Learning multimodal interaction models in mixed societies

Muhammad Usman Malik

Normandie Université

THESE

Pour obtenir le diplôme de doctorat

Spécialité Informatique

Préparée au sein de l'INSA ROUEN NORMANDIE

Learning Multimodal Interaction Models in Mixed Societies

Présentée et soutenue par

Muhammad Usman MALIK

Thèse soutenue publiquement le 24-11-2020 devant le jury composé de		
Mme. Chloé CLAVEL	Professeur des Universités / Université, Telecom Paris, France	Rapporteur
Mr. Mohamed CHETOUANI	Professeur des Universités / Université, Sorbonne Université, France	Rapporteur
Mme. Elisabeth ANDRE	Professeur des Universités / Université University of Augsburg, Germany	Examinatrice
Mr. Pierre CHEVAILLIER	Professeur des Universités / Université ENIB, (Ecole Nationale d'Ingénieurs de Brest) , France	Examinateur
Mr. Kotaro FUNAKOSHI	Maitre de conférences Tokyo Institute of Technology (TITECH), Japan	Examinateur
Mr. Alexandre PAUCHET	Maitre de conférences HDR / INSA Rouen Normandie, Rouen, France	Directeur de thèse
Mr. Julien SAUNIER	Maitre de conférences / INSA Rouen Normandie, Rouen, France	Encadrant de thèse

Thèse dirigée par Alexandre PAUCHET, laboratoire LITIS

INSA ROUEN NORMANDIE

MIIS

litis

Abstract

Human -Agent Interaction and Machine learning are two different research domains. Human-agent interaction refers to techniques and concepts involved in developing smart agents, such as robots or virtual agents, capable of seamless interaction with humans, to achieve a common goal. Machine learning, on the other hand, exploits statistical algorithms to learn data patterns. The proposed research work lies at the crossroad of these two research areas.

Human interactions involve multiple modalities, which can be verbal such as speech and text, as well as non-verbal i.e. facial expressions, gaze, head and hand gestures, etc. To mimic real-time human-human interaction within human-agent interaction, multiple interaction modalities can be exploited. With the availability of multimodal human-human and human-agent interaction corpora, machine learning techniques can be used to develop various interrelated human-agent interaction models. In this regard, our research work proposes original models for addressee detection, turn change and next speaker prediction, and finally visual focus of attention behaviour generation, in multiparty interaction.

Our addressee detection model predicts the addressee of an utterance during interaction involving more than two participants. The addressee detection problem has been tackled as a supervised multiclass machine learning problem. Various machine learning algorithms have been trained to develop addressee detection models. The results achieved show that the proposed addressee detection algorithms outperform a baseline.

The second model we propose concerns the turn change and next speaker prediction in multiparty interaction. Turn change prediction is modeled as a binary classification problem whereas the next speaker prediction model is considered as a multiclass classification problem. Machine learning algorithms are trained to solve these two interrelated problems. The results depict that the proposed models outperform baselines.

Finally, the third proposed model concerns the visual focus of attention (VFOA) behaviour generation problem for both speakers and listeners in multiparty interaction. This model is divided into various sub-models that are trained via machine learning as well as heuristic techniques. The results testify that our proposed systems yield better performance than the baseline models developed via random and rule-based approaches. The proposed VFOA behavior generation model is currently implemented as a series of four modules to create different interaction scenarios between multiple virtual agents. For the purpose of evaluation, recorded videos for VFOA generation models for speakers and listeners, are presented to users who evaluate the baseline, real VFOA behaviour

and proposed VFOA models on the various naturalness criteria. The results show that the VFOA behaviour generated via the proposed VFOA model is perceived more natural than the baselines and as equally natural as real VFOA behaviour.

RÉSUMÉ

Les travaux de recherche proposés se situe au carrefour de deux domaines de recherche, l'interaction humain-agent et l'apprentissage automatique. L'interaction humain-agent fait référence aux techniques et concepts impliqués dans le développement des agents intelligents, tels que les robots et les agents virtuels, capables d'interagir avec les humains pour atteindre un objectif commun. L'apprentissage automatique, d'autre part, exploite des algorithmes statistiques pour apprendre des modèles de données.

