capacité de notre système à détecter les signaux multimodaux influents. Seule une petite partie des tranches multimodales est partagée avec les modalités du langage (coefficient Jaccard de 0,18) et audio (coefficient Jaccard de 0,12).

Mesure	Similarité Jaccard			
	1.	2.	3.	4.
Modalité				
1. Langage	1			
2. Audio	0.024	1		
3. Vidéo	0.028	0.067	1	
4. Multimodalité	0.176	0.118	0.042	1

TABLE 7.1 – Similarité de Jaccard entre les tranches d'attention de différentes modalités.

7.3. A QUOI CORRESPONDENT LES PICS D'ATTENTION TEMPORELLE AU SEIN D'UNE RÉPONSE ?

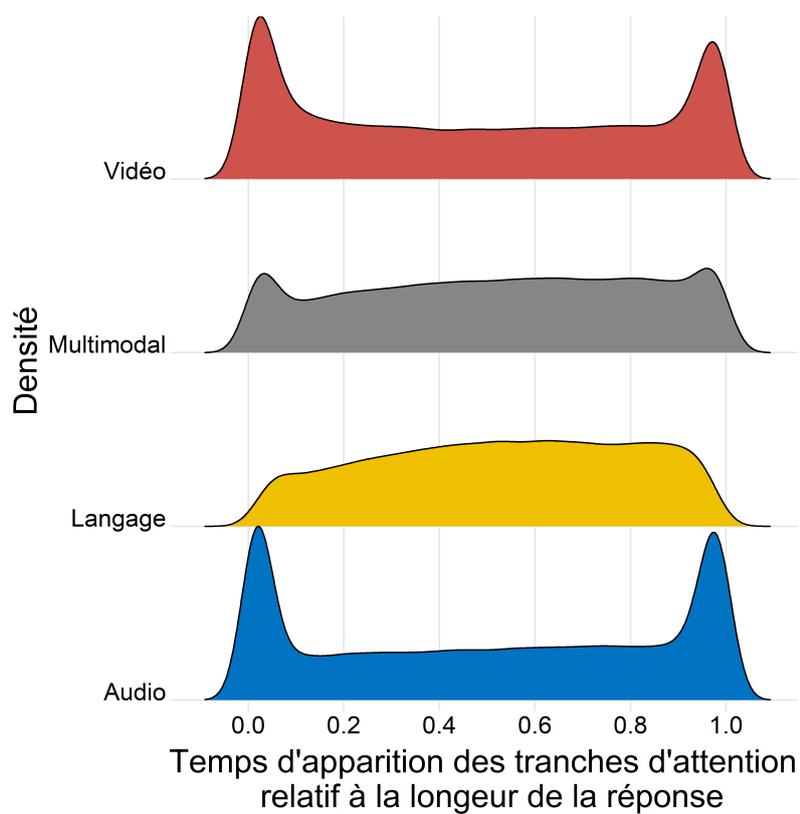


FIGURE 7.5 – Densité du nombre de tranches d'attention en fonction de leur moment d'apparition relatif à la longueur totale de la réponse.

7.4. COMMENT LES MODALITÉS SONT-ELLES FUSIONNÉES AU SEIN DU GMU PENDANT LES TRANCHES D'ATTENTION ?

Modalité	Audio			Langage			Vidéo			Multimodal		
Nombre de tranches d'attention extraites	62400			65056			45636			104821		
Moyenne de tranches d'attention par question	3.20			2.61			2.23			3.92		
Moyenne de tranches d'attention par entretien	12.63			12.86			9.54			20.36		
	D1	Moyenne	D10	D1	Moyenne	D10	D1	Moyenne	D10	D1	Moyenne	D10
Durée des tranches d'attention	0.5s	1.2s	2s	2 mots	5.8 mots	10 mots	0.5s	1.45s	2.5s	0.9s	1.81s	3.3s

TABLE 7.2 – Statistiques descriptives des tranches d'attention extraites.

D_1 et D_{10} correspondent au premier et au dernier déciles.

7.4 Comment les modalités sont-elles fusionnées au sein du GMU pendant les tranches d'attention ?

Dans le contexte multimodal, afin de comprendre quelle modalité contribue le plus aux tranches d'attention, nous étudions le mécanisme de fusion multimodale. Pour rappel, nous avons fait le choix d'utiliser un mécanisme de fusion interprétable qui consiste à projeter toutes les modalités d'un espace joint et d'inférer automatiquement la contribution de chacune des modalités (voir equation 6.18).

Méthode : Statistiques sur les valeurs de porte des unités multimodales Les modalités sont fusionnées par le biais du GMU et une représentation multimodale $z_t^{multimodal}$ est calculée selon l'équation (6.18), comme une somme pondérée de chaque modalité. Comme le modèle est assez simple, nous pouvons étudier la norme du vecteur $\sigma_t^m * h_t^m$ pour chaque modalité m pendant les tranches d'attention, ce qui reflète le degré de contribution de chaque modalité à la représentation multimodale pour ces tranches.

Résultats La figure 7.6 affiche les boîtes à moustaches de la contribution de chaque modalité pour constituer le vecteur multimodal. La valeur moyenne des valeurs de porte pour l'audio est supérieure à celle des modalités linguistiques et vidéo, qui présentent à leurs tours des valeurs extrêmes plus élevées. Nous interprétons cela comme une preuve que la modalité audio contribue davantage à la sélection de tranches d'attention élevées, alors que les modalités langue et vidéo sont plus ponctuelles.

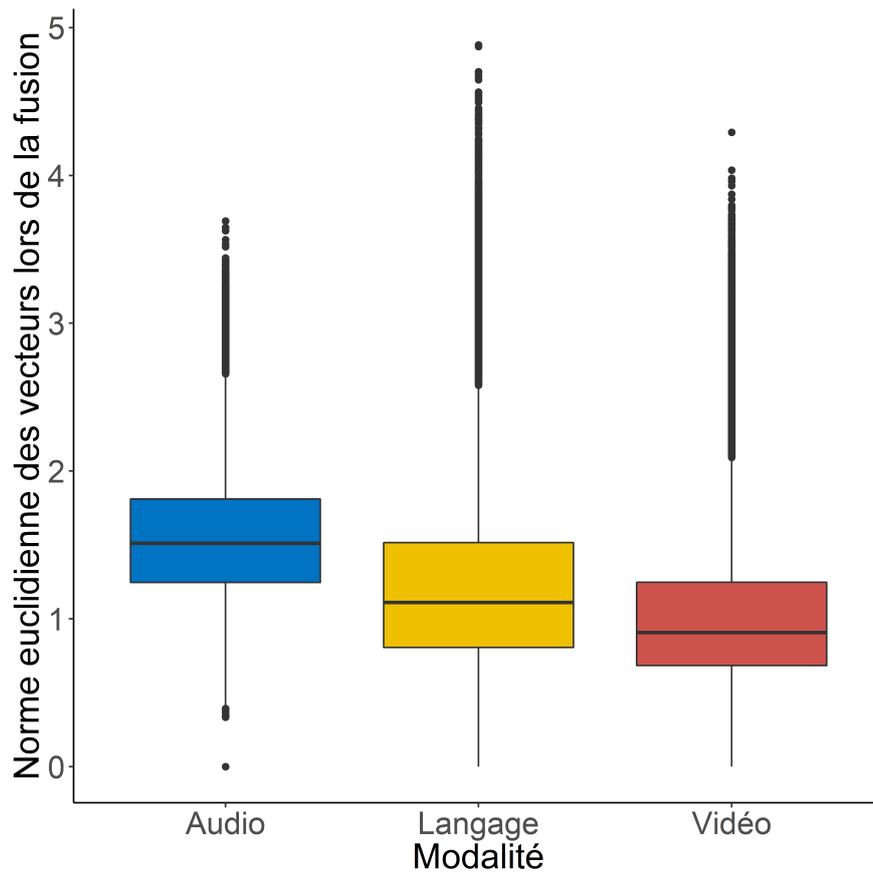


FIGURE 7.6 – Boîtes à moustaches de la norme du vecteur $\sigma_t^m * h_t^m$ pour chaque modalité m (décrite en équation 6.18) pendant les tranches d'attention. Ceci reflète le degré de contribution de chaque modalité à la représentation multimodale pour ces tranches.

7.5 Le contenu des tranches d'attention est-il différent de celui des tranches aléatoires ?

L'attention à un pas de temps t est calculée à la suite d'un traitement par des Bi-GRU selon l'équation 6.6 et ne dépend pas directement des descripteurs d'entrée. Les pics d'attention peuvent fortement dépendre du contexte et de la mémoire des cellules GRU et très peu des descripteurs d'entrée au moment de leur apparition. Pour que l'attention soit utile en tant qu'outil d'interprétation, il est nécessaire que les tranches d'attention soient également pertinentes du point de vue de leur contenu. En ce sens, nous vérifions que les tranches d'attention sont identifiables par leur contenu en terme des descripteurs d'entrée. Nous mettons également en évidence les descripteurs qui contribuent le plus à la tâche d'identification des tranches d'attention, ce qui nous aidera à mieux comprendre ce qui différencie ces tranches d'attention des tranches aléatoires.

Méthode : Classification supervisée entre les tranches d'attention et les tranches aléatoires. Pour s'assurer que les pics d'attention proviennent principalement de ce qui se passe dans les tranches de temps concernées, nous construisons une tâche de classification binaire. Cette tâche consiste à distinguer les moments de forte attention des autres moments aléatoires en se basant sur les comportements verbaux et non verbaux qui se produisent au cours de ceux-ci. Ainsi, pour notre tâche, nous prenons comme étiquette positive les tranches d'attention extraites dans chacune des réponses du candidat. Comme étiquette négative, pour chaque tranche d'attention détectée, nous échantillonnons quatre moments dans la réponse du candidat avec la même durée que la tranche d'attention précédemment sélectionnée (voir figure 7.7). Nous échantillonnons ces moments selon une distribution proportionnelle à $1 - \alpha_t$, où t désigne le pas de temps de la trame. Grâce à cet échantillonnage, nous visons à sélectionner des moments d'importance variable, et pas seulement les moins ou les plus importants.

Notre objectif est de comprendre si les tranches d'attention sont différentes et en quoi elles diffèrent. Nous avons donc décidé d'utiliser des classifieurs classiques avec une méthodologie bien établie pour analyser l'importance des descripteurs à savoir Ridge (modèle linéaire et transparent) et XGBoost [Chen and Guestrin, 2016], une méthode d'ensemble d'arbres de décision (modèle non linéaire).

Représentations utilisées pour la classification. Comme ces classifieurs prennent en entrée un vecteur fixe, nous construisons pour chaque modalité un ensemble de descripteurs interprétables.

Pour les descripteurs audio et vidéo, nous utilisons les descripteurs prosodiques et d'ex-

7.5. LE CONTENU DES TRANCHES D'ATTENTION EST-IL DIFFÉRENT DE CELUI DES TRANCHES ALÉATOIRES ?

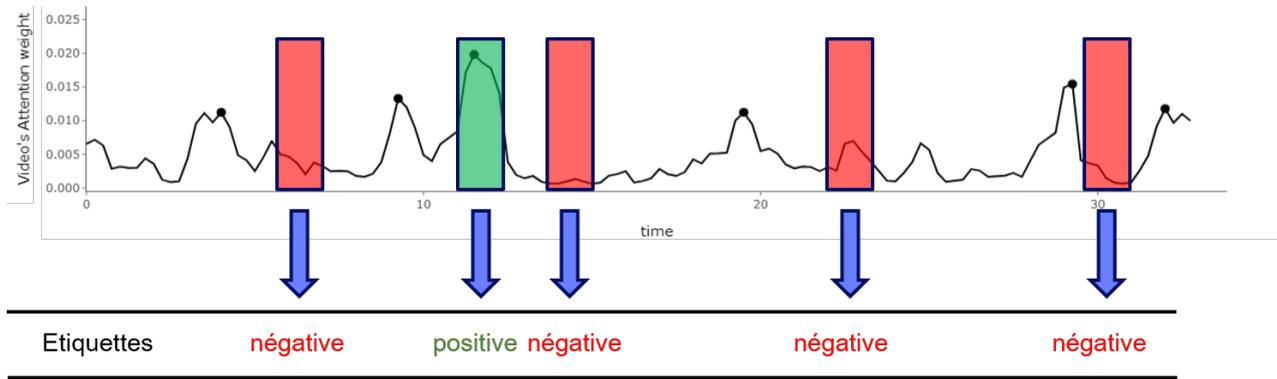


FIGURE 7.7 – Processus d'échantillonnage de moments aléatoires et d'étiquetages par rapport à un pic d'attention. Pour chaque tranche d'attention extraite, nous échantillonnons quatre tranches aléatoires de même durée dans la réponse du candidat.

pressions faciales (déjà interprétables) initialement utilisés pour entraîner HireNet. Nous effectuons un pré-traitement des descripteurs en appliquant une normalisation Z intra-réponse sur toute la séquence [Nguyen and Gatica-Perez, 2015]. Puis nous intégrons les descripteurs sur les fenêtres temporelles des moments choisis en utilisant les fonctions suivantes : moyenne, moyenne des gradients positifs et moyenne des gradients négatifs.

Pour les descripteurs de la modalité langage, afin d'explorer le contenu des tranches d'attention et d'analyser leur spécificité, nous utilisons plusieurs représentations en gardant à l'esprit à la fois l'interprétabilité et la performance, elles sont au nombre de trois.

— *Représentation BERT moyennée sur les tranches.*

Premièrement nous utilisons la représentation originale de BERT moyennée sur les tranches, c'est la représentation la plus proche des descripteurs originaux utilisés par le modèle HireNet. Elle demeure la représentation comportant le plus d'information mais consiste en la représentation la moins interprétable.

— *Représentation par descripteurs experts.*

Nous construisons une représentation experte en extrayant des descripteurs issus de dictionnaires ou d'outils externes. Cette représentation fournit un jeu de descripteurs entièrement interprétables.

Plus précisément, nous extrayons le style linguistique du candidat [Mairesse et al., 2007] en extrayant l'histogramme des catégories grammaticales utilisées (par exemple verbes, noms ou pronoms personnels) en utilisant l'outil TreeTagger [Schmid, 1994]. Ensuite, nous utilisons le dictionnaire LIWC [Piolat et al., 2011], qui classe les mots selon 58 catégories de haut niveau telles que "travail" ou "accomplissement", et nous calculons l'histogramme de leur utilisation dans les tranches. Ceci nous permettra

de savoir si certaines catégories sont plus influentes par rapport à d'autres sur le mécanisme d'attention.

Troisièmement, nous utilisons le dictionnaire FEEL [Poncelet, 2016]. Ce dictionnaire classe les mots selon les 6 émotions de base définies par Ekman. Ce dictionnaire pourrait potentiellement nous informer sur l'utilité des émotions dans le contenu verbal lors de l'entretien d'embauche.

Quatrièmement, nous utilisons le dictionnaire LEXIQUE3 [New, 2006]. Ce dictionnaire fournit des informations sur la fréquence de l'utilisation des mots dans la langue française, comme la fréquence du lemme dans un énorme corpus de livres et de films, selon 42 caractéristiques. Ainsi, nous essayons de savoir si l'utilisation de mots plus rares dans la langue française a un effet sur le mécanisme d'attention. Dans l'ensemble, cette représentation donne un vecteur de 116-d.

— *Représentation par BOW BERT.*

Enfin, à mi-chemin entre l'interprétabilité et la représentation de la boîte noire, nous calculons une représentation BOW des précédents vecteurs BERT, calculée à l'aide de la boîte à outils OpenXBOW [Schmitt and Schuller, 2017] sur 300 dimensions. Plus précisément, nous construisons un dictionnaire grâce à une méthode non supervisée (K-means) pour regrouper des vecteurs BERT similaires dans une même catégorie. Nous affectons chaque mot de chaque tranche d'attention aux classes de ce dictionnaire, puis nous calculons la représentation BOW sur chaque tranche. De cette façon, nous pouvons examiner si une classe de mots (une classe du dictionnaire) est caractéristique des tranches d'attention. De cette manière, nous pouvons ensuite visualiser les mots regroupés dans les classes influentes identifiées.

Pour la représentation multimodale, nous concaténons les caractéristiques monomodales précédentes à la manière d'une fusion précoce. Pour cette expérience, nous conservons les mêmes ensembles d'entraînement, de développement et de test que dans la section 6.5.1 pour éviter toute fuite de données.

Résultats et analyse des descripteurs importants pour la tâche d'identification Le tableau 7.3 présente les résultats de cette expérience, en indiquant les scores F1 de la classe positive et négative, et l'AUC pour chacun des classifieurs (Ridge et XGBoost) et chacune des modalités. A noter que, par construction, la tâche de classification est non balancée avec une répartition de 80% d'étiquettes négatives et 20% d'étiquettes positives. Dans l'ensemble, les deux classifieurs présentent de bons résultats pour les tâches monomodales : nous interprétons les scores F1 positifs élevés et l'AUC très élevée comme preuve que, malgré l'influence de la modélisation des séquences et l'utilisation d'informations contex-

7.5. LE CONTENU DES TRANCHES D'ATTENTION EST-IL DIFFÉRENT DE CELUI DES TRANCHES ALÉATOIRES ?

Modalité	Représentation	Ridge			XGBoost		
		F1 \oplus	F1 \ominus	AUC	F1 \oplus	F1 \ominus	AUC
Audio		0.779	0.945	0.961	0.829	0.955	0.973
Video		0.558	0.913	0.873	0.620	0.917	0.906
Langage	Dictionnaires	0.353	0.614	0.720	0.175	0.889	0.755
	BERT Moyenné	0.628	0.917	0.923	0.585	0.909	0.903
	BOW BERT	0.308	0.893	0.823	0.288	0.892	0.808
Multimodal	Dictionnaires	0.113	0.889	0.679	0.930	0.982	0.995
	BERT Moyenné	0.107	0.886	0.705	0.870	0.969	0.986
	BOW BERT	0.047	0.886	0.682	0.866	0.968	0.985

TABLE 7.3 – Résultats de classification pour la tâche d'identification des tranches d'attention et aléatoires.

tuelles, l'importance d'un moment est toujours principalement définie par les événements qui s'y produisent. Ainsi, les classifieurs peuvent discriminer les moments précis où des pics d'attention se produisent des moments aléatoires de la même réponse. Les résultats sont particulièrement élevés pour l'audio, que nous interprétons comme une différence de contenu plus prononcée entre les tranches aléatoires et les tranches d'attention pour cette modalité. Pour l'entrée multimodale, le classifieur Ridge présente de mauvaises performances par rapport à XGBoost, comme le montrent les scores positifs F1 et AUC beaucoup plus faibles. Ce résultat pourrait potentiellement s'expliquer par les interactions entre modalités dans les tranches d'attention, que le classifieur linéaire ne peut pas saisir.

Analyse des descripteurs importants. Afin de mettre en évidence les descripteurs qui contribuent le plus à l'identification des tranches d'attention, nous utilisons les classifieurs entraînés et nous inspectons les descripteurs les plus discriminants, que nous répertorions dans le tableau 7.4.

Pour les expressions faciales, le relèvement de la partie externe des sourcils (AU02), le relèvement du menton (AU17), la chute de la mâchoire (AU26), la rotation de la tête (Rx et Rz) sont classés comme importants par les deux classifieurs. De plus, ils présentent une taille d'effet modérée selon le test de Cohen. En se basant sur le signe des coefficients du modèle Ridge, on remarque que les tranches d'attention sont principalement induites par : 1) sourcils externes relevés, 2) la non-activation de la chute de la mâchoire, 3) l'activation du relèvement du menton, et 4) l'inclinaison de la tête vers l'avant. Une interprétation possible est que les signes de confusion (menton levé et silence) lors d'un entretien d'embauche par vidéo ont un impact sur la perception du recruteur [Cahya et al., 2019]. Une deuxième

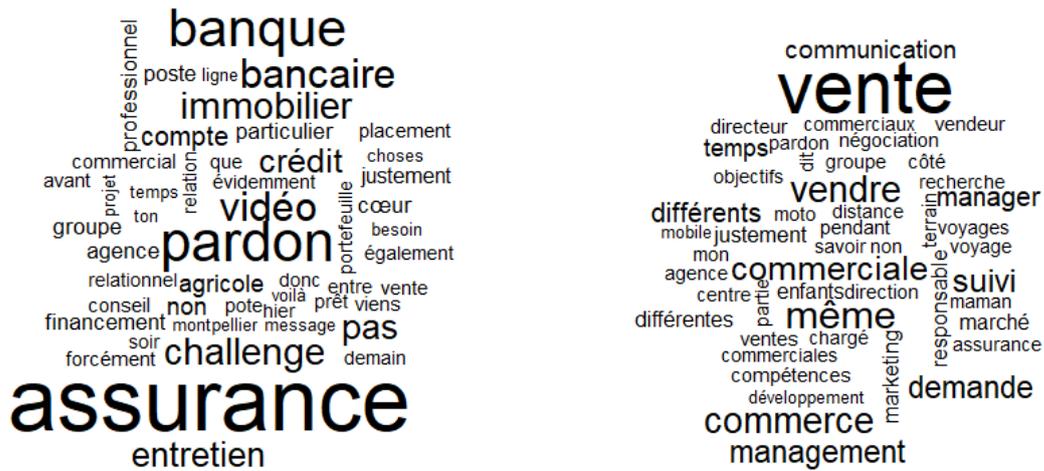
interprétation possible est que les signes d'insistance (sourcil relevé et tête fléchie vers l'avant) ont également un impact sur la décision du recruteur [McClave, 2000].

Pour les indices audio, l'intensité sonore, le flux spectral et l'amplitude de f1 sont considérés comme importants et leurs signes de coefficient (négatifs) dans le modèle Ridge montrent que les silences sont importants. De même, le rapport de bruit harmonique, `logrelf0.h1.a3` sont classés comme important, ce qui conduit à l'hypothèse que la respiration et la tension de la voix sont importantes (voir l'analyse des descripteurs de [Clavel and Richard, 2011]) dans le contexte d'un entretien vidéo asynchrone. Après l'écoute des tranches d'attention, une analyse qualitative met en évidence les silences, les disfluences (par exemple les mots remplisseurs tels que « euh ») et la respiration. Cette écoute qualitative soutient l'analyse quantitative de l'importance des descripteurs.

Pour les indices linguistiques, l'analyse de l'importance des descripteurs de la représentation moyennée de BERT n'est pas explorée, car les dimensions de BERT sont encore difficilement compréhensibles [Shin et al., 2018]. Pour la représentation basée sur les dictionnaires, seul un faible effet se produit pour la fréquence lexicale dans la langue française (fréquence dans les corpus de films et de livres), les pronoms et les pronoms personnels. Pour la représentation BERT BOW, les dimensions `v100`, `v29`, `v76` et `v299` sont considérées comme importantes pour les deux algorithmes. Nous pouvons analyser le contenu de chaque dimension en examinant les mots contenus dans ces catégories. Un nuage de mots de chacune de ces dimensions est affiché figure 7.8. `v100` semble être lié au secteur (banque, assurance, immobilier), `v76` semble être lié aux compétences (vente, communication, gestion). Ces deux caractéristiques sont liées aux expériences antérieures et à l'auto présentation des compétences des candidats. Les coefficients négatifs (`v29` et `v299`) contiennent principalement des mots fréquents (moi, être, penser, etc.), ce qui souligne l'hypothèse selon laquelle les mots communs n'ont pas une grande influence (voir figure figure 7.9).

Pour les indices multimodaux, le tableau 7.6 indique les 15 caractéristiques les plus importantes prises en compte par le classifieur XGBoost. Ces caractéristiques sont uniquement issues des modalités audio et linguistiques. L'audio et la langue semblent être les éléments qui contribuent le plus aux tranches d'attention multimodales, ce qui est en cohérence avec les résultats précédents de la Table 7.1 et de la figure 7.6. Il est intéressant de noter qu'aucun descripteur n'a une taille d'effet d supérieure à 0,5. Cette constatation renforce l'idée que l'attention et les tranches aléatoires ne sont pas séparables de manière linéaire dans le cadre multimodal. Cela corrobore les résultats précédents, où le classifieur linéaire (Ridge) ne pouvait pas faire la distinction entre les tranches d'attention et les tranches aléatoires par rapport à la non-linéarité (XGBoost). De la même manière que l'analyse monomodale, nous reportons en appendices les nuages de mots associés aux dimensions textuelles.

7.5. LE CONTENU DES TRANCHES D'ATTENTION EST-IL DIFFÉRENT DE CELUI DES TRANCHES ALÉATOIRES ?



(a) Nuage de mots de la dimension v100 (b) Nuage de mots de la dimension v76

FIGURE 7.8 – Nuage de mots des dimensions positivement associés avec les tranches d’attention.



(a) Nuage de mots de la dimension v29 (b) Nuage de mots de la dimension v299

FIGURE 7.9 – Nuage de mots des dimensions négativement associés avec les tranches d’attention.

7.6 Est-ce que les tranches d'attention contiennent plus d'information pour inférer la convocabilité que des tranches aléatoires ?

Nous avons établi que notre modèle pouvait mettre en évidence un certain nombre de moments dans la réponse du candidat et avons constaté que les *tranches d'attention* étaient différentes des autres tranches de la réponse. Nous étudions maintenant leur utilité pour la tâche de prédiction de la convocabilité, en vérifiant que les moments contenus dans ces *tranches d'attention* ont un contenu prédictif plus élevé que les tranches aléatoires.

Méthode : Classification supervisée concernant la convocabilité en utilisant des tranches fines aléatoires ou des tranches d'attention. En utilisant les mêmes descripteurs que dans la section 7.5, nous construisons une tâche de classification basée sur une fenêtre de temps minimale dans la réponse du candidat. Pour ce faire, nous constituons deux sous-ensembles de l'ensemble de données précédent : le premier ne contient que les *tranches d'attention* ; le second est constitué des tranches aléatoires sélectionnées dans la Section 7.5.

Pour chaque sous-ensemble, nous effectuons la procédure de bootstrap suivante : nous constituons 100 nouveaux jeu de données d'apprentissage, chacun étant un échantillonnage composé d'une seule *tranche d'attention* ou d'une seule tranche aléatoire par candidat respectivement pour le premier et le deuxième sous-ensemble.

Nous entraînons les classifieurs Ridge et XGBoost sur ces ensembles de données et les testons sur deux sous-ensembles d'une seule *tranche d'attention* ou d'une seule tranche aléatoire par candidat pour le premier et le deuxième sous-ensemble respectivement. Notez que nous suivons toujours la même répartition d'entraînement, de développement et de test que dans la section 6.5.1 pour éviter toute fuite de données.

Résultats et discussions. Le tableau 7.5 présente les résultats. Pour chaque modalité, nous indiquons la moyenne et l'écart-type de l'AUC des deux classifieurs, chacun étant entraîné et évalué respectivement sur les *tranches* aléatoires et sur les *tranches d'attention*. Nous observons une AUC moyenne significativement plus élevée pour les *tranches d'attention* que pour les tranches aléatoires, pour les modalités audio, linguistiques et multimodales des deux classifieurs ($p\text{-values} < 0,01$). La comparaison des moyennes pour la modalité vidéo ne montre pas de différence significative de performance entre les deux situations. Nous concluons que les *tranches d'attention* contiennent effectivement plus d'informations que les tranches aléatoires pour prédire la convocabilité du candidat pour les modalités audio, linguistiques et multimodales.

7.7 Conclusion

En ce qui concerne l'interprétabilité, nous avons proposé une étude approfondie sur les courbes d'attention. Plus précisément, nous considérons les pics de valeurs d'attention, qui se produisent sur de courtes durées. Nous définissons ces pics comme des tranches d'attention, et : 1) nous montrons qu'elles sont systématiquement différentes des tranches aléatoires et qu'elles sont identifiables par leur contenu ; 2) nous caractérisons les tranches d'attention par une analyse d'importance des descripteurs ; 3) nous étudions les valeurs prédictives de ces tranches en ce qui concerne l'inférence de la convocabilité. De plus, nos résultats semblent indiquer que les tranches d'attention se produisent plus souvent au début et à la fin d'une réponse ; que les tranches multimodales ne chevauchent que légèrement les tranches monomodales ; et que les tranches d'attention multimodales sont meilleures que les tranches aléatoires pour prédire la convocabilité. De manière générale, nous avons développé une méthodologie pour la détection et l'interprétation d'instantanés clés dans des vidéos monologues d'entretien d'embauche.

Groupe de descripteurs	Ridge		XGBoost
	Coefficients positifs	Coefficients négatifs	
Bas du visage	AU17 ³ , AU14 ⁹	AU10 ² , AU26 ⁷ , AU20 ⁸ , AU28 ¹⁰	AU17 ⁶ , AU26 ⁷ , AU10 ⁸ , AU25 ¹⁰
Haut du visage	AU02 ⁵	AU04 ⁴	AU02 ¹ , AU04 ⁹
Tête	R_z ¹ , R_x ⁶		R_x ³ , R_z ⁴ ,
Confidence d'OpenFace			↑ success ² , success ⁵ ,
Cepstral		mfcc2 ² , mfcc3 ⁸	mfcc2 ⁵ , ↑ mfcc2 ¹⁰
Spectral	f1bandwidth ⁹	spectralflux ¹ , f1amplitude ⁷ , f3frequency ¹⁰	f1amplitude ² , spectralflux ⁴ , f3amplitude ⁷ , ↑ spectralflux ⁸ , f1bandwidth ⁹
Prosodie		Intensité ⁵	Intensité ⁶
Qualité de la voix	h1.a ⁴ , slope500.1500 ⁵ , h1.h ⁶	hnr ³	hnr ¹ , slope500.1500 ³
LIWC			I ² , present ⁵ , future ¹⁰
LEXIQUE3	freqlemfilms_sd ² , freqlemlivres_mean ³ , freqlemlivres_max ¹⁰	freqlemfilms_mean ¹ , freqlemlivres_sd ⁴ , freqlemfilms_max ⁸	freqlemfilms_max ³ , nbsyll_max ⁴ , freqlemfilms_q25 ⁶ , nbhomoph_max ⁹
POS		auxverb ⁵ , verb ⁶ , pron ⁷ , article ⁹	pron ¹ , auxverb ⁷ , conj ⁸
BOW BERT	v100 ¹ , v76 ² , v188 ⁶ , v102 ⁷	v29 ³ , v299 ⁴ , v242 ⁵ , v123 ⁸ , v262 ⁹ , v298 ¹⁰	v100 ¹ , v29 ² , v11 ³ , v191 ⁴ , v123 ⁵ , v147 ⁶ , v299 ⁷ , v279 ⁸ , v76 ⁹ , v286 ¹⁰

TABLE 7.4 – Analyse de l'importance des descripteurs pour les modèles monomodaux. F^i dénote le descripteur F classé en i -ème position. ↑ et ↓ dénotent respectivement la moyenne des gradients positifs et négatifs. Les descripteurs avec un important coefficient positif (ou respectivement négatif) pour le modèle Ridge ont plus de chance d'apparaître (ou respectivement ne pas apparaître) dans les tranches d'attentions. Les descripteurs en gras F dénotent les descripteurs avec au minimum un effet moyen d de test de Cohen ($d > 0.5$).

7.7. CONCLUSION

Descripteurs	Représentation	Ridge		XGBoost	
		Tranche aléatoire	<i>Tranche d'attention</i>	Tranche aléatoire	<i>Tranche d'attention</i>
Audio		0.509 ± 0.024	0.544 ± 0.015	0.514 ± 0.022	0.565 ± 0.023
Vidéo		0.514 ± 0.018	0.516 ± 0.030	0.515 ± 0.019	0.518 ± 0.022
Langage	Dictionnaires	0.525 ± 0.017	0.529 ± 0.019	0.517 ± 0.019	0.524 ± 0.023
	BERT moyenné	0.535 ± 0.021	0.546 ± 0.019	0.546 ± 0.018	0.553 ± 0.017
	BOW BERT	0.523 ± 0.020	0.522 ± 0.021	0.524 ± 0.022	0.521 ± 0.016
Multimodal	Dictionnaires	0.529 ± 0.017	0.588 ± 0.021	0.538 ± 0.020	0.603 ± 0.018
	BERT moyenné	0.545 ± 0.018	0.580 ± 0.023	0.549 ± 0.020	0.606 ± 0.018
	BOW BERT	0.527 ± 0.020	0.582 ± 0.017	0.537 ± 0.020	0.604 ± 0.018

TABLE 7.5 – Résultats pour la tâche de l’employabilité en utilisant un jeu de données comprenant des tranches aléatoires ou un jeu de données utilisant des tranches d’attention.

Groupes de descripteurs	XGBoost
Spectral	f3amplitude, ↑ f1frequency, ↓ f1amplitude, ↓ f1frequency, ↓ f1bandwidth, ↑ hnr
BOW BERT	v82, v191, v248, v164, v94, v202, v39, v165, v146

TABLE 7.6 – Analyse d’importance des descripteurs multimodaux. ↑ and ↓ dénotent respectivement la moyenne des gradients positifs et négatifs.

Section 8

Équité individuelle pour la convocabilité

L'équité et l'absence de biais des algorithmes sont des éléments clés pour l'éthique des algorithmes en apprentissage automatique.

De nombreux acteurs industriels commercialisent déjà des outils d'évaluation automatique dans le domaine de l'analyse automatique d'EVD, sans fournir pour autant des éléments concrets et techniques garantissant des garde-fous au regard des potentiels biais [Raji et al., 2020, Raghavan et al., 2019].

Ainsi, il est important d'auditer de telles solutions afin de vérifier qu'elles ne font pas leurs évaluations en utilisant des informations sensibles des candidats. Dans ce chapitre, nous étudions les performances et le biais potentiel des systèmes d'analyse automatique de la convocabilité sur un ensemble de données vidéo de monologues accessibles au public.

Selon les pratiques en vigueur dans l'industrie, les systèmes traitent généralement plusieurs modalités : les expressions faciales, le langage et les indices vocaux des candidats, ce qui rend leurs biais difficiles à anticiper dans ces multiples modalités. Dans le cadre d'un système multimodal d'apprentissage profond pour la prédiction de l'embauche, nous montrons que le sexe et l'origine ethnique peuvent être retrouvés en utilisant des classifieurs simples à partir de la représentation latente, en particulier pour les indices vocaux et visuels qui pourraient être critiques pour le processus de sélection.

Alors que les approches actuelles se concentrent sur l'évaluation de l'équité des systèmes d'analyse automatique des entretiens vidéo (voir section 3.5.2 p. 39), nous nous intéressons ici à la fois au développement d'un modèle qui puisse s'affranchir de variables sensibles et à l'évaluation de ce modèle. Nous proposons une première approche qui exploite une branche adversaire pour apprendre une représentation ignorant les variables sensibles (dans notre cas le sexe et l'ethnicité). Nous montrons expérimentalement qu'elle assure une meilleure représentation sans perte significative de performances sur la tâche principale.

Cette procédure nécessite encore l'annotation explicite des variables protégées, ce qui

n'est pas possible dans la pratique et même strictement interdit dans certains pays européens (par exemple la France). Nous proposons donc une deuxième approche qui ne nécessite pas de collecte d'informations supplémentaires : cette fois, nous entraînons le système à ne pas utiliser la représentation faciale des candidats. Avec cette approche, nous espérons éviter que le modèle n'intègre des informations sur les caractéristiques faciales du candidat ainsi que sur son sexe ou son origine ethnique.

La suite du chapitre est découpée en quatre parties. Premièrement, nous posons le cadre législatif dans lequel s'inscrit notre étude. Deuxièmement, nous présentons le matériel utilisé dans ce chapitre. Troisièmement, nous présentons l'architecture adverse pour obtenir une représentation égalitaire entre candidats. Finalement, nous présentons les expériences et les résultats obtenus concernant les performances d'une telle architecture.

Publication associée à ce chapitre : En soumission, Hemamou Léo, Guillon Arthur, Martin Jean-claude, Clavel Chloé. Using Adversarial Learning for Removing Sensitive Information in Neural Representation for Hireability Prediction in Monologue Videos

8.1 Influence du cadre législatif français

Nous nous appuyons sur le cadre législatif français dans les choix effectués pour la collecte de données et l'utilisation des méthodes d'apprentissage automatique. Nous décrivons succinctement l'influence de ce cadre sur les choix méthodologiques de ce chapitre.

8.1.1 Collecte de données sensibles

Tout d'abord, la collecte de données sensibles est à la fois nécessaire pour auditer le modèle d'apprentissage et à la fois difficile, voire impossible, par rapport au cadre juridique. En effet, dans le contexte du Règlement Général de la Protection des Données, il est interdit de collecter des données concernant l'origine raciale ou ethnique, les opinions politiques, les croyances religieuses ou philosophiques, ou l'appartenance à un syndicat (article 9[a] du RGPD). En théorie, ces questions pourraient être contournées en demandant aux candidats un consentement explicite pour l'utilisation de leurs informations personnelles, mais cette demande est problématique dans un contexte industriel. Ainsi, une telle demande pourrait détériorer la confiance du candidat envers l'entreprise et même l'amener à quitter le processus de sélection. Dans ce sens, nous n'avons pas trouvé de clients prêts à demander ce consentement à leurs candidats.

Nous avons donc choisi d'utiliser le seul jeu de données public annoté en convocabilité

et ayant déjà collecté des variables sensibles des candidats (genre et ethnie).

8.1.2 Équité individuelle ou équité de groupe ?

La question de l'équité et sa traduction en termes de métriques pour l'apprentissage automatique sont largement discutées dans la littérature [Corbett-Davies and Goel, 2018]. Au sein de ces métriques, nous pouvons distinguer les métriques dites individuelles et les métriques de groupe. Les métriques individuelles consistent à déterminer si deux individus, avec des variables sensibles différentes, sont traités ou non de la même façon. Les métriques de groupe consistent à évaluer si certaines mesures (taux de sélection, erreur des algorithmes, etc.) diffèrent d'un groupe à l'autre. Dans le contexte où l'ethnicité n'est pas reconnue par l'état républicain français, les mesures liées aux disparités de groupe sont encore très compliquées à mettre en oeuvre [Streiff-Fénart, 2012, Lieberman, 2001]. Ainsi, au contraire de l'équité de groupe largement employé aux États-Unis (voir section 2.7), l'équité individuelle demeure mieux adaptée au cadre européen.

8.1.3 Des méthodes proscrites

Au-delà de la métrique, certaines méthodes peuvent être problématiques au regard de la loi, et notamment vis-à-vis de la discrimination positive interdite en France. En effet, plusieurs travaux utilisent les variables sensibles pour obtenir un résultat plus juste entre candidats, soit directement, soit en proposant l'utilisation de modèles spécifiques pour chacun des groupes. Cependant, de telles méthodes sont problématiques au regard de la loi française, où chaque personne doit être traitée de la même façon. La discrimination positive interdite en France contraint dès lors à ne pas utiliser de modèles spécifiques ou de valeurs seuils spécifiques à chacun des groupes minoritaires [Lipton et al., 2018].

Ainsi, nous avons opté pour une méthodologie visant à retirer toute information sensible de la représentation latente utilisée pour effectuer l'inférence du classifieur. Une telle méthodologie a trois intérêts : 1) elle permet d'assurer un traitement égalitaire, 2) elle renforce l'anonymisation des données, 3) elle n'est pas en contradiction avec les mesures législatives françaises.

De plus, nous proposons une avancée pour l'utilisation d'un tel système sans la collecte spécifique de données sensibles permettant ainsi d'obtenir un système viable dans le cadre législatif français.

8.2 Matériel

Les données utilisées dans cet article sont des vidéos provenant de l'ensemble de données « ChaLearn Looking at People (LaP) » [Escalante et al., 2020]. Cette section présente l'ensemble des données et les descripteurs multimodaux utilisés.

8.2.1 Description du jeu de données

L'ensemble de données ChaLearn LaP est constitué de 10000 clips vidéo extraits de plus de 3000 vidéos YouTube de personnes faisant face à la caméra et parlant en anglais. Le nombre de clips par vidéo varie de 1 à 5 et leur durée est de 15 s. L'objectif de cet ensemble de données est de modéliser les premières impressions exprimées grâce à l'annotation des traits de personnalité Big 5, et l'aptitude à l'embauche de la personne perçue par des personnes tierces. Chaque vidéo a été annotée par des travailleurs d'Amazon Mechanical Turk, à qui il a été demandé d'évaluer les traits de personnalité et de décider s'ils inviteraient ou non les personnes à un entretien d'embauche.