Les interactions humaines impliquent plusieurs modalités, qui peuvent être verbales comme la parole et le texte, ainsi que les comportements non-verbaux, c'est-à-dire les expressions faciales, le regard, les gestes de la tête et des mains, etc. Afin d'imiter l'interaction humain-humain en temps réel en interaction humain-agent, plusieurs modalités d'interaction peuvent être exploitées. Avec la disponibilité de corpus d'interaction multimodales humain-humain et humain-agent, les techniques d'apprentissage automatique peuvent alors être utilisées pour développer des modèles interdépendants participant à l'interaction humain-agent. À cet égard, nos travaux de recherche proposent des modèles originaux pour la détection de destinataires d'énoncés, le changement de tour de parole et la prédiction du prochain locuteur, et enfin la génération de comportement d'attention visuelle en interaction multipartie.

Notre modèle de détection de destinataire prédit le destinataire d'un énoncé lors d'interactions impliquant plus de deux participants. Le problème de détection de destinataires a été traité comme un problème d'apprentissage automatique multiclasse supervisé. Plusieurs algorithmes d'apprentissage ont été entrainés pour développer des modèles de détection de destinataires. Les résultats obtenus montrent que ces propositions sont plus performants qu'un algorithme de référence.

Le second modèle que nous proposons concerne le changement de tour de parole et la prédiction du prochain locuteur dans une interaction multipartie. La prédiction du changement de tour est modélisée comme un problème de classification binaire alors que le modèle de prédiction du prochain locuteur est considéré comme un problème de classification multiclasse. Des algorithmes d'apprentissage automatique sont entraînés pour résoudre ces deux problèmes interdépendants. Les résultats montrent que les modèles proposés sont plus performants que les modèles de référence.

Enfin, le troisième modèle proposé concerne le problème de génération du comportement d'attention visuelle (CAV) pour les locuteurs et les auditeurs dans une interaction multipartie. Ce modèle est divisé en plusieurs sous-modèles qui sont entrainés par l'apprentissage machine ainsi que par des techniques heuristiques. Les résultats attestent que les systèmes que nous proposons sont plus performants que les modèles de

référence développés par des approches aléatoires et à base de règles.

Le modèle de génération de comportement CAV proposé est mis en œuvre sous la forme d'une série de quatre modules permettant de créer différents scénarios d'interaction entre plusieurs agents virtuels. Afin de l'évaluer, des vidéos enregistrées pour les modèles de génération de CAV pour les orateurs et les auditeurs, sont présentées à des évaluateurs humains qui évaluent les comportements de référence, le comportement réel issu du corpus et les modèles proposés de CAV sur plusieurs critères de naturalité du comportement. Les résultats montrent que le comportement de CAV généré via le modèle est perçu comme plus naturel que les bases de référence et aussi naturel que le comportement réel.

To my family,
My mother, and my late father, who must be proud when he looks down at me from heavens,
for your relentless support, love and affection.

Acknowledgements

When I started my Ph.D. I expected it to be a dull and labourious Journey - however, as it turned out, I can absolutely say it was a life-changing experience for me and that is because of the people who made this journey worthwhile.

In this regard, I owe a heavy debt of gratitude to my supervisor Mr. Alexandre Pauchet for welcoming me to his team. His scientific guidance, motivation and relentless support at each stage of my P.hD. helped me overcome not only my research obstacles but my personal slumps as well. Without his continuous assistance, I would never have been able to complete this research work.

I would also like to thank my co-supervisor Mr. Julien Saunier for his constant help and sound scientific advice throughout my research period. His encouragement and confidence in my abilities has been a driving force during my P.hD.

I am also indebted to my thesis reviewers Mme Chloe Clavel and Mr. Mohamed Chetouani for their insightful reviews that helped improve my thesis manuscript. I am also grateful to Mme Elisabeth Andre, Mr Pierre Chevaillier and Mr Kotaro Funakoshi for agreeing to be the examiners of my P.hD. defense.

I would like to thank Maël for helping me with system implementation and Arfa for her help with distribution and conduct of surveys. Without their help, I would not have been able to complete this research work within stipulated time period.

My lab fellows and colleagues also deserve a big thanks for all the moments that we shared in and outside of the lab which made my P.hD. journey even more pleasant. In this regard, I would like to thank Mukesh, Franco, Qiu Chi, Benjamin, Imad, Imen, Marwa, Henrique, and many others. I expect to be excused if I forget some names.

I would like to express my gratitude to the secretaries of the laboratory Brigitte and Sandra who were always ready to help me with administrative tasks.

Finally, I am indebted to my mother Amina, my sisters Saima and Asma, and my brothers Amir and Khurram for always being there for me whenever I needed them. Their continuous encouragement, and financial and moral support has always been an impetus that kept me going during my P.hD.

In the end, I would like to thank my late father Mr. Obaid Ullah Malik, who did everything in his capacity to make sure that I achieve all the milestones that I want in my life. Thank you *Abbu Jee*.