Variables sensibles. Chaque vidéo a été annotée avec le sexe (homme ou femme) et l'origine ethnique (asiatique, caucasienne ou afro-américaine), selon la répartition suivante : 5460 annotations masculines et 4538 annotations féminines, et 331 annotations asiatiques, 8598 caucasiennes et 1071 afro-américaines. Dans ce chapitre, nous considérons ces annotations (sexe et ethnicité) comme les variables protégées qui ne doivent pas être utilisées lors de l'inférence de la convocabilité.

Représentations statiques des visages des candidats. Nous envisageons également le cas où nous n'aurions pas accès aux variables protégées (sexe et ethnicité). Dans ce cas, nous visons à forcer le système à n'utiliser aucune information relative aux attributs du visage. Nous utilisons les niveaux de confiance d'OpenFace pour extraire quelques images (c'est-à-dire 5) lorsque les unités d'action du visage sont minimales (pour obtenir le visage le plus neutre). Ensuite, une représentation du visage est obtenue en utilisant un réseau neuronal pré entraîné pour la tâche de reconnaissance faciale, à savoir ArcFace [Deng et al., 2019]. Nous obtenons un vecteur à 512 dimensions par candidat en faisant la moyenne des représentations de visages appartenant à chaque candidat. Enfin, nous extrayons une représentation bidimensionnelle de ces représentations de visages en utilisant l'algorithme de réduction de dimension UMAP [McInnes et al., 2018]. De la même manière que pour l'entraînement contre le sexe et l'ethnicité, nous examinons l'entraînement adversaire contre la représentation encodée des visages.

8.2.2 Descripteurs des vidéos monologues

Les descripteurs utilisés pour décrire les vidéos sont similaires à ceux extraits dans les expériences partie II.

Indices prosodiques (audio) : nous extrayons les descripteurs audio à chaque trame de vidéo grâce à la bibliothèque OpenSmile [Eyben et al., 2013a]. Nous avons utilisé l'ensemble de descripteurs ComParE [Schuller et al., 2013], qui est un standard dans la communauté de l'informatique affective, utilisé à l'origine dans le premier défi ChaLearn [Escalante et al., 2020]. Cet ensemble donne un vecteur de 130 dimensions extrait à la fréquence de 100 Hz. Nous avons utilisé une représentation lissée de ces vecteurs audio, en utilisant une fenêtre temporelle de 0,5 s et un chevauchement de 0,25 s, qui sont des paramètres classiques dans la recherche en informatique sociale [Varni et al., 2018].

Expressions faciales (OpenFace) : Les descripteurs des expressions faciales sont extraits à chaque trame de la vidéo à l'aide de la bibliothèque OpenFace [Baltrusaitis et al., 2018]. Elles comprennent : la position et la rotation de la tête, la présence et l'intensité des unités d'action du visage et la direction du regard résultant en un vecteur de dimension 52. Comme pour l'audio, ces valeurs sont lissées dans le temps, en utilisant les mêmes paramètres de fenêtre temporelle et de chevauchement.

Contenu verbal (BERT) : Il existe des transcriptions manuelles de chaque clip vidéo, que nous utilisons pour calculer une représentation BERT de chaque transcription [Devlin et al., 2019]. Nous utilisons le modèle BERT de base fourni par HuggingFace [Wolf et al., 2020] et calculons une représentation contextualisée de chaque mot en utilisant l'avant-dernière couche. Nous construisons une représentation de la transcription en séquençant les représentations mot par mot. Chaque mot est représenté par un vecteur de dimension 768.

8.3 Atténuer les biais dans la prédiction de la convocabilité par un apprentissage adversaire

8.3.1 Formalisation

Dans notre modèle, nous désignons chaque instance par le triplet X, Y, Z où X désigne la réponse du candidat, Y l'étiquette de convocabilité de ce candidat et Z la variable protégée qui ne doit pas être utilisée pour l'inférence de Y . Soit $X_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$ la séquence de descripteurs de bas niveau décrivant la réponse du candidat respectivement à partir des trois modalités : langue (L), vidéo (V) et audio (A). T_m et d_m représentent la longueur de la séquence et la dimension des éléments de la modalité $m \in \{L, V, A\}$

car chaque modalité a une fréquence d'échantillonnage et une dimension de descripteurs différentes.

Le but de notre modèle est de maximiser la performance de la tâche principale (la prédiction de la convocabilité Y) tout en minimisant la performance d'un classifieur adverse de prédire Z . La variable Y est une variable binaire (c'est-à-dire convocable ou non convocable). Enfin, la variable protégée Z est soit une variable catégorielle (par exemple, le sexe ou l'origine ethnique), soit un vecteur continu (par exemple, la représentation des visages).

8.3.2 Architecture

Notre système est composé de trois parties : a) le réseau principal qui encode les informations de la réponse multimodale X à une représentation latente \tilde{h} , b) le classifieur de convocabilité qui classe la réponse sur la base de la représentation \tilde{h} et c) la branche adverse qui essaie de prédire Z sur la base de la représentation \tilde{h} . Un schéma de l'architecture est disponible en figure 8.1.

Réseau principal.

Le réseau principal est largement inspiré de HireNet chapitre 6 à la différence de l'absence de contexte (intitulé des questions et du poste) et une fusion multimodale au niveau de la vidéo et non des signaux sociaux.

Nous encodons chaque modalité séparément par un encodeur de modalité spécifique afin d'obtenir une meilleure représentation de chaque élément pour la représentation intra-modalité [Poria et al., 2017]. Pour ce faire, nous utilisons un composant de réseau neuronal récurrent, à savoir un Bidirectional Gated Recurrent Unit (BiGRU).

$$z_t^m = BiGRU(x_t), t \in [1, T_m] \quad (8.1)$$

Afin d'obtenir un vecteur de taille fixe pour chaque modalité et de tenir compte des moments saillants dans la réponse du candidat chapitre 5, nous utilisons un mécanisme d'attention additive, décrit par les équations suivantes :

$$\begin{aligned} u_t^m &= \tanh(W_A^m z_t^m + b^m) \\ \alpha_t^m &= \frac{\exp(u_p^{m\top} u_t^m)}{\sum_{t'} \exp(u_p^{m\top} u_{t'}^m)} \\ o^m &= \sum_t \alpha_t^m z_t^m \end{aligned} \quad (8.2)$$

où W_A^m est une matrice de poids, u_p^m et b^m sont des vecteurs de poids et $u_p^{m\top}$ dénote la transposée de u_p^m pour chaque $m \in \{L, A, V\}$.

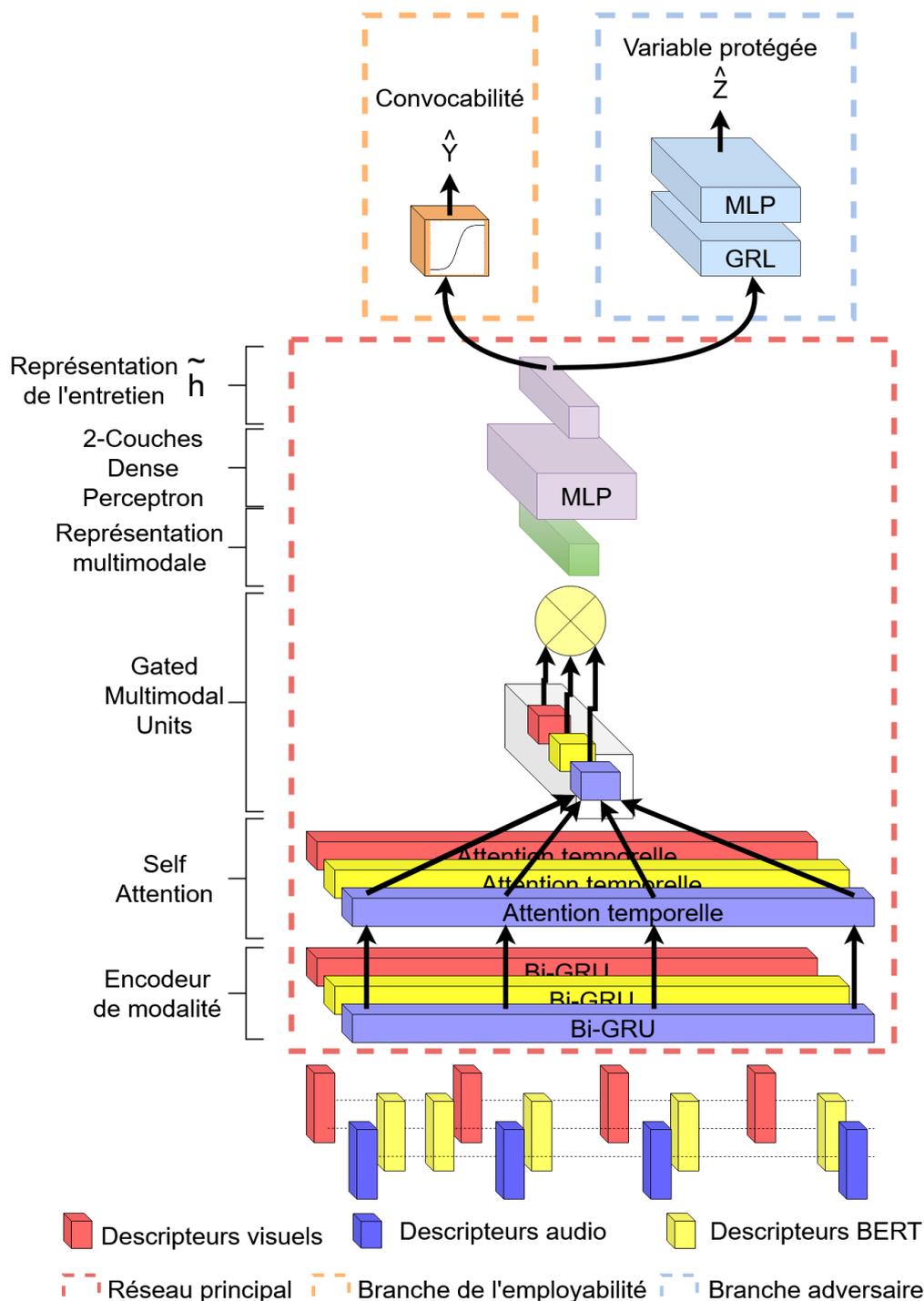


FIGURE 8.1 – Architecture multimodale équitable proposée. Les versions monomodales sont obtenues en utilisant uniquement une modalité en entrée et en retirant le GMU.

Nous proposons d'améliorer la représentation multimodale en fusionnant les trois modalités grâce à l'utilisation d'une unité multimodale à portes (GMU [Arevalo et al., 2020]). Le GMU fonctionne en projetant toutes les modalités sur le même espace et en apprenant la contribution de chacune d'entre elles par le biais d'un mécanisme à portes.

$$\begin{aligned}
 \tilde{o}^a &= \tanh(W_{Aproj}o^a) \\
 \tilde{o}^l &= \tanh(W_{Lproj}o^l) \\
 \tilde{o}^v &= \tanh(W_{Vproj}o^v) \\
 \sigma^a &= \sigma(W_{Agating}[o^a, o^l, o^v]) \\
 \sigma^v &= \sigma(W_{Lgating}[o^a, o^l, o^v]) \\
 \sigma^l &= \sigma(W_{Vgating}[o^a, o^l, o^v])
 \end{aligned}$$

$$o^{multimodal} = \sigma^a * \tilde{o}^a + \sigma^l * \tilde{o}^l + \sigma^v * \tilde{o}^v \quad (8.3)$$

où W_{Aproj} , W_{Lproj} , W_{Vproj} , $W_{Agating}$, $W_{Lgating}$, $W_{Vgating}$ sont des matrices de poids, σ est la fonction sigmoïde et $[x, y]$ est la concaténation de x et y . La sortie de l'unité multimodale est alors donnée par $o^{multimodal}$, qui représente l'information multimodale de la réponse.

En plus de la représentation multimodale, nous avons choisi d'utiliser une simple combinaison de deux couches denses, ce qui est une simplification des architectures trouvées dans la littérature pour l'analyse des EVD ([Leong et al., 2019] et chapitre 6)

$$h_1 = \tanh(W_1^\top o^{multimodal} + b_1) \quad (8.4)$$

$$\tilde{h} = \tanh(W_2^\top h_1 + b_2) \quad (8.5)$$

où W_1 et W_2 sont des matrices de poids et b_1 et b_2 sont des vecteurs de poids. Nous dénotons par θ_e les paramètres du réseau principal.

Classifieur pour la convocabilité. Une fois \tilde{h} obtenu, nous l'utilisons comme représentation afin d'inférer la convocabilité :

$$\hat{Y} = \sigma(W_v \tilde{h} + b_v) \quad (8.6)$$

où W_v est une matrice de poids et b_v un vecteur de poids. Nous dénotons par θ_h les paramètres du classifieur pour la convocabilité.

Comme le problème auquel nous sommes confronté est celui d'une classification binaire, la fonction de coût de la tâche principale est :

$$\mathcal{L}_T = -\frac{1}{N} \sum_{i=1}^N Y_i \log \hat{Y}_i + (1 - Y_i) \log(1 - \hat{Y}_i)$$

où Y dénote l'étiquette de la convocabilité des candidats.

Branche neuronale adverse.

8.3. ATTÉNUER LES BIAIS DANS LA PRÉDICTION DE LA CONVOCABILITÉ PAR UN APPRENTISSAGE ADVERSAIRE

Variable sensible	Jeu de données	Jeu d'entraînement (6991 clips)	Jeu de validation (1448 clips)	Jeu de test (1559 clips)
Genre - Femme	0.560	0.560 (3053)	0.575 (702)	0.541 (783)
Genre - Hommes	0.495	0.482 (3938)	0.524 (746)	0.519 (776)
Disparate Impact	0.883	0.860	0.911	0.959
Ethnicité - Asiatique	0.570	0.584 (236)	0.718 (32)	0.444 (63)
Ethnicité - Caucasien	0.541	0.537 (5998)	0.545 (1273)	0.554 (1327)
Ethnicité - Afro-Américains	0.434	0.424 (757)	0.559 (143)	0.374 (171)
Disparate Impact	0.761	0.726	0.759	0.675

TABLE 8.1 – Résumé des biais initiaux par rapport au nouveau découpage des données proposée

Afin de forcer le réseau à ne pas utiliser les informations sensibles, nous utilisons une approche adversaire pour entraîner une seconde branche, à partir de la représentation latente \tilde{h} . Cette branche tente de retrouver les informations sensibles des candidats. Cette deuxième branche se greffe sur le réseau principal par le biais d'une couche d'inversion de gradient (GRL, Ganin2015UnsupervisedBackpropagation). Une GRL est une couche spéciale sans vecteur de poids ni matrice, qui se comporte comme la fonction d'identité lors de la passe en avant, mais qui multiplie le gradient par -1 lors de la passe en arrière, ce qui fait que le réseau *désapprend* l'information correspondante. Afin d'obtenir une représentation plus juste, nous avons décidé d'utiliser un réseau plus profond que la branche de la convocabilité. De plus, comme le type de variables protégées change, nous utilisons une fonction d'activation ou de coût différente selon le type d'informations à protéger. Plus précisément :

$$\hat{Z} = \begin{cases} \text{softmax}(W_4 \sigma(W_3 \tilde{h} + b_3) + b_4) & \text{pour les variables catégoriques } Z \\ W_4(W_3 \tilde{h} + b_3) + b_4 & \text{pour les variables continues } Z \end{cases}$$

où W_3, W_4 sont des matrices de poids et b_3, b_4 sont des vecteurs de poids. Nous dénotons par θ_a les paramètres de la branche adverse. La fonction de coût pour la branche adverse est définie comme suit :

$$\mathcal{L}_A = \begin{cases} -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C Z_{ij} \log(\hat{Z}_{ij}) & \text{pour les variables catégoriques } Z \\ \frac{1}{N} \sum_{i=1}^N (Z_i - \hat{Z}_i)^2 & \text{pour les variables continues } Z \end{cases}$$

où C indique le nombre de classes pour la variable Z si catégorique (par exemple 2 pour le sexe et 3 pour l'appartenance ethnique dans notre cas).

Enfin, le modèle complet est entraîné en optimisant l'objectif min-max suivant :

$$\min_{\theta_e, \theta_h} \max_{\theta_a} \mathcal{L}_T(\theta_e, \theta_h) - \lambda \mathcal{L}_A(\theta_e, \theta_a)$$

où $\lambda \geq 0$ est un hyperparamètre de compromis entre l'objectif de convocabilité et l'objectif adversaire.

8.3.3 Stratégie d'entraînement pour le réseau adverse

Nous avons suivi cette stratégie afin d'entraîner le modèle :

1. Premièrement nous entraînons le modèle principal et la branche liée à la convocabilité (θ_e et θ_h).
2. Ensuite, nous entraînons uniquement la branche adverse (θ_a)
3. L'entraînement adverse est ensuite organisé par l'entraînement alternatif de l'ensemble complet du réseau (θ_e, θ_h , et θ_a) ou uniquement de la branche adverse (θ_a)
 - (a) Entraînement du réseau complet (réseau principal, branche pour la convocabilité, branche adverse connectée par le GRL) pour une époque (θ_e, θ_h , et θ_a)
 - (b) réinitialisation des poids de la branche adverse (θ_a)
 - (c) entraînement de la branche adverse jusqu'il n'y ait plus d'amélioration sur la fonction coût pour le jeu de validation (θ_a)

Nous réitérons cette boucle jusqu'à ce que la fonction coût définie par

$$\min_{\theta_e, \theta_h} \max_{\theta_a} \mathcal{L}_T(\theta_e, \theta_h) - \lambda \mathcal{L}_A(\theta_e, \theta_a)$$

ne diminue plus sur le jeu de validation.

8.4 Expériences

Nous avons mené plusieurs expériences pour évaluer notre approche. Nous détaillons les métriques d'évaluation ci-dessous. Nous illustrons ensuite les performances du réseau principal sur la seule tâche de prédiction de la convocabilité, sans entraînement adverse. Enfin, nous décrivons les expériences liées aux deux approches proposées pour la suppression des informations sensibles.

8.4.1 Métriques d'évaluation

Les mesures d'évaluation utilisées pour la tâche de convocabilité sont la justesse et l'aire sous la courbe (AUC). Ces mesures d'évaluation sont bien adaptées à la classification binaire et ont été largement utilisées dans l'analyse automatique des EVD [Chen et al., 2017, Escalante et al., 2020]. Ensuite, nous utilisons deux mesures pour l'évaluation de l'équité. La première mesure consiste à mesurer la quantité d'informations sensibles présentes dans la représentation latente \tilde{h} définie dans l'équation (8.5).

Nous utilisons le protocole suivant : tout d’abord, nous entraînons deux *classifieurs diagnostiques*, la régression logistique (LR) et XGBoost (XGB) pour inférer la variable protégée (sexe ou ethnicité) à partir de la représentation latente du réseau, en utilisant les mêmes ensembles d’entraînement, de validation et de test. Nous utilisons l’AUC de ces classifieurs comme mesure d’évaluation. Dans le cas d’un problème multi-classes, nous rapportons la moyenne d’AUC d’un contre le reste (one-vs-rest en anglais). Ce protocole est largement utilisé dans la littérature sur la protection des données et la littérature en apprentissage machine équitable [Jaiswal and Mower Provost, 2020, Elazar and Goldberg, 2018] et nous permet d’estimer le degré d’*anti-classification* et d’équité individuelle du classifieur de la convocabilité (voir section 3.5.2). De mauvaises performances des classifieurs diagnostiques signifient que moins d’informations sensibles sont contenues dans la représentation latente.

La deuxième mesure est l’effet disparate ou plus communément appelé en anglais *disparate impact* (DI) défini comme

$$DI = \frac{Pr(Y = 1 | Z = \text{non privilégié})}{Pr(Y = 1 | Z = \text{privilégié})}$$

Bien qu’elle ne soit pas explicitement optimisée par l’approche que nous proposons, elle nous permet de mesurer une potentielle inégalité dans le traitement de groupe dans leur ensemble. Cette mesure est utilisée dans les ressources humaines aux États-Unis (voir section 2.7 p. 20) et dans la littérature sur l’apprentissage machine équitable [Sánchez-Monedero et al., 2020]. Un DI plus proche de 1 induit un taux de sélection plus équitable entre les groupes.

8.4.2 Performance sur la tâche de convocabilité

Bien que l’ensemble de données ChaLearn LaP soit fourni avec une proposition de découpage entre les ensembles d’entraînements, de validation et de test, nous avons constaté que 84 % des clips de l’ensemble de test ont au moins un clip dans l’ensemble d’entraînement extrait de la même vidéo YouTube. Ce chevauchement est potentiellement problématique, car il pourrait permettre aux classifieurs d’obtenir de bons résultats sur la tâche de prédiction de la convocabilité en la réduisant à une tâche d’identification. Ainsi, en apprenant uniquement les caractéristiques spécifiques individuelles des intervenants (issu par exemple du visage ou l’audio), un classifieur pourrait obtenir une bonne performance pour inférer la convocabilité, une stratégie qui n’a pas de sens du point de vue du recrutement.

Pour cette raison, nous proposons d’utiliser une répartition différente des trois ensembles de données (entraînement, validation et test) pour toutes les expériences. Ce découpage as-

	LR		XGBoost	
	ACC	AUC	ACC	AUC
Découpage original	0.680	0.742	0.745	0.811
Découpage proposé	0.650	0.695	0.622	0.672

TABLE 8.2 – Différences entre les deux découpages de données pour la prédiction de la convocabilité en utilisant la représentation faciale.

sure qu’aucune vidéo n’est partagée entre les différents ensembles. Dans cette section, nous donnons un aperçu supplémentaire des raisons pour lesquelles changer la répartition des ensembles de données semble raisonnable et nous comparons les performances des modèles étudiés sur les deux découpages de l’ensemble de données concernant la tâche de convocabilité.

Pour illustrer les différences entre les deux ensembles, nous menons plusieurs expériences en comparant l’utilisation des deux découpages.

Modèle naïf par vote

Premièrement, nous montrons qu’il est possible d’obtenir de bons résultats par rapport au découpage original de ChaLearn en utilisant un modèle naïf. Pour cela, pour chaque vidéo de l’ensemble test, nous prédisons un score de convocabilité en calculant la moyenne des étiquettes de convocabilité originales des clips extraits de la même vidéo apparaissant dans le jeu d’entraînement. Les valeurs manquantes (candidats présents uniquement dans l’ensemble de test) sont remplacées par la valeur de la moyenne des étiquettes dans les jeux de données d’entraînement et de validation. En comparant ces valeurs aux étiquettes de vérité terrain de l’ensemble de test, nous obtenons une AUC de 0.797 et une justesse de 0,7215, qui sont comparables aux performances de la plupart des modèles proposés lors du défi initial ChaLearn.

Classification grâce à la représentation faciale

Ensuite, nous extrayons les représentations faciales des candidats (grâce au réseau pré-entraîné ArcFace [Deng et al., 2019]) et entraînons deux classifieurs (LR et XGBoost) à prédire la convocabilité sur chacun des découpages (original et proposé). Nous présentons les résultats obtenus dans le tableau 8.2. Les classifieurs entraînés et testés sur le découpage original ont un score plus élevé (AUC de 0,811) que le score obtenu par les classifieurs entraînés et évalués sur le découpage proposé (AUC de 0,695). Nous interprétons ce résultat

	Language		Audio		Video		Multimodal	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Découpage original	0.601	0.607	0.664	0.721	0.686	0.751	0.701	0.766
Découpage proposé	0.584	0.610	0.642	0.695	0.647	0.700	0.682	0.741

TABLE 8.3 – Différences entre les deux découpages de données pour la prédiction de la convocabilité en utilisant les réseaux proposés.

comme la preuve que le partage des vidéos survenant dans le découpage original facilite la tâche de prédiction de la convocabilité et pourrait potentiellement réduire cette tâche à celle de reconnaissance de l’identité d’une personne. De plus, cette expérience nous conforte dans le choix qu’il est parfois dangereux d’utiliser de tels descripteurs pour l’inférence de la convocabilité ce qui peut résulter en l’inférence de biais inhérent au jeu de données ou à la méthodologie.

Classification par notre réseau

Enfin, nous présentons dans le tableau 8.3 les performances des 4 modèles étudiés dans cette section (les 3 modèles unimodaux et le modèle multimodal) sur les découpages originaux et proposés.

Nous observons que la baisse de performance entre les deux découpages n’est pas très importante, sauf pour la modalité vidéo. Nous interprétons cela comme la preuve que les classifieurs ne s’appuient pas beaucoup sur les informations partagées entre les deux ensembles et qu’ils sont assez stables. Cette impression est renforcée par la comparaison de ces résultats avec le tableau 8.2.

De plus, les résultats sur la tâche de recrutement (voir tableau 8.3) présentent des différences significatives entre les modalités : les scores pour la modalité vidéo sont étonnamment élevés et ceux pour la modalité linguistique étonnamment bas par rapport aux autres modalités, ce qui est inhabituel selon la littérature [Rasipuram and Jayagopi, 2018, Chen et al., 2017] et nos travaux (voir chapitre 5) et chapitre 6) .

Nous interprétons ce résultat comme une spécificité de l’ensemble du jeu de données ChaLearn. Les premières impressions véhiculées par la modalité visuelle prennent le pas sur le contenu verbal, moins important ici tant le fond du discours n’est pas relié à des questions posées en entretien d’embauche. Enfin, le modèle multimodal est nettement plus performant que les autres, ce qui montre clairement les avantages de cette approche.

	Langage	Audio	Video	Multimodal
Justesse	0.584	0.642	0.647	0.682
AUC	0.613	0.695	0.700	0.741
DI Genre	0.984	0.837	0.591	0.695
DI Ethnicité	0.867	0.913	0.708	0.656

TABLE 8.4 – Résultats pour la tâche de la convocabilité et l’effet disparate du système automatique.

Les prédictions automatiques amplifient-elles le biais présent dans l’annotation humaine ?

Le tableau 8.1 affiche la proportion moyenne de l’étiquette employable pour chacune des classes protégées dans l’ensemble des données ainsi que dans les ensembles d’entraînement, de validation et de test. On peut voir dans ce tableau que l’ensemble de données est fortement biaisé en ce qui concerne la variable de l’ethnicité avec un effet disparate de 0,761 (comme indiqué ci-dessus, l’EEOC considère qu’un effet disparate ne devrait pas être inférieur à 0,8). En outre, le tableau 8.4 présente l’effet disparate des prédictions des systèmes automatiques sur l’ensemble de test. En ce qui concerne le sexe, les prédictions basées sur des modèles audio, vidéo et multimodaux sont moins justes que les annotations de vérité terrain, alors que les prédictions basées sur le langage sont plus justes. En ce qui concerne l’ethnicité, les modèles de langue et audio montrent un meilleur DI significativement plus élevé que les annotations humaines réelles, alors que les modèles vidéo et multimodaux semblent dégrader l’effet disparate. Il est intéressant de noter que le modèle multimodal est le pire modèle en ce qui concerne l’effet disparate alors qu’il présente les meilleures performances en ce qui concerne la capacité de convocabilité.

8.4.3 Entraînement adversaire contre le genre et l’ethnicité.

En utilisant l’approche expliquée dans la section section 8.3.2 et les réseaux pré entraînés décrits dans la section section 8.4.2, nous avons entraîné le réseau contre les variables protégées du sexe et de l’ethnicité dans deux expériences. Les expériences ont été effectuées 3 fois et nous faisons la moyenne des mesures d’évaluation. Les résultats des expériences sont présentés dans le tableau 8.5 : ce tableau compare les métriques d’évaluation des réseaux protégés et non protégés (c’est-à-dire avant entraînement adversaire), pour chaque modalité. Les résultats sont présentés sous la forme d’une série de questions et de réponses :

Pouvons-nous retrouver le sexe et l’ethnicité à partir des représentations latentes entraînés uniquement pour la tâche de convocabilité ?

Nous discutons ici des scores AUC des classifieurs diagnostiques pour le sexe et l’ethnicité sur les réseaux non protégés. Comme nous pouvons le voir de par les valeurs élevées de l’AUC, le sexe peut être récupéré à partir des modèles audio, vidéo et multimodaux, mais pas à partir du modèle linguistique. Les scores AUC pour l’ethnicité ne sont significatifs que pour les modèles vidéo et, dans une moindre mesure, multimodaux, ce qui signifie que certaines informations peuvent toujours être récupérées à partir de leur représentation latente.

Comme OpenFace ne renvoie que l’intensité des unités d’action faciale et la position et la rotation de la tête, ce résultat peut potentiellement s’expliquer soit par un éventuel biais dans OpenFace, soit par une différence dans la production des expressions faciales entre les hommes et les femmes. Enfin, contrairement à [Jaiswal and Mower Provost, 2020], notre modèle multimodal ne montre pas de fuite plus importante que nos modèles unimodaux.

Quel est l’impact de l’entraînement adverse sur les performances pour la prédiction de convocabilité ?

Pour les variables protégées et toutes les modalités, les scores des mesures d’évaluation sur la tâche de convocabilité entre les modèles protégés et non protégés semblent indiquer une perte de performance très mineure, qui est légèrement plus perceptible pour le modèle multimodal.

Quel est l’impact de l’entraînement adverse sur le genre ?

Nous discutons ici de l’AUC des classifieurs diagnostiques et des scores de l’effet disparate entre les modèles protégés et non protégés affichés dans la colonne du genre. Notre méthode montre une réduction de l’AUC pour toutes les modalités. Cette réduction est très significative pour les modalités audio, vidéo et multimodales. Nous interprétons cette baisse de performance comme une preuve qu’une grande partie des informations sensibles de la représentation cachée a été supprimée. Les valeurs d’effets disparates sont également affectées de manière significative par l’entraînement adverse, ce qui signifie que le réseau équilibre les scores des groupes privilégiés et non privilégiés. Dans le cas multimodal, cette amélioration amène la mesure DI au-dessus du seuil de 0,8.

Quel est l'impact de l'entraînement adversaire sur l'ethnicité ?

Selon les remarques précédentes, seules les représentations cachées de la vidéo et les modèles multimodaux contiennent des informations significatives sur l'ethnicité. La comparaison des scores des modèles non protégés et protégés de la colonne Ethnicité montre peu de différence entre les AUC des classifieurs diagnostiques pour les modalités linguistiques et audio. Une variation significative se produit pour les modalités vidéo et multimodales, où l'AUC chute jusqu'à une valeur proche de 0,5, c'est-à-dire une performance de classification de type aléatoire. Nous considérons que les informations sur l'ethnicité ont été retirées avec succès de la représentation latente.

Même si les informations sensibles ont été retirées des couches cachées, l'effet disparate se détériore. Il est important de souligner que même si le modèle protégé est plus proche d'un traitement plus équitable des candidats, une certaine inégalité entre les classes demeure.

8.4.4 Entraînement adversaire contre la représentation des visages

Cette section présente des expériences sur la deuxième approche proposée, qui consiste à appliquer l'entraînement adverse contre la représentation des visages statiques des candidats afin de supprimer les informations concernant les attributs du visage. Comme dans la section précédente, nous utilisons les réseaux pré entraînés décrits dans la section section 8.4.2, en évaluant les mêmes paramètres pour les deux variables protégées que sont le sexe et l'ethnicité et en réalisant les expériences 3 fois.

Les résultats sont présentés dans le tableau 8.6. Comme pour l'expérience précédente, nous observons que cette procédure a un très léger impact sur les performances du réseau en matière de convocabilité.

Les AUC des classifieurs diagnostiques sont à nouveau réduites pour les modèles vidéo et multimodaux en ce qui concerne le sexe et l'ethnicité, et pour le modèle audio en ce qui concerne le sexe, bien que cette réduction soit globalement moins importante que dans le cas précédent.

Les observations concernant les scores DI diffèrent de l'expérience précédente : aucune augmentation n'est observée pour le DI du sexe pour la modalité audio entre les modèles non protégés et protégés, alors qu'une augmentation apparaît dans le tableau 8.5. En revanche, le DI de l'ethnicité pour la modalité vidéo augmente maintenant fortement.

Dans l'ensemble, nous interprétons ces résultats comme très prometteurs en ce qui concerne l'applicabilité de cette technique aux modèles d'analyse automatique des EVD. Bien que le gain en termes de confidentialité soit un peu moins important que dans le cas précédent, aucune collecte supplémentaire d'informations sensibles n'a été nécessaire

et d'autres attributs faciaux potentiels (par exemple, l'attrait du visage ou la couleur des cheveux) [Bahng et al., 2020] sont possiblement retirés de la représentation latente.

8.5 Conclusion

Il est essentiel de garantir l'égalité de traitement des candidats lors d'un entretien d'embauche, en particulier pour les systèmes d'embauche automatiques. Bien que la formation des recruteurs et l'adoption de certaines pratiques telles que les entretiens structurés puissent aider, il est également nécessaire pour les acteurs industriels de s'assurer que leur algorithme ne discrimine pas des groupes de population. Comme ces acteurs n'ont partagé aucune preuve de non-discrimination ni de procédure standard pour protéger leurs algorithmes, nous proposons d'évaluer un système de recrutement de pointe sur un ensemble de données publiques. Bien que n'utilisant pas de caractéristiques sensibles, nous montrons que le sexe et l'ethnicité sont récupérables à partir de la représentation latente du système.

Les travaux précédents n'avaient évalué les résultats qu'en termes d'équité de groupe sans proposer de méthodes. Nous faisons un pas supplémentaire dans l'obtention d'un modèle plus juste et proposons une méthode adversaire pour garantir l'égalité de traitement. Pour résoudre ce problème, nous utilisons l'apprentissage adversaire pour retirer les informations sensibles de la représentation latente. L'approche que nous proposons est polyvalente et peut être utilisée avec différentes variables sensibles. Nous l'appliquons d'abord directement aux annotations de sexe et d'origine ethnique des candidats. Nous montrons que notre approche réussit à supprimer ces informations des représentations latentes des modèles. Cependant, comme ces annotations peuvent être difficiles à obtenir, nous proposons d'appliquer la même approche à la représentation statique des visages des candidats, qui peut être facilement extraite des entretiens vidéo.

De plus, nous observons expérimentalement que les algorithmes proposés réduisent généralement de manière significative l'effet disparate pour le sexe mais présentent des résultats plus mitigés pour l'ethnicité. Dans l'ensemble, nous proposons une méthodologie pour les systèmes d'embauche automatisés qui peut conduire à un traitement plus équitable des candidats.

			Genre		Ethnicité	
			Non protégé	Protégé	Non protégé	Protégé
Language	Convocabilité	AUC	0.613	0.611	0.613	0.609
		ACC	0.584	0.582	0.584	0.576
	AUC Diagnostique	LR	0.584	0.567	0.500	0.522
		XGB	0.532	0.521	0.496	0.512
	DI		0.984	0.957	0.867	0.832
Audio	Convocabilité	AUC	0.695	0.691	0.695	0.695
		ACC	0.642	0.638	0.642	0.640
	AUC Diagnostique	LR	0.850	0.584	0.507	0.508
		XGB	0.819	0.579	0.498	0.508
	DI		0.837	0.954	0.913	0.892
Video	Convocabilité	AUC	0.700	0.693	0.700	0.698
		ACC	0.647	0.649	0.647	0.652
	AUC Diagnostique	LR	0.745	0.589	0.661	0.482
		XGB	0.730	0.650	0.629	0.520
	DI		0.591	0.774	0.708	0.690
Multimodal	Convocabilité	AUC	0.741	0.735	0.741	0.731
		ACC	0.682	0.675	0.682	0.669
	AUC Diagnostique	LR	0.762	0.637	0.582	0.530
		XGB	0.748	0.628	0.557	0.519
	DI		0.695	0.865	0.656	0.636

TABLE 8.5 – Résultat pour l’entraînement adverse contre le genre et l’ethnicité. Pour chaque variable protégée et modalité, les mesures de la convocabilité (AUC et justesse), l’AUC des classifieurs diagnostiques et l’effet disparate sont rapportés pour les réseaux protégés et non protégés.

		Convocabilité		AUC Genre		AUC ethnicité		DI genre	DI ethnicité
		AUC	ACC	LR	XGB	LR	XGB		
Langage	Non protégé	0.613	0.584	0.584	0.534	0.500	0.497	0.984	0.867
	Protégé	0.611	0.580	0.578	0.538	0.500	0.503	0.986	0.818
Audio	Non protégé	0.695	0.642	0.850	0.819	0.506	0.498	0.837	0.913
	Protégé	0.674	0.597	0.593	0.586	0.501	0.500	0.880	0.946
Video	Non protégé	0.700	0.647	0.745	0.730	0.661	0.629	0.591	0.708
	Protégé	0.700	0.650	0.642	0.665	0.646	0.587	0.704	0.864
Multimodal	Non protégé	0.741	0.682	0.762	0.748	0.582	0.557	0.695	0.656
	Protégé	0.730	0.667	0.623	0.644	0.506	0.510	0.847	0.640

TABLE 8.6 – Effets de l’entraînement adversaire contre la représentation des visages sur les métriques d’évaluation.

Section 9

Conclusions et perspectives

9.1 Apport de notre travail

Dans cette thèse, nous avons identifié les défis posés par la mise en place de méthodes pour l'analyse automatique des entretiens vidéos différés et proposé des réponses à certains de ces défis.

Bien que de nombreuses recherches ont eu pour objet l'étude de signaux sociaux lors d'entretiens d'embauche, la plupart d'entre elles se déroulent dans un contexte de simulation d'entretien d'embauche. Notre travail de thèse s'inscrit par une collaboration avec un acteur industriel du monde des ressources humaines. Nous avons pu ainsi contribuer au domaine à travers deux jeux de données originaux dans la communauté par le nombre de candidats enregistrés et qui postulaient à de vrais postes. De plus, les entretiens vidéo différés ont une structure particulière et peu étudiées jusque maintenant alors que ce type d'entretien se développe de plus en plus. Ainsi, l'étude des EVD soulève des questions scientifiques spécifiques que nous rappellerons dans cette section. Nous présentons par la suite les réponses que notre travail apporte à ces questions.

Proposition d'un modèle neuronal pour l'analyse automatique d'entretien structuré

Nous avons proposé un modèle neuronal original adapté aux entretiens structurés nommé HireNet. Ce modèle s'appuie sur le fait qu'au cours d'une même campagne de recrutement, les mêmes questions sont posées à chaque candidat, dans le même ordre. La particularité de ce modèle réside dans son aspect hiérarchique et dans l'intégration d'informations contextuelle telle que l'intitulé des questions ou du poste. HireNet modélise l'entretien d'embauche comme une hiérarchie à deux niveaux de séquences. Le premier niveau consiste en la prise en

compte de la séquence des comportements verbaux (mots) ou non verbaux (trame audio ou d'expression faciale) de chaque réponse. Le deuxième niveau consiste en la prise en compte des séquences de question-réponse. Nous avons montré que la prise en compte de la séquentialité et de l'aspect hiérarchique des entretiens était bénéfique pour la modélisation d'une architecture neuronale. Enfin, l'ajout de mécanisme d'attention temporelle permet d'une part d'obtenir une amélioration de performances et d'autre part d'ajouter une interprétabilité locale pour l'identification de potentiels moments clés lors de l'entretien d'embauche.

- *Q.R.1 => Peut-on proposer une architecture neuronale adaptée à modéliser les entretiens structurés ?* Nous avons montré qu'une architecture prenant en compte la séquentialité des comportements bas niveaux, la hiérarchie de l'entretien structuré et les informations contextuelles de l'entretien était adaptée pour la modélisation des entretiens structurés en vidéo différée.
- *Q.R.2 => Quels sont les comportements influents en entretien d'embauche ?* L'ajout d'une attention temporelle permet d'obtenir une interprétabilité locale intrinsèque des comportements influents en entretien d'embauche.

Etude d'une fonction d'attention temporelle contextuelle

Nous avons proposé une fonction d'attention contextuelle permettant de prendre en compte à la fois l'importance intrinsèque de chaque pas de temps de la séquence à pondérer et à la fois l'importance de leur interaction avec un vecteur contexte. Cette fonction d'attention a montré des résultats significatifs avec les données simulées. Cette conclusion est cependant mitigée lorsque cette fonction est utilisée au sein de HireNet, montrant une légère amélioration uniquement pour la modalité du langage.

- *Q.R.1 => Peut-on proposer une architecture neuronale adaptée à modéliser les entretiens structurés ?* Nous avons construit une fonction d'attention contextuelle adaptée aux entretiens structurés. Cependant, l'utilisation de cette fonction au sein de HireNet montre uniquement une légère amélioration pour la modalité langage.
- *Q.R.2 => Quels sont les comportements influents en entretien d'embauche ?* : Le précédent résultat semble indiquer que l'importance des comportements non verbaux est peu influencée par le type de question.