23 Octobre 2020
Usman

TABLE OF CONTENTS

List of Tables

List of Figures

Figure **Page**

Introduction

L'interaction multimodale humain-agent se situe au carrefour de différents domaines de recherche, notamment l'intelligence artificielle, la vision par ordinateur, la psychologie, etc. En raison de l'essor de l'informatique, les ordinateurs se sont intégrés dans notre vie quotidienne. Dans de nombreuses applications, les humains doivent interagir avec des agents d'une manière similaire à celle dont ils interagissent entre eux. Par exemple, dans les interactions entre humains, en moyenne 65% de la signification du message est déduite des comportements non verbaux du locuteur [Foley and Gentile, 2010]. Par conséquent, pour assurer une interaction naturelle entre les humains et les agents, les agents doivent être capables à la fois d'interpréter des stimuli multimodaux et d'exprimer des comportements également multimodaux.

La Figure I contient un exemple de cycle typique d'interaction multimodale humain-agent. Elle montre que l'humain exprime son intention, son attention et ses émotions à l'aide d'actions telles que la parole, le geste, etc. L'agent perçoit les actions humaines par le biais de dispositifs d'entrée tels que les microphones, la caméra vidéo, etc. Pour générer une réponse, les données d'entrée reçues des humains et de l'environnement sont traitées pour prendre une décision finale concernant l'action à entreprendre en réponse à l'entrée. Pour exprimer une sortie ou une action, les agents utilisent des dispositifs tels qu'un convertisseur de texte en parole, etc. Ces actions sont perçues par l'humain par le biais de ses sens tels que la vue, l'ouïe, etc. Ce processus se poursuit tout au long de l'interaction entre les humains et les agents.

De plus, les interactions humain-agent peuvent être dyadiques, impliquant deux participants, ou multiparties, dans lesquelles plus de deux participants interagissent (voir Figure II). Les tâches d'interaction sont plus simples en interaction dyadique qu'en multipartie. Par exemple, en interaction dyadique, le destinataire d'un énoncé est connu puisqu'il n'y a qu'un seul auditeur. De même, le centre d'attention des participants est relativement simple car il n'y a qu'un seul participant à regarder, en plus des objets.

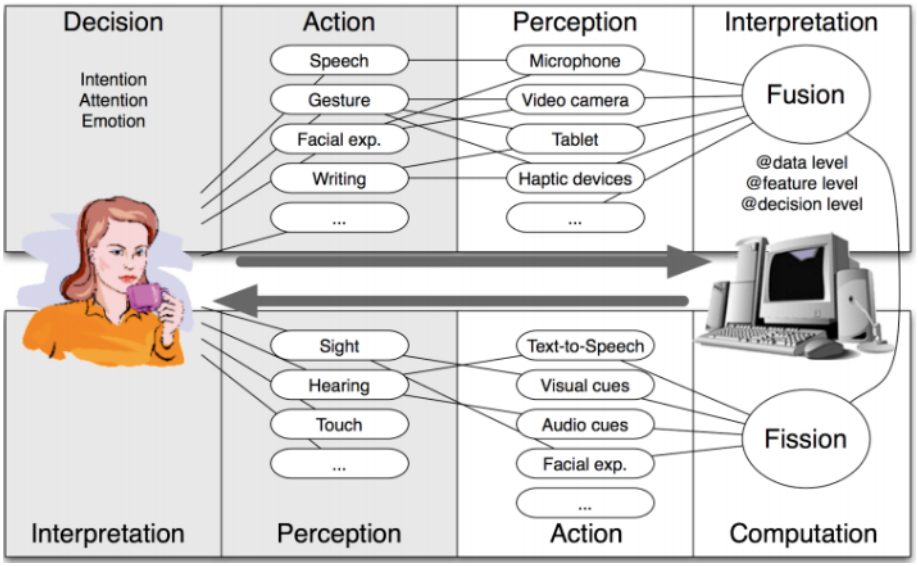

Figure I – Interaction multimodale entre humain et agent.