Proposition d'un mécanisme de fusion multimodale interprétable pour modalités bruitées

Nous avons proposé une méthode de fusion des modalités à pas de temps régulier. Cette méthode permet d'obtenir une fusion plus fin grain que les méthodes par concaténation de représentation des modalités. De plus, elle permet de s'extraire de la nécessité d'une retranscription parfaite de la parole, nécessaire par exemple par les méthodes fusionnant au niveau du mot. Nos expériences montrent que l'utilisation d'une telle méthode au sein d'HireNet améliore les performances par rapport aux modèles monomodaux. L'utilisation des GMU permet d'obtenir une interprétabilité locale dans la façon dont chaque modalité contribue pour l'obtention de la représentation multimodale. Enfin, l'utilisation du GMU suivi d'une attention temporelle permet de mettre en exergue des moments multimodaux singuliers.

- *Q.R.2 => Quels sont les comportements influents en entretien d'embauche ?* Une fusion bas niveau des modalités associées à une attention temporelle permet d'obtenir une interprétation locale intrinsèque des comportements multimodaux influents.
- *Q.R.3 => Comment fusionner plusieurs modalités potentiellement bruitées ?* Nous fusionnons les modalités à pas de temps régulier. Une unité GMU permet de contrôler la contribution de chacune des modalités. Cette méthode permet de fusionner à un niveau fin grain tout en surmontant l'aspect localement bruité de certaines modalités comme par exemple la modalité textuelle obtenue par des systèmes de transcription automatique de la parole.

Proposition de méthodes d'analyse approfondie des mécanismes d'attention

Nous avons proposé une méthode d'analyse pour l'attention temporelle afin de comprendre d'une façon plus générale les tendances apprises. Nos expériences semblent tout d'abord montrer que les questions-réponses les plus influentes se situent au début de l'entretien vidéo différé. Nous nous sommes intéressés par la suite à des instants considérés comme importants par notre système que nous avons nommés les tranches d'attention. Nos expériences montrent que ces instants se déroulent plus souvent au début et à la fin de réponse pour les modalités audio et vidéo tandis qu'elles sont distribuées de façon quasi uniforme pour les modalités langage et multimodales. Nous avons ensuite montré que ces

tranches d'attention différent de moments aléatoires par leur contenu. Nous avons étudié la spécificité du contenu de ces tranches d'attention en terme de descripteurs influents. En ce qui concerne les tranches d'attention multimodales, le mécanisme de fusion semble indiquer que la modalité audio contribue le plus, suivies de la modalité du langage, puis celle de la vidéo. Enfin, nous montrons que les tranches d'attention comportent plus d'information pertinente que les tranches aléatoires pour la tâche de prédiction de la convocabilité.

- *Q.R.2 => Quels sont les comportements influents en entretien d'embauche ?* Nous avons proposé une méthodologie pour détecter et caractériser automatiquement les comportements jugés les plus influents en entretien d'embauche par un modèle d'apprentissage profond. Nous mettons en exergue certains comportements verbaux et non verbaux considérés comme importants par notre système.

Proposition d'une méthodologie pour l'obtention d'un traitement plus égalitaire entre candidats avec ou sans la nécessité d'obtention d'informations sensibles.

Nous avons proposé une représentation multimodale neuronale insensible au genre et à l'ethnicité à l'aide d'une méthode adversaire pour l'analyse automatique de la convocabilité. Cette première méthode nécessite la collecte d'informations sensibles, souvent impossible en pratique, notamment en France. Nous avons donc proposé une deuxième méthodologie ne nécessitant pas d'annotation de ces informations. Nos expériences montrent qu'il est possible d'obtenir une représentation plus équitable sans perte de performances pour la tâche de convocabilité.

- *Q.R.4 => Comment limiter les potentiels biais de modèles neuronaux ?* Nous avons proposé une méthode pour obtenir une représentation multimodale sans information liée au genre et à l'ethnicité grâce à une méthode adversaire. Cette méthode permet de ne pas prendre en compte des éléments non pertinents lors de la prédiction de la convocabilité.

9.2 Perspectives de recherche

Interprétabilité de la modélisation et de l'influence du contexte

Notre travail sur l'interprétabilité s'est principalement focalisé sur la compréhension des comportements bas niveaux influents. Cependant une large partie de notre modèle s'intéresse à la modélisation du contexte de la question posée et plus largement du contexte du type de poste.

Il serait intéressant de comprendre dans quelle mesure ce contexte est encodé. Retrouvons-nous, comme dans la littérature de la psychologie du travail, des groupes distincts de questions comportementales ou situationnelles? Une telle constatation permettrait de retrouver automatiquement une typologie déjà bien adoptée dans la communauté de la psychologie du travail. De plus, des nouveaux types de questions pourraient apparaître et permettre d'affiner cette topologie.

Il serait aussi intéressant de comprendre dans quelle mesure le contexte influence l'attention temporelle au niveau de la réponse et au niveau de l'entretien : certaines questions impliquent-elles un pic d'attention au niveau de l'entretien? Quels mots sont influents en fonction du contexte?

Un début de réponse à ces questions est envisageable par une visualisation des états cachés des encodeurs de contexte (section 6.1.2) et une analyse fine de la valeur d'attention temporelle associée à l'interaction entre les pas de temps et le contexte (section 6.1.2).

De la prédiction de la convocabilité vers la prédiction d'attributs spécifiques

Nous avons considéré tout au long de cette thèse la tâche de prédiction de convocabilité comme une tâche de classification binaire. Cependant, la dimension de la convocabilité va bien au-delà d'une simple classification binaire entre bons ou mauvais candidats. Tout d'abord, cette dimension pourrait être échelonnée selon plusieurs niveaux graduels ou évaluée de façon continue comme cela a été proposé dans de précédentes études [Rasipuram and Jayagopi, 2018, Naim et al., 2018].

Deuxièmement, nous pourrions concentrer notre analyse sur les critères d'évaluation de l'entretien d'embauche choisis par le recruteur au sein de la plateforme (compétences communicationnelles, expériences professionnelles, etc.). De cette façon, nous pourrions comprendre si certains critères sont évaluables au travers d'un outil automatique, comprendre quels signaux sociaux ou mots sont associés spécifiquement à ces critères au travers de la méthode développée et explorer la relation entre critères et évaluation de la convocabilité. De plus, ce fractionnement de la notion de convocabilité permettra d'ajouter une interpré-

tabilité au modèle automatique en fournissant des informations complémentaires lors des prédictions [Escalante et al., 2020]. Finalement, il serait intéressant de ne pas s'intéresser uniquement à la performance en entretien mais directement à la performance au travail dans le sens où certaines dimensions (comme l'anxiété) peuvent être influentes pour la première performance mais pas nécessaire pour la seconde [Schneider et al., 2019].

Dans ce sens, un premier travail avec un élève de master, Orson Jay, a été initié afin de proposer un système multi-tâches interprétable pour répondre à ces questions de recherche.

Vers une amélioration de l'extraction et de l'évaluation des éléments clés de l'entretien : représentation, fonction d'attention temporelle et évaluation humaine

Les mécanismes d'attention temporelle sont un élément clé pour l'amélioration des performances du système, mais surtout pour l'exploration des moments clés de l'entretien d'embauche. Au cours de cette thèse, nous avons proposé une méthodologie pour extraire des moments clés choisis au regard des valeurs d'attention temporelles. Cependant, ces valeurs d'attentions sont très dépendantes de la séquence d'entrée et par conséquent de la représentation choisie pour modéliser la réponse des candidats. Cette représentation peut être améliorée en considérant par exemple une nouvelle modalité (par l'annotation automatique de la partie supérieure du corps ou des gestes du candidat) ou en améliorant la façon dont les modalités sont fusionnées (l'utilisation d'un système multi vues pour la fusion des modalités pourrait être intéressante [Zadeh et al., 2019]).

Deuxièmement, les valeurs d'attention temporelle sont largement dépendantes de la fonction d'attention choisie. Dès lors, on peut se poser la question de l'utilisation du softmax pour une telle fonction qui va plus naturellement sélectionner un nombre restreint de pas temps. Ainsi, la conception de fonctions ou de méthodes spécifiques pour l'obtention ou l'affinement de l'attention temporelle [Li et al., 2018, Long et al., 2019, Doughty et al., 2018] serait une direction intéressante. De plus, il serait intéressant d'améliorer l'attention afin de déterminer le caractère positif ou négatif des comportements extraits.

Nous avons limité notre évaluation de la pertinence de ces moments à une évaluation automatique. Il serait important de mener une étude perceptive humaine sur l'utilité de ces moments par soit l'obtention d'une annotation temporelle des moments importants de la part des recruteurs soit par un questionnaire sur l'utilité des moments extraits de la vidéo. De plus, l'identification de groupes de comportements de façon non supervisée et par annotations humaines permettrait de constituer un dictionnaire des comportements influents en entretien d'embauche. Enfin, beaucoup de travail est encore nécessaire pour passer d'une interprétabilité à une explicabilité. L'expertise et la disponibilité des recruteurs

et des chercheurs en psychologie sociale et du travail sont nécessaires afin de comprendre au mieux le rôle des signaux sociaux (signification des sourires selon leurs dynamiques, signification du sourire selon le moment du tour de parole, identification des gestions de l'impression, importance du contexte, etc.).

Aide à l'entraînement aux candidats

Nous avons construit au travers de cette thèse un outil automatique pour analyser automatiquement un entretien d'embauche vidéo différé. Cet outil pourrait être un élément clé pour une aide à l'entraînement aux candidats afin de comprendre quels comportements étaient impactants lors de sessions d'entraînement. De plus un tel outil pourrait être déployé en complémentarité d'une plateforme collaborative [Zhao et al., 2017], ou intégré à un agent conversationnel animé [Hoque et al., 2013]. Des conseils automatiques basés sur les tranches d'attention, pour chacune des modalités, pourraient être intéressants.

Généralisation et personnalisation

Une des limites de notre travail réside dans le fait qu'un seul type de poste, nommé commercial, a été exploré. Nous partageons le point de vue de [Nguyen and Gatica-Perez, 2016] concernant le fait que l'analyse automatique des entretiens d'embauche reste à privilégier pour les postes nécessitant des compétences interpersonnelles et communicationnelles. Aussi, il serait intéressant d'évaluer la pertinence d'un tel outil sur d'autres types de poste comme chefs de projet ou employés des ressources humaines. En ce sens, la généralisation du modèle à plus de positions pourrait être envisagée. De plus, il serait intéressant d'étudier si des comportements sont généraux ou spécifiques à des postes en particulier [Ruben et al., 2015, Liu et al., 2017].

D'autre part, certaines informations n'ont pas été modélisées au travers de notre modèle et leur étude constitue une direction de recherche intéressante. Ainsi, la modélisation du recruteur et de la culture de l'entreprise pourrait être explorée. Cette modélisation pourrait permettre de mettre en exergue des profils de recruteurs, d'exposer des possibles biais personnels et d'évaluer la dynamique recruteur-candidat dans la modalité de l'entretien vidéo différé.

Vers un traitement encore plus juste des candidats

Nous avons, au travers de ce travail, proposé une méthode permettant une analyse automatique des réponses des candidats qui s'assure de n'utiliser aucune information liée à leur apparence faciale. Cependant, de nombreux autres marqueurs constituent d'autres sources

de discrimination menant à un traitement inégalitaire comme les accents étrangers, l'obésité, les tatouages ou les handicaps. Dans un premier temps, il serait intéressant d'améliorer et d'étendre la méthode à la suppression d'informations liées à l'empreinte vocale. Dans un deuxième temps, nous pensons que la question du handicap est une question primordiale à traiter, mais qui nécessite un travail bien plus important tant la nature du handicap peut être diverse, dynamique, et subtile. Enfin, nous avons traité le cas particulier de l'égalité qui est justifié dans le cadre juridique européen. Cependant, en fonction de la législation, l'étude d'une équité de groupe pourrait être envisageable.

Notre travail de recherche ouvre de nombreuses possibilités pour l'analyse automatique de comportements verbaux et non verbaux. Pour l'étude exploratoire des signaux sociaux influents, nos travaux pourraient permettre de pointer vers des comportements courts qui peuvent être par la suite étudiés de façon plus classique. Ainsi, nos travaux pourraient être appliqués à une tâche connexe comme la détection de mensonges et des stratégies d'impressions dans les entretiens d'embauches. Pour l'analyse automatique des entretiens structurés, nos travaux proposent une architecture hiérarchique adaptée à de tels entretiens. Dans ce sens, une adaptation de notre système à l'analyse automatique des entretiens d'embauche face à face est envisageable. Pour terminer, les travaux décrits dans cette thèse visent à mieux comprendre les possibilités et limites des traitements informatiques de signaux notamment sociaux dans le cas d'entretiens vidéos différés avec des vrais candidats postulant pour des vrais postes.

Bibliographie

- [Aafaq et al., 2019] Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., and Shah, M. (2019). Video description : A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 52(6) :1–28.
- [Agarwal et al., 2018] Agarwal, A., Beygelzimer, A., Dudfk, M., Langford, J., and Hanna, W. (2018). A reductions approach to fair classification. *35th International Conference on Machine Learning, ICML 2018*, 1 :102–119.
- [Ali et al., 2015] Ali, M. R., Crasta, D., Jin, L., Baretto, A., Pachter, J., Rogge, R. D., and Hoque, M. E. (2015). LISSA - Live Interactive Social Skill Assistance. *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 173–179.
- [Aloufi et al., 2020] Aloufi, R., Haddadi, H., and Boyle, D. (2020). Privacy-preserving Voice Analysis via Disentangled Representations.
- [Ancona et al., 2017] Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. pages 1–16.
- [Anderson et al., 2013] Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., and Sabouret, N. (2013). The TARDIS Framework : Intelligent Virtual Agents for Social Coaching in Job Interviews. pages 476–491.
- [Arevalo et al., 2020] Arevalo, J., Solorio, T., Montes-y Gómez, M., and González, F. A. (2020). Gated multimodal networks. *Neural Computing and Applications*, 32(14) :10209–10228.
- [Bahng et al., 2020] Bahng, H., Chung, S., Yoo, S., and Choo, J. (2020). Exploring Unlabeled Faces for Novel Attribute Discovery. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5820–5829. IEEE.

-
- [Baltrusaitis et al., 2019] Baltrusaitis, T., Ahuja, C., and Morency, L. P. (2019). Multi-modal Machine Learning : A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2) :423–443.
- [Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. P. (2018). OpenFace 2.0 : Facial behavior analysis toolkit. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 59–66.
- [Bambach et al., 2015] Bambach, S., Lee, S., Crandall, D. J., and Yu, C. (2015). Lending A Hand : Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1957. IEEE.
- [Barrick et al., 2009] Barrick, M. R., Shaffer, J. A., and DeGrassi, S. W. (2009). What You See May Not Be What You Get : Relationships Among Self-Presentation Tactics and Ratings of Interview and Job Performance. *Journal of Applied Psychology*, 94(6) :1394–1411.
- [Basch and Melchers, 2019] Basch, J. and Melchers, K. (2019). Fair and Flexible?! Explanations Can Improve Applicant Reactions Toward Asynchronous Video Interviews. *Personnel Assessment and Decisions*, 5(3).
- [Basch et al., 2020] Basch, J. M., Melchers, K. G., Kegelmann, J., and Lieb, L. (2020). Smile for the camera! The role of social presence and impression management in perceptions of technology-mediated interviews. *Journal of Managerial Psychology*, 35(4) :285–299.
- [Batinca et al., 2013] Batinca, L., Stratou, G., Shapiro, A., Morency, L. P., and Scherer, S. (2013). Cicero - Towards a multimodal virtual audience platform for public speaking training. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8108 LNAI :116–128.
- [Ben-younes et al., 2017] Ben-younes, H., Cadene, R., Cord, M., and Thome, N. (2017). MUTAN : Multimodal Tucker Fusion for Visual Question Answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2631–2639. IEEE.
- [Boersma and van Heuven, 2001] Boersma, P. and van Heuven, V. (2001). Speak and unSpeak with Praat. *Glott International*, 5(9-10) :341–347.
- [Brenner et al., 2016] Brenner, F. S., Ortner, T. M., and Fay, D. (2016). Asynchronous video interviewing as a new technology in personnel selection : The applicant’s point of view. *Frontiers in Psychology*, 7(JUN) :1–11.

- [Brenner and DeLamater, 2016] Brenner, P. S. and DeLamater, J. (2016). Lies, Damned Lies, and Survey Self-Reports? Identity as a Cause of Measurement Bias. *Social Psychology Quarterly*, 79(4) :333–354.
- [Buehl et al., 2019] Buehl, A. K., Melchers, K. G., Macan, T., and Kühnel, J. (2019). Tell Me Sweet Little Lies : How Does Faking in Interviews Affect Interview Scores and Interview Validity? *Journal of Business and Psychology*, 34(1).
- [Cahya et al., 2019] Cahya, D. E., Ramakrishnan, R., and Giuliani, M. (2019). Static and Temporal Differences in Social Signals Between Error-Free and Erroneous Situations in Human-Robot Collaboration. In *Social Robotics - 11th International Conference, ICSR 2019, Madrid, Spain, November 26-29, 2019, Proceedings*, pages 189–199.
- [Calmon et al., 2017] Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips) :3993–4002.
- [Campion et al., 1997] Campion, M. A., Palmer, D. K., and Campion, J. E. (1997). A review of Structure in the Selection Interview.
- [Cao et al., 2019] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S.-E., and Sheikh, Y. A. (2019). OpenPose : Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.
- [Carney et al., 2007] Carney, D. R., Colvin, C. R., and Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5) :1054–1072.
- [Cassell et al., 2001] Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., and Rich, C. (2001). Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, pages 114–123, Morristown, NJ, USA. Association for Computational Linguistics.
- [Chen et al., 2018] Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. (2018). This looks like that : Deep learning for interpretable image recognition. *arXiv*, (NeurIPS) :1–12.
- [Chen et al., 2016a] Chen, L., Feng, G., Leong, C. W., Lehman, B., Martin-Raugh, M., Kell, H., Lee, C. M., and Yoon, S.-Y. (2016a). Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, number October, pages 161–168, New York, New York, USA. ACM Press.

-
- [Chen et al., 2016b] Chen, L., Feng, G., Martin-Raugh, M., Leong, C. W., Kitchen, C., Yoon, S. Y., Lehman, B., Kell, H., and Lee, C. M. (2016b). Automatic scoring of monologue video interviews using multimodal cues. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 08-12-Sept(September) :32–36.
- [Chen et al., 2017] Chen, L., Zhao, R., Leong, C. W., Lehman, B., Feng, G., and Hoque, M. E. (2017). Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 504–509. IEEE.
- [Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). XGBoost : A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August :785–794.
- [Cho et al., 2014] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*.
- [Chollet and Scherer, 2017] Chollet, M. and Scherer, S. (2017). Assessing Public Speaking Ability from Thin Slices of Behavior. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, pages 310–316.
- [Clavel and Richard, 2011] Clavel, C. and Richard, G. (2011). Recognition of acoustic emotion. In *Emotion-Oriented Systems*, chapter 5, pages 139–167.
- [Corbett-Davies and Goel, 2018] Corbett-Davies, S. and Goel, S. (2018). The Measure and Mismeasure of Fairness : A Critical Review of Fair Machine Learning. (Ec).
- [Degroot and Gooty, 2009] Degroot, T. and Gooty, J. (2009). Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of Business and Psychology*, 24(2) :179–192.
- [Delobelle et al., 2020] Delobelle, P., Temple, P., Perrouin, G., Frénay, B., Heymans, P., and Berendt, B. (2020). Ethical Adversaries : Towards Mitigating Unfairness with Adversarial Machine Learning. pages 1–17.
- [Deng et al., 2019] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace : Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June :4685–4694.
-

- [Devlin et al., 2019] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies - Proceedings of the Conference*, 1(Mlm) :4171–4186.
- [Dibia, 2017] Dibia, V. (2017). HandTrack : A Library For Prototyping Real-time Hand Tracking Interfaces using Convolutional Neural Networks.
- [Doughty et al., 2018] Doughty, H., Mayol-Cuevas, W., and Damen, D. (2018). The Pros and Cons : Rank-aware Temporal Attention for Skill Determination in Long Videos. *arXiv*.
- [Ekman et al., 1990] Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). The Duchenne Smile : Emotional Expression and Brain Physiology II. *Journal of Personality and Social Psychology*, 58(2) :342–353.
- [Elazar and Goldberg, 2018] Elazar, Y. and Goldberg, Y. (2018). Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Escalante et al., 2020] Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Gucluturk, Y., Guclu, U., Baro, X., Guyon, I., Jacques, J. C., Madadi, M., Ayache, S., Viegas, E., Gurpinar, F., Wicaksana, A. S., Liem, C., Van Gerven, M. A., and Van Lier, R. (2020). Modeling, Recognizing, and Explaining Apparent Personality from Videos. *IEEE Transactions on Affective Computing*, 3045(c).
- [Ester et al., 1996] Ester, M., Kriegel, H.-p., Xu, X., and Miinchen, D. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD : Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.
- [Eyben et al., 2016] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2) :190–202.
- [Eyben et al., 2013a] Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013a). Recent developments in openSMILE, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, (May) :835–838.

- [Eyben et al., 2013b] Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013b). Recent developments in openSMILE, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, (May) :835–838.
- [Fauconnier, 2015] Fauconnier, J.-P. (2015). French Word Embeddings.
- [Feiler and Powell, 2015] Feiler, A. R. and Powell, D. M. (2015). Behavioral Expression of Job Interview Anxiety.
- [Feiler and Powell, 2016] Feiler, A. R. and Powell, D. M. (2016). Behavioral Expression of Job Interview Anxiety. *Journal of Business and Psychology*, 31(1) :155–171.
- [Finnerty et al., 2016] Finnerty, A. N., Muralidhar, S., Nguyen, L. S., Pianesi, F., and Gatica-Perez, D. (2016). Stressful first impressions in job interviews. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, (October) :325–332.
- [Forbes and Jackson, 1980] Forbes, R. J. and Jackson, P. R. (1980). Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology*, 53(1) :65–72.
- [Frauendorfer and Mast, 2015] Frauendorfer, D. and Mast, M. S. (2015). The Impact of Nonverbal Behavior in the Job Interview. In *The Social Psychology of Nonverbal Communication*, pages 220–247. Palgrave Macmillan UK, London.
- [Freitag et al., 2017] Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., and Schuller, B. (2017). auDeep : Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks.
- [Galassi et al., 2020] Galassi, A., Lippi, M., and Torroni, P. (2020). Attention in Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–18.
- [Garcia et al., 2019a] Garcia, A., Colombo, P., Essid, S., D’Alché-Buc, F., and Clavel, C. (2019a). From the Token to the Review : A Hierarchical Multimodal approach to Opinion Mining.
- [Garcia et al., 2019b] Garcia, N. C., Morerio, P., and Murino, V. (2019b). *Cross-modal Learning by Hallucinating Missing Modalities in RGB-D Vision*. Elsevier Inc.
- [Gavand, 2013] Gavand, A. (2013). *Le recrutement : enjeux, outils, meilleures pratiques et nouveaux standards*. Eyrolles edition.
- [Giannakopoulos, 2015] Giannakopoulos, T. (2015). PyAudioAnalysis : An open-source python library for audio signal analysis. *PLoS ONE*, 10(12) :1–17.

- [Gifford et al., 1985] Gifford, R., Ng, C. F., and Wilkinson, M. (1985). Nonverbal cues in the employment interview : Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*, 70(4) :729–736.
- [Gilpin et al., 2019] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2019). Explaining explanations : An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89.
- [Goodrich, 1979] Goodrich, W. (1979). Face-to-Face Interaction : Research, Methods, and Theory. *Family Process*, 18(3) :355–356.
- [Gorman et al., 2018] Gorman, C. A., Robinson, J., and Gamble, J. S. (2018). An investigation into the validity of asynchronous web-based video employment-interview ratings. *Consulting Psychology Journal*, 70(2) :129–146.
- [Gosling et al., 2002] Gosling, S. D., Ko, S., Mannarelli, T., and Morris, M. E. (2002). A room with a cue : Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, 82(3) :379–398.
- [Gu et al., 2018] Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., and Marsic, I. (2018). Multi-modal affective analysis using hierarchical attention strategy with word-level alignment. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1 :2225–2235.
- [Guchait et al., 2014] Guchait, P., Ruetzler, T., Taylor, J., and Toldi, N. (2014). Video interviewing : A potential selection tool for hospitality managers - A study to understand applicant perspective. *International Journal of Hospitality Management*, 36 :90–100.
- [Hamdani et al., 2014] Hamdani, M. R., Valcea, S., and Buckley, M. R. (2014). The relentless pursuit of construct validity in the design of employment interviews. *Human Resource Management Review*, 24(2) :160–176.
- [Hemamou et al., 2019a] Hemamou, L., Felhi, G., Martin, J.-c., and Clavel, C. (2019a). Slices of Attention in Asynchronous Video Job Interviews. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.
- [Hemamou et al., 2019b] Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J.-c., and Clavel, C. (2019b). HireNet : A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 :573–581.

-
- [Hemamou et al., 2020] Hemamou, L., Guillon, A., Martin, J.-c., and Clavel, C. (2020). Attention Slices dans les Entretiens d'Embauche Vidéo Différés. In *Workshop sur les "Affects, Compagnons Artificiels et Interactions" (ACAI)*, pages 1–9.
- [Hoque et al., 2013] Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., and Picard, R. W. (2013). MACH. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13*, page 697, New York, New York, USA. ACM Press.
- [Huffcutt, 2011] Huffcutt, A. I. (2011). An Empirical Review of the Employment Interview Construct Literature. 19(1).
- [Huffcutt et al., 2001] Huffcutt, A. I., Conway, J. M., Roth, P. L., and Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, 86(5) :897–913.
- [Huffcutt et al., 2013] Huffcutt, A. I., Culbertson, S. S., and Weyhrauch, W. S. (2013). Employment interview reliability : New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, 21(3) :264–276.
- [Huffcutt et al., 2011] Huffcutt, A. I., Van Iddekinge, C. H., and Roth, P. L. (2011). Understanding applicant behavior in employment interviews : A theoretical model of interviewee performance. *Human Resource Management Review*, 21(4) :353–367.
- [Hutchinson and Mitchell, 2019] Hutchinson, B. and Mitchell, M. (2019). 50 Years of Test (Un)fairness : Lessons for machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 49–58.
- [Jain and Wallace, 2019] Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. In *North American Chapter of the Association for Computational Linguistics*.
- [Jaiswal and Mower Provost, 2020] Jaiswal, M. and Mower Provost, E. (2020). Privacy Enhanced Multimodal Neural Representations for Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05) :7985–7993.
- [Janssoone et al., 2016] Janssoone, T., Clavel, C., Bailly, K., and Richard, G. (2016). Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. *Proceedings of International Conference on Intelligent Virtual Agents*.
- [Jetley et al., 2018] Jetley, S., Lord, N. A., Lee, N., and Torr, P. H. S. (2018). Learn To Pay Attention. In *International Conference on Learning Representations*, pages 1–14.
- [Kroll and Ziegler, 2016] Kroll, E. and Ziegler, M. (2016). Discrimination due to Ethnicity and Gender : How susceptible are video-based job interviews? *International Journal of Selection and Assessment*, 24(2) :161–171.
-

- [Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *NAACL HLT*, 9(2) :29–30.
- [Leong et al., 2019] Leong, C. W., Roohr, K., Ramanarayanan, V., Martin-Raugh, M. P., Kell, H., Ubale, R., Qian, Y., Mladineo, Z., and McCulla, L. (2019). To Trust, or Not to Trust? A Study of Human Bias in Automated Video Interview Assessments.
- [Levashina et al., 2014] Levashina, J., Hartwell, C. J., Morgeson, F. P., and Campion, M. A. (2014). The Structured Employment Interview : Narrative and Quantitative Review of the Research Literature. *Personnel Psychology*, 67(1) :241–293.
- [Li et al., 2018] Li, K., Wu, Z., Peng, K.-c., Ernst, J., and Fu, Y. (2018). Tell Me Where to Look : Guided Attention Inference Network. In *Conference on Computer Vision and Pattern Recognition*.
- [Liang et al., 2018] Liang, P. P., Liu, Z., Bagher Zadeh, A., and Morency, L.-P. (2018). Multimodal Language Analysis with Recurrent Multistage Fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Lieberman, 2001] Lieberman, R. (2001). A Tale of Two Countries : The Politics of Color Blindness in France and the United States. *French Politics, Culture & Society*, 19(3).
- [Lipton et al., 2018] Lipton, Z. C., Chouldechova, A., and McAuley, J. (2018). Does mitigating ML’s impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, 2018-Decem(ML) :8125–8135.
- [Littlewort et al., 2011] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M. (2011). The computer expression recognition toolbox (CERT). In *Face and Gesture 2011*, pages 298–305. IEEE.
- [Liu et al., 2017] Liu, P., Qiu, X., and Huang, X. (2017). Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Liu et al., 2019] Liu, X., Wang, X., and Matwin, S. (2019). Improving the interpretability of deep neural networks with knowledge distillation. *IEEE International Conference on Data Mining Workshops, ICDMW*, 2018-Novem :905–912.
- [Long et al., 2019] Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., and Mei, T. (2019). Gaussian temporal awareness networks for action localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June :344–353.

-
- [Louizos et al., 2015] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015). The Variational Fair Autoencoder. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pages 1–11.
- [Louppe et al., 2017] Louppe, G., Kagan, M., and Cranmer, K. (2017). Learning to pivot with adversarial networks. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips) :982–991.
- [Lukacik et al., 2020] Lukacik, E. R., Bourdage, J. S., and Roulin, N. (2020). Into the void : A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, (September 2019) :100789.
- [Madras et al., 2018] Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. *35th International Conference on Machine Learning, ICML 2018*, 8 :5423–5434.
- [Mairesse et al., 2007] Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30 :457–500.
- [Martin et al., 2020] Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, , Seddah, D., and Sagot, B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 7203–7219, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [McCarthy et al., 2017] McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., and Ahmed, S. M. (2017). Applicant Perspectives During Selection : A Review Addressing “So What?,” “What’s New?,” and “Where to Next?”. *Journal of Management*, 43(6) :1693–1725.
- [McClave, 2000] McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32(7) :855–878.
- [McDuff et al., 2010] McDuff, D., El Kaliouby, R., Kassam, K., and Picard, R. (2010). Affect valence inference from facial action unit spectrograms. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 17–24. IEEE.
- [McInnes et al., 2018] McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP : Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29) :861.
-

- [Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26 :3111–3119.
- [Miller, 2019] Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, 267 :1–38.
- [Muralidhar and Gatica-perez, 2017] Muralidhar, S. and Gatica-perez, D. (2017). Examining Linguistic Content and Skill Impression Structure for Job Interview Analytics in Hospitality. pages 339–343.
- [Muralidhar et al., 2018] Muralidhar, S., Nguyen, L., and Gatica-Perez, D. (2018). Words Worth : Verbal Content and Hirability Impressions in YouTube Video Resumes. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 322–327, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Muralidhar et al., 2016] Muralidhar, S., Nguyen, L. S., Frauendorfer, D., Odobez, J.-M., Schmid Mast, M., and Gatica-Perez, D. (2016). Training on the job : behavioral analysis of job interviews in hospitality. *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, pages 84–91.
- [Muralidhar et al., 2020] Muralidhar, S., Patricia, E., Mayor, E., Bangerter, A., Mast, M. S., and Gatica-perez, D. (2020). Understanding Applicants ’ Reactions to Asynchronous Video Interviews Through Self-reports and Nonverbal Cues.
- [Muralidhar Idiap et al., 2018] Muralidhar Idiap, S., Switzerland smuralidhar, E., Odobez Idiap, J.-M., and Switzerland odobez, E. (2018). Facing Employers and Customers : What Do Gaze and Expressions Tell About Soft Skills ?
- [Murdoch et al., 2018] Murdoch, W. J., Liu, P. J., and Yu, B. (2018). Beyond Word Importance : Contextual Decomposition to Extract Interactions from LSTMs.
- [Murphy et al., 2015] Murphy, N. A., Hall, J. A., Schmid Mast, M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., Roter, D. L., and Nguyen, L. (2015). Reliability and Validity of Nonverbal Thin Slices in Social Interactions. *Personality and Social Psychology Bulletin*, 41(2) :199–213.
- [Naim et al., 2018] Naim, I., Tanveer, M. I., Gildea, D., and Hoque, M. E. (2018). Automated Analysis and Prediction of Job Interview Performance. *IEEE Transactions on Affective Computing*, 9(2) :191–204.
- [Narayanan, 2018] Narayanan, A. (2018). Tutorial : 21 Fairness Definitions and their Politics. In *Conference on Fairness, Accountability, and Transparency*.

-
- [New, 2006] New, B. (2006). Lexique 3 : Une nouvelle base de données lexicales. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006), avril 2006, Louvain, Belgique*.
- [Nguyen, 2015] Nguyen, L. S. (2015). Computational Analysis Of Behavior In Employment Interviews And Video Resumes. 6567(January 2015) :1–158.
- [Nguyen et al., 2014] Nguyen, L. S., Frauendorfer, D., Mast, M. S., and Gatica-Perez, D. (2014). Hire me : Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Transactions on Multimedia*, 16(4) :1018–1031.
- [Nguyen and Gatica-Perez, 2015] Nguyen, L. S. and Gatica-Perez, D. (2015). I Would Hire You in a Minute. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pages 51–58, New York, New York, USA. ACM Press.
- [Nguyen and Gatica-Perez, 2016] Nguyen, L. S. and Gatica-Perez, D. (2016). Hirability in the Wild : Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia*, 18(7) :1422–1437.
- [Palaskar et al., 2019] Palaskar, S., Libovický, J., Gella, S., and Metze, F. (2019). Multi-modal Abstractive Summarization for How2 Videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Pascanu et al., 2020] Pascanu, S., Jayakumar, M., and Wojciech M. Czarnecki and Jacob Menick and Jonathan Schwarz and Jack Rae and Simon Osindero and Yee Whye Teh and Tim Harley and Razvan (2020). Multiplicative Interactions and Where to Find Them. In *ICLR*.
- [Peeters and Lievens, 2006] Peeters, H. and Lievens, F. (2006). Verbal and nonverbal impression management tactics in behavior description and situational interviews. *International Journal of Selection and Assessment*, 14(3) :206–222.
- [Pentland, 2004] Pentland, A. (2004). Social dynamics : Signals and behavior. In *International Conference on Developmental Learning.*, volume 5.
- [Peters et al., 2018] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, Stroudsburg, PA, USA. Association for Computational Linguistics.
-

- [Pham et al., 2019] Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., and Póczos, B. (2019). Found in Translation : Learning Robust Joint Representations by Cyclic Translations between Modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 :6892–6899.
- [Piolat et al., 2011] Piolat, A., Booth, R. J., Chung, C. K., Davids, M., and Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation. *Psychologie Française*, 56(3) :145–159.
- [Poh and San, 2015] Poh, W. Y. F. and San (2015). EVALUATING CANDIDATE PERFORMANCE AND REACTION IN ONE-WAY VIDEO INTERVIEWS. (May).
- [Poncelet, 2016] Poncelet, P. (2016). FEEL : a French Expanded Emotion Lexicon.
- [Poria et al., 2017] Poria, S., Mazumder, N., Cambria, E., Hazarika, D., Morency, L. P., and Zadeh, A. (2017). Context-dependent sentiment analysis in user-generated videos. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1 :873–883.
- [Posthuma et al., 2002] Posthuma, R. A., Morgeson, F. P., and Campion, M. A. (2002). Beyond employment interview validity : A comprehensive narrative review of recent research and trends over time. *Personnel Psychology*, 55(1) :1–81.
- [Raghavan et al., 2019] Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2019). Mitigating Bias in Algorithmic Hiring : Evaluating Claims and Practices.
- [Raji et al., 2020] Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020). Saving Face : Investigating the ethical concerns of facial recognition auditing. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151.
- [Rao S. B et al., 2017] Rao S. B, P., Rasipuram, S., Das, R., and Jayagopi, D. B. (2017). Automatic assessment of communication skill in non-conventional interview settings : a comparative study. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017*, number November, pages 221–229, New York, New York, USA. ACM Press.
- [Rasipuram et al., 2017a] Rasipuram, S., Das, R., Rao, P., and Jayagopi, D. B. (2017a). Online Peer-to-peer Discussions : A Platform for Automatic Assessment of Communication Skill. pages 1–6.
- [Rasipuram and Jayagopi, 2016] Rasipuram, S. and Jayagopi, D. B. (2016). Automatic assessment of communication skill in interface-based employment interviews using audio-

- visual cues. *2016 IEEE International Conference on Multimedia and Expo Workshop, ICMEW 2016*, (September).
- [Rasipuram and Jayagopi, 2018] Rasipuram, S. and Jayagopi, D. B. (2018). Automatic assessment of communication skill in interview-based interactions. *Multimedia Tools and Applications*, 77(14) :18709–18739.
- [Rasipuram et al., 2017b] Rasipuram, S., Rao, S. B., and Jayagopi, D. B. (2017b). Automatic prediction of fluency in interface-based interviews. *2016 IEEE Annual India Conference, INDICON 2016*, (December).
- [Ribeiro et al., 2016] Ribeiro, M., Singh, S., and Guestrin, C. (2016). “Why Should I Trust You?” : Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, pages 97–101, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Roulin et al., 2014] Roulin, N., Bangerter, A., and Levashina, J. (2014). Interviewers’ perceptions of impression management in employment interviews. *Journal of Managerial Psychology*, 29(2) :141–163.
- [Roulin et al., 2015] Roulin, N., Bangerter, A., and Levashina, J. (2015). Honest and Deceptive Impression Management in the Employment Interview : Can It Be Detected and How Does It Impact Evaluations? *Personnel Psychology*, 68(2) :395–444.
- [Ruben et al., 2015] Ruben, M. A., Hall, J. A., and Schmid Mast, M. (2015). Smiling in a job interview : When less is more. *Journal of Social Psychology*, 155(2) :107–126.
- [Rupasinghe et al., 2017] Rupasinghe, A. T., Gunawardena, N. L., Shujan, S., and Atukorale, D. A. (2017). Scaling personality traits of interviewees in an online job interview by vocal spectrum and facial cue analysis. *16th International Conference on Advances in ICT for Emerging Regions, ICTer 2016 - Conference Proceedings*, (September) :288–295.
- [Ryoo et al., 2015] Ryoo, M. S., Rothrock, B., and Matthies, L. (2015). Pooled motion features for first-person videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June(Figure 1) :896–904.
- [Salgado, 2017] Salgado, J. F. (2017). Personnel Selection. In *Oxford Research Encyclopedia of Psychology*, volume 1. Oxford University Press.
- [Sánchez-monedero and Dencik, 2019] Sánchez-monedero, J. and Dencik, L. (2019). The datafication of the workplace. *Data Justice Project, Cardiff University*, pages 1–46.

- [Sánchez-Monedero et al., 2020] Sánchez-Monedero, J., Dencik, L., and Edwards, L. (2020). What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 458–468.
- [Schmid, 1994] Schmid, H. I.-C. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *European Conference on Machine Learning*.
- [Schmidt and Hunter, 1998] Schmidt, F. L. and Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology : Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2) :262–274.
- [Schmitt and Schuller, 2017] Schmitt, M. and Schuller, B. (2017). openXBOW – Introducing the passau open-source crossmodal bag-of-words toolkit. *Journal of Machine Learning Research*, 18 :1–5.
- [Schmitt, 2012] Schmitt, N., editor (2012). *The Oxford Handbook of Personnel Assessment and Selection*. Oxford University Press.
- [Schneider et al., 2019] Schneider, L., Powell, D. M., and Bonaccio, S. (2019). Does interview anxiety predict job performance and does it influence the predictive validity of interviews? *International Journal of Selection and Assessment*, (April) :1–9.
- [Schneider et al., 2015] Schneider, L., Powell, D. M., and Roulin, N. (2015). Cues to deception in the employment interview. *International Journal of Selection and Assessment*, 23(2) :182–190.
- [Schuller et al., 2013] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge : Social signals, conflict, emotion, autism. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August) :148–152.
- [Shin et al., 2018] Shin, J., Madotto, A., and Fung, P. (2018). Interpreting Word Embeddings with Eigenvector Analysis. *NIPS 2018 Interpretability and Robustness in Audio, Speech, and Language (IRASL) Workshop : 32nd Annual Conference on Neural Information Processing Systems*, (Nips).
- [Singhania et al., 2020] Singhania, A., Unnam, A., and Aggarwal, V. (2020). Grading video interviews with fairness considerations.

- [Srivastava et al., 2019] Srivastava, B. M. L., Bellet, A., Tommasi, M., and Vincent, E. (2019). Privacy-Preserving Adversarial Representation Learning in ASR : Reality or Illusion? In *Interspeech 2019*, volume 2019-Septe, pages 3700–3704, ISCA. ISCA.
- [Stöckli et al., 2018] Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., and Samson, A. C. (2018). Facial expression analysis with AFFDEX and FACET : A validation study. *Behavior Research Methods*, 50(4) :1446–1460.
- [Streiff-Fénart, 2012] Streiff-Fénart, J. (2012). A French dilemma : Anti-discrimination policies and minority claims in contemporary France. *Comparative European Politics*, 10(3) :283–300.
- [Suen et al., 2019a] Suen, H. Y., Chen, M. Y. C., and Lu, S. H. (2019a). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, 98(43) :93–101.
- [Suen et al., 2019b] Suen, H.-y., Hung, K.-e., and Lin, C.-l. (2019b). TensorFlow-Based Automatic Personality Recognition Used in Asynchronous Video Interviews. *IEEE Access*, 7(c) :61018–61023.
- [Swider et al., 2016] Swider, B. W., Barrick, M. R., and Brad Harris, T. (2016). Initial impressions : What they are, what they are not, and how they influence structured interview outcomes. *Journal of Applied Psychology*, 101(5) :625–638.
- [Tanaka et al., 2015] Tanaka, H., Sakti, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H., and Nakamura, S. (2015). Automated Social Skills Trainer. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 17–27.
- [Tanveer et al., 2015] Tanveer, M. I., Lin, E., and Hoque, M. E. (2015). Rhema. In *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, pages 286–295, New York, New York, USA. ACM Press.
- [Tanveer et al., 2016] Tanveer, M. I., Zhao, R., Chen, K., Tiet, Z., and Hoque, M. E. (2016). AutoManner : An Automated Interface for Making Public Speakers Aware of Their Mannerisms. *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 385–396.
- [Torres and Gregory, 2018] Torres, E. N. and Gregory, A. (2018). Hiring manager’s evaluations of asynchronous video interviews : The role of candidate competencies, aesthetics, and resume placement. *International Journal of Hospitality Management*, 75(April) :86–93.

- [Torres and Mejia, 2017] Torres, E. N. and Mejia, C. (2017). Asynchronous video interviews in the hospitality industry : Considerations for virtual employee selection. *International Journal of Hospitality Management*, 61 :4–13.
- [Trigeorgis et al., 2016] Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016-May :5200–5204.
- [Tripathy et al., 2019] Tripathy, A., Wang, Y., and Ishwar, P. (2019). Privacy-Preserving Adversarial Networks. *2019 57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019*, pages 495–505.
- [Tsai et al., 2019] Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Varni et al., 2018] Varni, G., Hupont, I., Clavel, C., and Chetouani, M. (2018). Computational Study of Primitive Emotional Contagion in Dyadic Interactions. *IEEE Transactions on Affective Computing*, 3045(c) :1–1.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. (Nips).
- [Voskuijl, 2017] Voskuijl, O. F. (2017). Job Analysis : Current and Future Perspectives. In *The Blackwell Handbook of Personnel Selection*, pages 25–46. Blackwell Publishing Ltd, Oxford, UK.
- [Wagner et al., 2013] Wagner, J., Lingensfelder, F., Baur, T., Damian, I., Kistler, F., and André, E. (2013). The social signal interpretation (SSI) framework. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, pages 831–834, New York, New York, USA. ACM Press.
- [Wang et al., 2019a] Wang, H., Wu, Z., Wang, Z., Wang, Z., and Jin, H. (2019a). Privacy-Preserving Deep Visual Recognition : An Adversarial Learning Framework and A New Dataset. pages 1–14.
- [Wang et al., 2019b] Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., and Morency, L.-P. (2019b). Words Can Shift : Dynamically Adjusting Word Representations Using Nonverbal Behaviors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 :7216–7223.

-
- [Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. V., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers : State-of-the-Art Natural Language Processing.
- [Wolf et al., 2019] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). HuggingFace’s Transformers : State-of-the-art Natural Language Processing.
- [Yang et al., 2019a] Yang, B., Li, J., Wong, D. F., Chao, L. S., Wang, X., and Tu, Z. (2019a). Context-Aware Self-Attention Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 :387–394.
- [Yang et al., 2019b] Yang, F., Du, M., and Hu, X. (2019b). Evaluating Explanation Without Ground Truth in Interpretable Machine Learning.
- [Yang et al.,] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. Hierarchical Attention Networks for Document Classification. Technical report.
- [Yang et al., 2016] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1480–1489.
- [Yosinski et al., 2015] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding Neural Networks Through Deep Visualization. In *In ICML Workshop on Deep Learning*.
- [Yu et al., 2017] Yu, H., Gui, L., Madaio, M., Ogan, A., Cassell, J., and Morency, L.-P. (2017). Temporally Selective Attention Model for Social and Affective State Recognition in Multimedia Content. *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1743–1751.
- [Zadeh et al., 2018a] Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., and Morency, L.-P. (2018a). Multi-attention Recurrent Network for Human Communication Comprehension. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.
- [Zadeh et al., 2019] Zadeh, A., Mao, C., Shi, K., Zhang, Y., Liang, P., Poria, S., and Morency, L. P. (2019). Factorized multimodal transformer for multimodal sequential learning. *arXiv*, pages 1–13.
-

- [Zadeh et al., 2018b] Zadeh, A., Poria, S., Liang, P. P., Cambria, E., Mazumder, N., and Morency, L. P. (2018b). Memory fusion network for multi-view sequential learning. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 5634–5641.
- [Zemel et al., 2013] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, 28(PART 2) :1362–1370.
- [Zhang et al., 2019] Zhang, D., Wu, L., Li, S., Zhu, Q., and Zhou, G. (2019). Multi-Modal Language Analysis with Hierarchical Interaction-Level and Selection-Level Attentions. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 724–729. IEEE.
- [Zhao et al., 2017] Zhao, R. U., Li, V., Barbosa, H., Ghoshal, G., and Ehsan, M. (2017). Semi-Automated & Collaborative Online Training Module for Improving Communication Skills. 1(2) :1–20.
- [Zibarras et al., 2018] Zibarras, L., Patterson, F., Holmes, J., Flaxman, C., and Kubacki, A. (2018). An exploration of applicant perceptions of asynchronous video MMIs in medical selection. *MedEdPublish*, 7(4) :1–16.

Annexes

Annexe A

Liste des positions références choisies dans le référentiel ROME pour la sélection des postes dans la base de données EASYRECRUE

- Conseil clientèle en assurances
- Agent commercial / Agente commerciale en assurances
- Attaché / Attachée de clientèle en assurances
- Chargé / Chargée de clientèle centre d'appels en assurances
- Chargé / Chargée de clientèle en assurances
- Collaborateur commercial / Collaboratrice commerciale en assurances
- Conseiller / Conseillère clientèle en assurances
- Conseiller / Conseillère en assurance produits d'épargne
- Conseiller / Conseillère en assurance vie
- Conseiller / Conseillère en assurances
- Conseiller / Conseillère gestion sinistres
- Conseiller / Conseillère Incendie, Accidents, Risques Divers
- Conseiller / Conseillère mutualiste
- Conseiller / Conseillère mutualiste Incendie, Accidents, Risques Divers
- Conseiller / Conseillère prévoyance santé
- Conseiller commercial / Conseillère commerciale en assurances
- Technico-commercial / Technico-commerciale en assurances
- Téléconseiller / Téléconseillère en assurances
- Assistant / Assistante de clientèle de banque
- Assistant / Assistante service clientèle bancaire
- Assistant / Assistante service clientèle de banque
- Caissier / Caissière de bureau de change
- Chargé / Chargée d'accueil en banque
- Chargé / Chargée d'accueil et de services clientèle bancaire
- Conseiller / Conseillère accueil en agence bancaire
- Conseiller / Conseillère d'accueil en banque
- Guichetier / Guichetière accueil banque
- Guichetier / Guichetière de banque
- Guichetier / Guichetière de la banque postale
- Guichetier payeur / Guichetière payeuse
- Guichetier vendeur / Guichetière vendeuse
- Téléconseiller / Téléconseillère en banque
- Attaché commercial / Attachée commerciale bancaire entreprise
- Attaché commercial / Attachée commerciale bancaire financements spécialisés
- Attaché commercial / Attachée commerciale banque d'affaires
- Chargé / Chargée d'affaires bancaires commerce international
- Chargé / Chargée d'affaires bancaires entreprise
- Chargé / Chargée d'affaires bancaires professionnelles
- Chargé / Chargée de clientèle bancaire grandes entreprises
- Chargé / Chargée de comptes bancaires professionnels
- Chargé / Chargée de développement clientèle bancaire entreprise
- Responsable clientèle bancaire entreprise
- Responsable grands comptes bancaires
- Conseiller / Conseillère en développement de patrimoine
- Conseiller / Conseillère en gestion de capitaux
- Conseiller / Conseillère en gestion de fortune
- Conseiller / Conseillère en gestion de patrimoine financier
- Conseiller / Conseillère en investissements financiers
- Conseiller / Conseillère en investissements privés
- Conseiller / Conseillère en patrimoine financier
- Conseiller / Conseillère gestion banque privée
- Gestionnaire de fortune
- Gestionnaire de patrimoine financier
- Attaché / Attachée de clientèle de banque
- Chargé / Chargée de clientèle bancaire
- Chargé / Chargée de clientèle commerciale de banque

-
- Chargé / Chargée de clientèle de banque
 - Chargé / Chargée de clientèle entreprises de banque
 - Chargé / Chargée de clientèle particuliers de banque
 - Chargé / Chargée de clientèle professionnelle de banque
 - Chargé / Chargée de clientèle rachat de crédits
 - Chargé / Chargée de gestion bancaire
 - Chargé / Chargée de relations clientèle bancaire
 - Conseiller / Conseillère de clientèle bancaire
 - Conseiller / Conseillère en crédit immobilier
 - Conseiller / Conseillère en produit épargne
 - Conseiller commercial professionnel / Conseillère commerciale professionnelle secteur bancaire
 - Conseiller financier / Conseillère financière banque postale
 - Conseiller financier / Conseillère financière clientèle professionnelle
 - Courtier / Courtière en prêts immobiliers
 - Gestionnaire de clientèle bancaire
 - Agent / Agente d'accueil de prestations sociales
 - Agent / Agente technique de banque
 - Agent / Agente technique des régimes de retraite complémentaire et de prévoyance
 - Agent administratif / Agente administrative back-office marché
 - Agent administratif / Agente administrative d'assurances
 - Agent administratif / Agente administrative de banque
 - Agent administratif / Agente administrative des opérations bancaires
 - Agent administratif / Agente administrative middle-office marché
 - Chargé / Chargée de clientèle principal / principale en immobilier
 - Chef d'agence locatif immobilier
 - Agent / Agente de location immobilière
 - Agent / Agente immobilier
 - Agent commercial / Agente commerciale en immobilier
 - Assistant commercial / Assistante commerciale en immobilier
 - Attaché commercial / Attachée commerciale en immobilier
 - Chasseur / Chasseuse immobilier
 - Conseiller / Conseillère de location en immobilier
 - Conseiller / Conseillère de transaction en immobilier
 - Conseiller / Conseillère de vente en immobilier
 - Conseiller / Conseillère en immobilier d'entreprise
 - Conseiller / Conseillère immobilier
 - Mandataire en vente de fonds de commerce
 - Marchand / Marchande de biens immobiliers
 - Négociateur / Négociatrice en immobilier
 - Négociateur / Négociatrice en immobilier d'entreprise
 - Négociateur / Négociatrice en location immobilière
 - Négociateur / Négociatrice immobilier
 - Négociateur / Négociatrice immobilier en bureau de vente
 - Prospecteur négociateur / Prospectrice négociatrice en immobilier
 - Responsable d'agence immobilière
 - Responsable de clientèle en transaction immobilière
 - Responsable de vente immobilière
 - Vendeur / Vendeuse en immobilier neuf
 - Vendeur / Vendeuse immobilier
 - Assistant / Assistante achat
 - Assistant / Assistante administration des ventes
 - Assistant / Assistante administration des ventes export
 - Assistant / Assistante commerce international
 - Assistant / Assistante des ventes
 - Assistant / Assistante export
 - Assistant / Assistante import
 - Assistant / Assistante import-export
 - Assistant / Assistante service clients
 - Assistant administratif et commercial / Assistante administrative et commerciale
 - Assistant commercial / Assistante commerciale
 - Attaché commercial / Attachée commerciale sédentaire
 - Collaborateur commercial / Collaboratrice commerciale
 - Commercial / Commerciale sédentaire
 - Conseiller commercial / Conseillère commerciale sédentaire
 - Délégué commercial / Déléguée commerciale sédentaire
 - Employé commercial / Employée commerciale sédentaire
 - Secrétaire commercial / commerciale
 - Technicien / Technicienne administration des ventes
 - Technicien / Technicienne de la vente par correspondance
 - Attaché commercial / Attachée commerciale en biens de consommation auprès des entreprises
 - Attaché commercial / Attachée commerciale en biens d'équipement professionnels
 - Attaché commercial / Attachée commerciale en biens intermédiaires et matières premières auprès des entreprises
 - Attaché commercial / Attachée commerciale en clientèle d'entreprises
 - Attaché commercial / Attachée commerciale en fournitures industrielles auprès des entreprises
 - Attaché commercial / Attachée commerciale en matériaux industriels auprès des entreprises
 - Attaché commercial / Attachée commerciale en matériel agricole auprès des entreprises
 - Attaché commercial / Attachée commerciale en matériel de bureau auprès des entreprises
 - Attaché commercial / Attachée commerciale en services auprès des entreprises
 - Attaché commercial / Attachée commerciale en transport-logistique
 - Attaché commercial / Attachée commerciale export
 - Attaché commercial / Attachée commerciale grandes et moyennes surfaces de vente (GMS)
 - Attaché commercial / Attachée commerciale tourisme
 - Commercial / Commerciale auprès d'une clientèle d'entreprises
 - Commercial / Commerciale en biens de consommation auprès des entreprises
 - Commercial / Commerciale en biens d'équipement auprès des entreprises
 - Commercial / Commerciale en biens intermédiaires et matières premières auprès des entreprises
 - Commercial / Commerciale en produits alimentaires secs en gros
 - Commercial / Commerciale en publicité auprès des entreprises
 - Commercial / Commerciale en services auprès des entreprises
 - Commercial / Commerciale export
 - Commercial vendeur / Commerciale vendeuse d'espaces publicitaires
 - Commercial vendeur / Commerciale vendeuse d'espaces publicitaires web
 - Conseiller commercial / Conseillère commerciale auprès d'une clientèle d'entreprises
 - Conseiller commercial / Conseillère commerciale en biens de consommation auprès des entreprises
 - Conseiller commercial / Conseillère commerciale en biens d'équipement auprès des entreprises
 - Conseiller commercial / Conseillère commerciale en biens intermédiaires et matières premières auprès des entreprises
 - Conseiller commercial / Conseillère commerciale en services auprès des entreprises
-

-
- Délégué commercial / Déléguée commerciale en biens de consommation auprès des entreprises
 - Délégué commercial / Déléguée commerciale en biens d'équipement auprès des entreprises
 - Délégué commercial / Déléguée commerciale en biens intermédiaires et matières premières auprès des entreprises
 - Délégué commercial / Déléguée commerciale en services auprès des entreprises
 - Représentant / Représentante en biens de consommation auprès des entreprises
 - Représentant / Représentante en biens d'équipement auprès des entreprises
 - Représentant / Représentante en biens intermédiaires et matières premières auprès des entreprises
 - Représentant / Représentante en services auprès des entreprises
 - Attaché commercial / Attachée commerciale auprès des particuliers
 - Chargé / Chargée de recouvrement de créances
 - Chef de groupe de vendeurs à domicile
 - Commercial / Commerciale auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en distribution d'énergie auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en équipement de l'habitat auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en gastronomie univers culinaire auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en produits culturels auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en produits d'entretien auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en solutions de télécommunications auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en textile et accessoires de mode auprès des particuliers
 - Conseiller commercial / Conseillère commerciale auprès des particuliers
 - Conseiller commercial / Conseillère commerciale auprès d'une clientèle de particuliers
 - Conseiller commercial / Conseillère commerciale en aménagement de cuisines auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en articles de décoration auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en bien-être et diététique auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en cheminées auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en déménagement
 - Conseiller commercial / Conseillère commerciale en mobile homes auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en piscines auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en portes/fermetures auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en produits cosmétiques auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en ravalement de façades auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en salles de bains auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en stores auprès des particuliers
 - Conseiller commercial / Conseillère commerciale en vérandas auprès des particuliers
 - Conseiller vendeur / Conseillère vendeuse à domicile
 - Conseiller vendeur / Conseillère vendeuse en laisser sur place auprès des particuliers
 - Délégué commercial / Déléguée commerciale auprès des particuliers
 - Recouvreur / Recouvreuse de créances
 - Assistant / Assistante de vente automobile
 - Attaché commercial / Attachée commerciale en automobiles
 - Chef de groupe véhicules d'occasion
 - Chef de groupe véhicules neufs
 - Commercial / Commerciale en automobiles
 - Conseiller / Conseillère des ventes automobiles
 - Conseiller commercial / Conseillère commerciale en automobiles
 - Conseiller commercial / Conseillère commerciale en véhicules industriels
 - Conseiller commercial / Conseillère commerciale en véhicules ou bateaux
 - Conseiller commercial / Conseillère commerciale motocycles
 - Conseiller vendeur / Conseillère vendeuse d'autocars
 - Conseiller vendeur / Conseillère vendeuse de bateaux de plaisance
 - Conseiller vendeur / Conseillère vendeuse de véhicules de loisirs
 - Conseiller vendeur / Conseillère vendeuse de véhicules poids lourds
 - Délégué commercial / Déléguée commerciale en automobiles
 - Gérant / Gérante de négoce automobile
 - Négociant / Négociante de véhicules d'occasion
 - Négociant / Négociante de véhicules neufs
 - Technicien / Technicienne de la vente automobile
 - Technicien / Technicienne de la vente de motocycles
 - Vendeur / Vendeuse automobile
 - Vendeur / Vendeuse de caravanes/camping car
 - Vendeur / Vendeuse de motocycles
 - Vendeur / Vendeuse de véhicules neufs
 - Vendeur / Vendeuse en véhicules anciens
 - Vendeur / Vendeuse en véhicules de collection
 - Vendeur / Vendeuse en véhicules d'occasion
 - Vendeur / Vendeuse en véhicules industriels
 - Vendeur / Vendeuse secteur véhicules d'occasion
 - Vendeur / Vendeuse secteur véhicules neufs
 - Chargé / Chargée d'expansion commerciale d'enseigne
 - Chef de secteur des ventes
 - Chef des ventes
 - Délégué régional / Déléguée régionale des ventes
 - Directeur / Directrice des ventes
 - Directeur / Directrice des ventes internationales
 - Directeur national / Directrice nationale des ventes
 - Directeur régional / Directrice régionale des ventes
 - Directeur régional / Directrice régionale des ventes export
 - Inspecteur / Inspectrice des ventes
 - Inspecteur / Inspectrice du cadre en assurances
 - Inspecteur commercial / Inspectrice commerciale
 - Manager commercial / Manageuse commerciale des forces de vente
 - Manager commercial junior / Manageuse commerciale junior des forces de vente
 - Responsable animateur / animatrice des forces de vente
 - Responsable animateur / animatrice des ventes
 - Responsable de la force de vente
 - Responsable des ventes
 - Responsable des ventes comptes-clés
 - Responsable des ventes zone export
 - Responsable régional / régionale des ventes
 - Responsable ventes indirectes
-

-
- Chargé / Chargée d'assistance
 - Conseiller / Conseillère clientèle à distance
 - Permanencier / Permanencière auxiliaire de régulation médicale
 - Responsable de centre d'appels
 - Responsable de plateau de centre d'appels
 - Superviseur / Superviseuse de centre d'appels
 - Technicien / Technicienne de la vente à distance
 - Téléacteur / Téléactrice
 - Téléconseiller / Téléconseillère
 - Téléopérateur / Téléopératrice
 - Téléprospecteur / Téléprospectrice
 - Télévendeur / Télévendeuse
 - Caissier / Caissière
 - Caissier / Caissière de parking
 - Caissier / Caissière de station-service
 - Caissier / Caissière en libre-service
 - Caissier / Caissière en restauration
 - Caissier / Caissière pompiste
 - Caissier / Caissière roller
 - Caissier / Caissière spectacle
 - Hôte / Hôtesse de caisse
 - Hôte / Hôtesse de caisse services clients
 - Péager / Péagère
 - Péagiste
 - Agent / Agente d'accueil touristique
 - Chargé / Chargée d'accueil en réceptif local
 - Chargé / Chargée d'accueil touristique
 - Conseiller / Conseillère en séjour touristique
 - Hôte / Hôtesse d'accueil et d'animation de croisière
 - Hôte / Hôtesse d'accueil tourisme
 - Hôte animateur / Hôtesse animatrice de croisière fluviale
 - Hôte animateur / Hôtesse animatrice de croisière maritime
 - Permanent local / Permanente locale d'agence réceptive
 - Représentant local / Représentante locale d'agence réceptive
 - Technicien / Technicienne d'accueil touristique
 - Agent / Agente de réservation en hôtellerie
 - Assistant / Assistante de réception en établissement hôtelier
 - Chef de brigade de réception hôtelière
 - Chef de réception en hôtellerie
 - Employé / Employée de réception en établissement hôtelier
 - Employé / Employée de réservation en hôtellerie
 - Night audit
 - Night auditor
 - Premier / Première de réception en hôtellerie
 - Réceptionniste de camping
 - Réceptionniste de nuit
 - Réceptionniste de village vacances
 - Réceptionniste en établissement touristique
 - Réceptionniste en hôtellerie
 - Réceptionniste tournant / tournante en établissement hôtelier
 - Responsable de réception hôtelière
 - Responsable des réservations en hôtellerie
 - Veilleur / Veilleuse de nuit en hôtellerie

Annexe B

Expérimentation sur données simulées pour l'étude des fonction d'attention

B.1 Formalisation du cadre d'étude

Nous voulons identifier un mot ou un signal social influent pour la décision de convocabilité, nous voulons retrouver cet élément clé au travers de la fonction d'attention en nous assurant que la valeur d'attention au pas de temps clé est bien maximale.

Pour la suite, nous posons le cadre d'étude suivant : nous devons effectuer une classification binaire de séquences $X_i = (x_t)_i$ où t désigne le t -ième élément de la i -ème séquence de X .

Dans nos expériences, un des éléments de la séquence (un certain t) contient l'information quant à l'étiquette de la séquence (X_i). Le but de nos différentes expériences est d'évaluer des fonctions d'attentions susceptibles de pouvoir retrouver cet élément clé afin de classifier la séquence.

Un vecteur de contexte C est disponible pour chaque séquence X_i et peut contenir de l'information utile quant à la classification de la séquence. Ce vecteur de contexte est similaire à celui qu'on retrouve lors de la modélisation du vecteur de l'intitulé des questions ou du titre de poste précédemment présenté en chapitre 5.

Nous présentons deux scénarios très similaires à ce que l'on pourrait obtenir dans notre problématique. Le premier scénario part du principe que le vecteur de contexte n'est pas du tout informatif. Dans ce cas, seules les valeurs intrinsèques de chaque pas de temps t influencent sur les valeurs d'attention. Cela pourrait être par exemple le cas pour un sourire en fin de réponse ayant une influence, peu importe le contexte de la question.

Le deuxième scénario part du principe que le vecteur de contexte est très informatif

quant à la désignation de l'élément clé de la réponse. Ceci correspond, par exemple, au fait que pour une question comme "*Quelles sont les aspects les plus importants lors d'un premier rendez-vous clientèle ?*", les termes liés au domaine de la relation client soient plus importants que les autres. Nous décrivons chacun des deux scénarios dans les sous-sections suivantes. Nous proposons par la suite une nouvelle fonction d'attention adaptée et nous comparons cette fonction avec deux autres fonctions d'attention populaires nommément la fonction d'attention additive et la fonction d'attention multiplicative.

B.2 Premier scénario : un contexte inutile

Nous décrivons la construction du jeu de données associée au premier scénario : le vecteur de contexte n'est pas du tout informatif, seules certaines valeurs intrinsèques de chaque pas de temps t influencent sur les valeurs d'attention.

Dans ce but nous effectuons les étapes suivantes pour la construction du jeu de données :

1. On génère N séquences (X) de taille T dont chaque élément est un vecteur de dimension D et est obtenu par tirage selon la loi uniforme entre -1 et 1.
2. On génère de la même manière N vecteurs de contexte (C) de dimension D . Dans ce scénario, le contexte est inutile.
3. On génère une clé K qui indique le moment important, c'est-à-dire l'élément qui contient l'étiquette de la séquence.
4. On attribue la valeur -1 ou 1 à la première dimension de chaque élément de chaque séquence.
5. On choisit un élément de chaque séquence au hasard et on attribue à la séquence, l'étiquette correspondant à la valeur de la première dimension.
6. On attribue à cet élément, les valeurs du vecteur clé K hormis pour la première dimension.

Ce jeu de données sera dénommé comme D_{S_1} .

Une image exemple du jeu de données du premier scénario est disponible en figure B.1

B.3 Deuxième scénario : un contexte indispensable

Nous décrivons la méthodologie associée au deuxième scénario : le vecteur de contexte est très informatif quant à la désignation de l'élément clé de la réponse.

Exemple Jeu de Données 1 : Contexte inutile.

$$K = [-0.3 \mid 0.2 \mid 0.4 \mid 0.5]$$

Exemple d'une séquence :

$$C = [0.5 \mid -0.8 \mid -0.2 \mid 0.1] \text{ } \left. \vphantom{C} \right\} \text{ aléatoire}$$

$$X = [-1 \mid -0.2 \mid 0.3 \mid 0.8] \quad [1 \mid 0.2 \mid 0.4 \mid 0.5] \quad [-1 \mid -1 \mid 0.9 \mid -0.1] \quad Y = [1]$$

Autre Exemple :

$$C = [0.3 \mid 0.8 \mid -0.1 \mid -1] \text{ } \left. \vphantom{C} \right\} \text{ aléatoire}$$

$$X = [-1 \mid 0.2 \mid 0.4 \mid 0.5] \quad [-1 \mid 0.8 \mid -0.1 \mid -0.2] \quad [1 \mid 0.2 \mid 0.2 \mid -0.9] \quad Y = [-1]$$

FIGURE B.1 – Exemple d'individus issus du jeu de données du scénario 1 : un contexte inutile.

Les vecteurs ayant les mêmes valeurs que la clé K contiennent l'information de l'étiquette Y de la séquence. Le vecteur de contexte C est ici complètement inutile.

Dans ce but, nous effectuons la méthodologie suivante pour la construction du jeu de données :

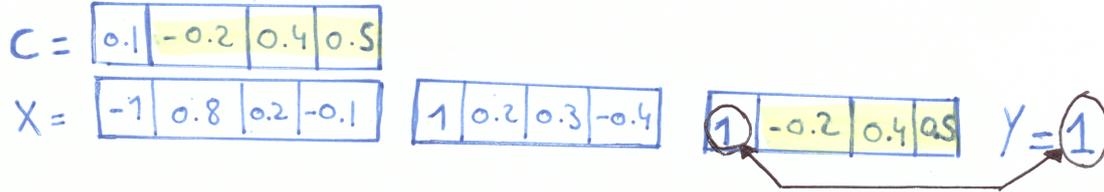
1. On génère N séquences (X) de taille T dont chaque élément est un vecteur de dimension D et est obtenu par tirage selon la loi uniforme entre -1 et 1.
2. On génère de la même manière N vecteurs de contexte (C) de dimension D . Dans ce scénario, le contexte est indispensable.
3. On attribue la valeur -1 ou 1 à la première dimension de chaque élément de chaque séquence.
4. On choisit un élément de chaque séquence au hasard et on attribue à la séquence, l'étiquette correspondant à la valeur de la première dimension.
5. On attribue à cet élément, les valeurs du vecteur clé C hormis pour la première dimension.

Ce jeu de données sera dénommé comme D_{S_2} .

Une image exemple du jeu de données du deuxième scénario est disponible en figure B.2

Exemple Jeu de Données 2: Un Contexte indispensable

Exemple d'une séquence :



Autre Exemple :

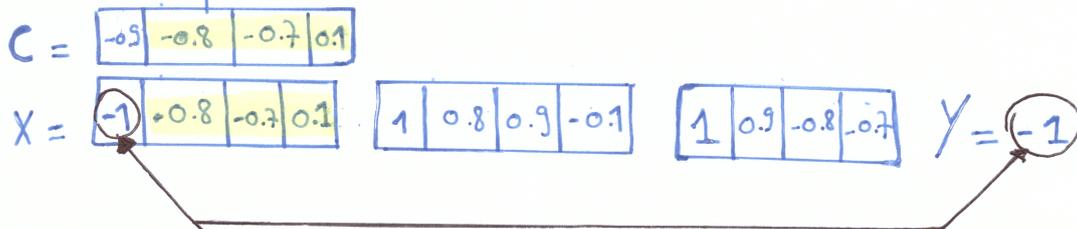


FIGURE B.2 – Exemple d'individus issus du jeu de données du scénario 2 : un contexte indispensable.

Les vecteurs ayant les mêmes valeurs que le vecteur de contexte C contiennent l'information de l'étiquette de la séquence. Le vecteur de contexte C est ici complètement nécessaire.

B.4 Architecture

Nous utilisons une architecture adaptée pour l'évaluation de la fonction d'attention. Plus précisément, pour une séquence X en entrée et un vecteur de contexte C , nous calculons tout d'abord des poids d'attention, puis une sigmoïde est appliquée à la première dimension du vecteur X issu de la somme pondérée par les poids d'attention.

$$\alpha_t = f(X; C) \quad (\text{B.1})$$

$$\hat{x} = \sum_t \alpha_t x_t \quad (\text{B.2})$$

$$\hat{y} = \sigma(\hat{x}[0]) \quad (\text{B.3})$$

Nous évaluons ainsi par la suite différentes fonctions d'attentions f sensibles de résoudre les problèmes jouets.

B.5 Fonctions d'attention évaluées

Nous évaluons trois formes d'attention lors des scénarios : deux formes d'attention classique (nommément attention additive et multiplicative) et nous proposons une troisième forme d'attention basée sur une fonction à portes.

L'attention additive est la même que celle définie dans le chapitre précédent. L'équation est définie comme telle :

$$u_t = \tanh(W_k x_t + W_c C + b_c) \quad (\text{B.4})$$

$$\alpha_t = \frac{\exp(u_p^\top u_t)}{\sum_{t'} \exp(u_p^\top u_{t'})} \quad (\text{B.5})$$

$$\hat{x} = \sum_t \alpha_t x_t \quad (\text{B.6})$$

Nous faisons l'hypothèse que cette forme d'attention est très efficace dans le premier scénario dans le sens où la valeur intrinsèque du vecteur est accessible par le terme $W_k x_t$. Cependant, comme évoqué en début de section le terme $W_c C$ est une constante parmi les pas de temps (il ne dépend pas de t) résultant en une modélisation sous-optimale de l'interaction avec le contexte.

La deuxième forme d'attention évaluée est l'attention multiplicative qui demeure une forme très répandue [Galassi et al., 2020] et constitue la forme principale d'attention de certaines architectures comme les transformers par exemple [Vaswani et al., 2017]. Pour rappel, la fonction d'attention s'écrit de cette façon :

$$u_t = (W_k x_t)^\top (W_c C) \quad (\text{B.7})$$

$$\alpha_t = \frac{\exp(u_t)}{\sum_{t'} \exp(u_{t'})} \quad (\text{B.8})$$

$$\hat{x} = \sum_t \alpha_t x_t \quad (\text{B.9})$$

Nous faisons l'hypothèse que cette forme d'attention est très efficace dans le deuxième scénario dans le sens où l'interaction avec le contexte est bien prise en compte à travers l'opération $(W_k x_t)^\top (W_c C)$. Cependant, l'accès à l'importance de la valeur intrinsèque du vecteur $W_k x_t$ peut être impossible de par l'interaction avec un vecteur de contexte non utile à la modélisation.

Finalement, nous proposons une troisième forme d'attention inspirée de [Yang et al., 2019a]. Cette forme d'attention intègre deux termes représentant les formes d'attention précédemment présentées : notamment un terme de self attention pour évaluer les événements intrinsèquement importants et un terme multiplicatif entre le contexte et les pas de temps pour évaluer les événements importants au regard du contexte. Cette forme d'attention est présentée ci-dessous :

$$u_t = \tanh(\lambda_t(W_c C(W_h x_t^\top) + (1 - \lambda_t)(b^\top(W_k x_t))) \quad (\text{B.10})$$

$$\lambda_t = \sigma([W_d C * W_g x_t; b * W_l x_t]) \quad (\text{B.11})$$

$$\alpha_t = \frac{\exp(u_t)}{\sum_{t'} \exp(u_{t'})} \quad (\text{B.12})$$

$$\hat{x} = \sum_t \alpha_t h_t^{\text{fusion}} \quad (\text{B.13})$$

où W_c , W_d , W_g , W_l , W_h et W_k sont des matrices de poids, b est un vecteur de poids et b^\top désigne la transposition de b , σ désigne la fonction sigmoïde, ; désigne l'opération de concaténation et $*$ désigne le terme opération de produit.

Nous concevons cette fonction d'attention pour prendre en compte l'interaction multiplicative de x avec le contexte C à travers le premier terme $W_c C(W_h x_t^\top)$ et l'importance intrinsèque de x_t à travers le second terme $b^\top(W_k x_t)$. Afin de pondérer chacun des deux termes, un mécanisme de porte (λ_t) est calculée à partir de la concaténation des deux précédents termes.

B.6 Résultats de l'exemple jouet

Nous prenons comme paramètres :

- le nombre de pas de temps $T = 70$
- la dimension de chaque élément $D = 64$
- le nombre de séquences du jeu de données $N = 5000$

Une répartition entre jeu d'entraînement, de validation et de test est effectuée aléatoirement pour chacun des jeux de données D_{S_1} et D_{S_2} (à hauteur respective de 80%,10%,10% du jeu de données initial). Nous constituons aussi un troisième jeu de données résultant de la concaténation de D_{S_1} et D_{S_2} afin d'évaluer que la fonction d'attention contextuelle peut s'adapter dynamiquement à l'un ou l'autre des scénarios.

B.6. RÉSULTATS DE L'EXEMPLE JOUET

	D_{S_1}	D_{S_2}	D_{S_1} et D_{S_2}
Attention additive	1.00	0.49	0.765
Attention multiplicative	0.654	0.997	0.798
Attention contextuelle	1.00	0.987	0.994

TABLE B.1 – Tableau des résultats de justesse en fonction différentes fonctions d'attention sur les jeux de données de l'exemple jouet. D_{S_1} et D_{S_2} consistent en la concaténation des deux jeux de données.

Les résultats pour l'exemple jouet sont disponible en tableau B.1. Les résultats sur l'exemple jouet confirment nos hypothèses. Ainsi l'attention additive prend bien en compte l'aspect intrinsèque des pas de temps, mais non l'interaction avec le contexte, tandis que l'attention multiplicative modélise l'interaction avec le contexte, mais échoue à prendre en compte l'aspect intrinsèque des pas de temps. L'attention contextuelle montre une capacité à modéliser les deux phénomènes.

Annexe C

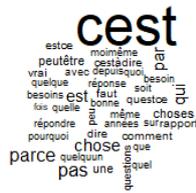
Nuages des mots de la représentation multimodale responsables de la discrimination des tranches d'attention



(a) Nuage de mots de la dimension v82



(b) Nuage de mots de la dimension v191



(c) Nuage de mots de la dimension v248



(d) Nuage de mots de la dimension v164



(e) Nuage de mots de la dimension v94



(f) Nuage de mots de la dimension v202



(g) Nuage de mots de la dimension v39



(h) Nuage de mots de la dimension v165

FIGURE C.1 – Nuage de mots des dimensions associées avec les tranches d'attention.



nous
merci
dans bon
isqui

(a) Nuage de mots de la dimension v146

FIGURE C.2 – Nuage de mots des dimensions associées avec les tranches d’attention.

Title : Automatic Analysis of Multimodal Behaviors during Asynchronous Video Interviews for Recruitment

Keywords : Job Interview, Social Signal Processing, Affective Computing, Neural Networks

Abstract : The development of new technologies influences all sectors of activity, including human resources and particularly in the recruitment process. The emergence of pre-recorded video interviews makes it possible to organize asynchronous interviews with candidates and evaluate them. Candidates connect to a platform and film themselves while they answer questions defined in advance by recruiters. The platform then allows several recruiters to evaluate the candidate, share notes and possibly invite the candidate to a face-to-face interview. As part of a project with an industrial partner, we collected two corpuses of more than 5,000 asynchronous video interviews for real jobs. This thesis studies the task of predicting the performance of candidates during asynchronous video interviews using three modalities (verbal content, prosody and facial expressions) using data from real interviews. For this purpose, we propose a new multimodal hierarchical attention model called HireNet. In HireNet, an interview

is viewed as a sequence of questions and answers containing salient social cues. A special feature of HireNet is the use of attention mechanisms, which aim to identify the most relevant parts of an answer. While most deep learning systems use attention mechanisms to provide a quick visualization of slices when an increase in attention values occurs, we perform an in-depth analysis to understand what happens during these moments. Overall, this method aims to improve the interpretability of such systems and to question their use as an exploratory tool. Our third contribution concerns biases in automatic video interview analysis systems. We propose a first approach that uses adversarial training to learn a representation that ignores the gender and ethnicity of the candidates. We then study the use of this adversarial training without the need to collect sensitive information about the candidates. In this way, we aim to improve the fairness of future automatic systems for processing job interview videos for equal job selection.

Titre : Analyse Automatique des Comportements Multimodaux lors d'Entretiens Vidéo Différés pour le Recrutement

Mots clés : Entretien d'Embauche, Traitement des Signaux Sociaux, Informatique Affective, Réseaux de Neurones.

Résumé : Le développement des nouvelles technologies influence tous les secteurs d'activités, y compris celui des ressources humaines et notamment dans le processus de recrutement. L'émergence des entretiens vidéo différés permet d'organiser en asynchrone des entretiens avec des candidats et de les évaluer. Les candidats se connectent à une plateforme et se filment pendant qu'ils répondent à des questions définies en amont par les recruteurs. La plateforme permet ensuite à plusieurs recruteurs d'évaluer le candidat, de partager des notes et d'inviter éventuellement le candidat à un entretien en face-à-face. Dans le cadre d'un projet avec un partenaire industriel, nous avons recueilli deux corpus de plus de 5000 entretiens d'embauche vidéo asynchrones pour des postes réels. Cette thèse étudie la tâche consistant à prédire les performances des candidats lors d'entretiens vidéo asynchrones en utilisant trois modalités (contenu verbal, prosodie et expressions faciales) en utilisant des données provenant d'entretiens réels. Dans ce but, nous proposons un nouveau modèle multimodal d'attention hiérarchique appelé HireNet. Dans HireNet, un entretien est considéré comme une séquence de

questions et de réponses contenant des signaux sociaux saillants. Une particularité de HireNet est l'utilisation de mécanismes d'attention, qui visent à identifier les parties les plus pertinentes d'une réponse. Alors que la plupart des systèmes d'apprentissage profond utilisent des mécanismes d'attention pour offrir une visualisation rapide des tranches lorsqu'une augmentation des valeurs d'attention se produit, nous effectuons une analyse approfondie pour comprendre ce qui se passe lors de ces moments. Dans l'ensemble, cette méthode vise à améliorer l'interprétabilité de tels systèmes et à s'interroger sur leur utilisation comme outil exploratoire. Notre troisième contribution concerne les biais dans les systèmes d'analyse automatique des entretiens vidéo. Nous proposons une première approche qui utilise un entraînement adversaire pour apprendre une représentation ignorant le sexe et l'ethnicité des candidats. Nous étudions ensuite l'utilisation de cet entraînement adversaire sans qu'il soit nécessaire de recueillir des informations sensibles sur les candidats. Ainsi, nous visons à améliorer l'équité des prochains systèmes automatiques de traitement des vidéos d'entretiens d'embauche pour une égalité dans la sélection des emplois.