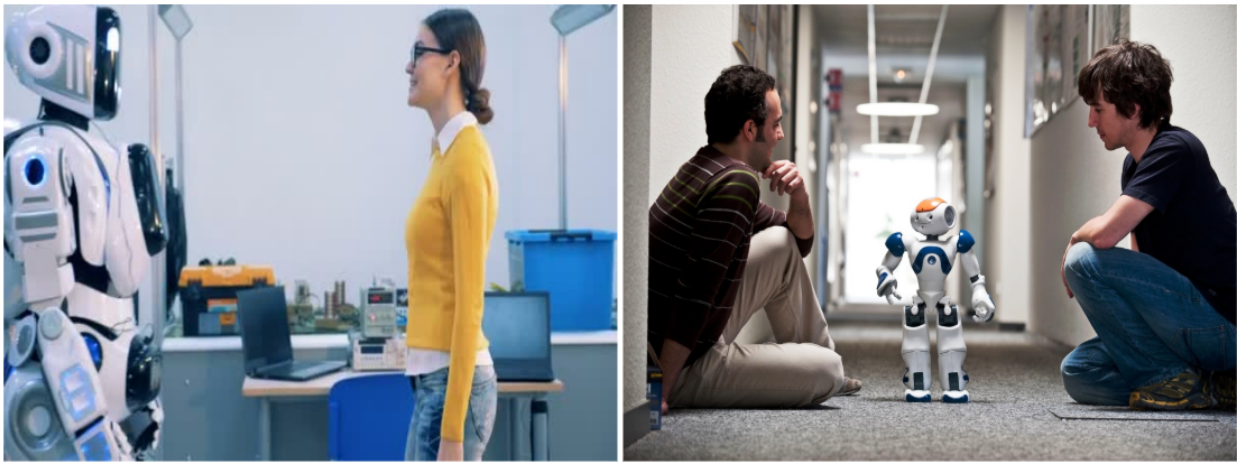

Figure II – Interaction dyadique (gauche) et interaction multipartie (droite)

Au contraire, la prise de décision lors des interactions multiparties est relativement complexe. Par exemple, prédire l'orateur du prochain énoncé après que l'orateur actuel a terminé est une tâche complexe en raison de la présence de plusieurs participants.

Bien que des techniques fondées sur des règles puissent être utilisées pour l'interaction humain-agent, la disponibilité de nombreuses données et les améliorations de la puissance de calcul amènent à considérer l'utilisation de techniques d'apprentissage machine pour l'interaction humain-agent.

Les techniques d'apprentissage machine permettent d'apprendre des modèles de comportement d'interaction d'agents intelligents à partir de corpus d'interactions entre humains. Les caractéristiques d'interaction apprises à partir des interactions humain-

humain peuvent ensuite être intégrées dans les agents intelligents. À cet égard, l'apprentissage machine dans l'interaction humain-agent vise à répondre à trois questions [Kaiser et al., 1997] :
Q1: Qu'est-ce que les agents intelligents apprennent des humains et comment ?
Q2: Qu'est-ce que les agents intelligents apprennent les uns des autres et comment ?
Q3: Qu'est-ce que les agents intelligents peuvent apprendre sans aide extérieure, et comment ?

Cette thèse se concentre sur la première question où nous proposons des modèles issus de l'apprentissage machine, parfois complétés par des modèles heuristiques, qui apprennent des interactions humain-humain et humain-agent pour améliorer les performances de différentes tâches liées à l'interaction humain-agent.

La motivation fondamentale de ce travail de recherche est qu'il est possible de proposer des solutions performantes pour différentes tâches d'interaction humain-agent grâce à la sélection de caractéristiques multimodales pertinentes et d'algorithmes d'apprentissage machine efficaces.

Dans cette recherche, nous proposons d'aborder un problème de perception, un problème de prise de décision et un problème de génération de comportement. Nous proposons des modèles pour trois tâches connexes dans l'interaction multipartie humains-agents : la détection du ou des destinataires d'un énoncé, le changement de tour de parole et la prédiction du prochain locuteur, et la génération du comportement d'attention visuelle. Les modèles que nous proposons sont basés sur des techniques d'apprentissage machine qui tirent parti de la nature multimodale des interactions humaines. Dans certains cas, des approches heuristiques sont utilisées pour traiter les résultats des modèles d'apprentissage machine.

Processus global du système

Les différentes tâches effectuées par les modèles proposés et leurs interactions sont représentées dans la Figure III. Au centre du schéma se trouvent trois modèles : la détection de destinataire, le changement de tour de parole et la prédiction du prochain locuteur et les modèles de génération du Comportement d'Attention Visuelle (CAV). Les trois modèles reposent sur l'apprentissage machine supervisé et les approches heuristiques.

L'objectif du modèle de détection de destinataire est de détecter le ou les destinataires de l'énoncé courrant. Cette détection des destinataires est conçue comme un problème d'apprentissage machine multiclasse supervisé, car dans une interaction multipartie, un

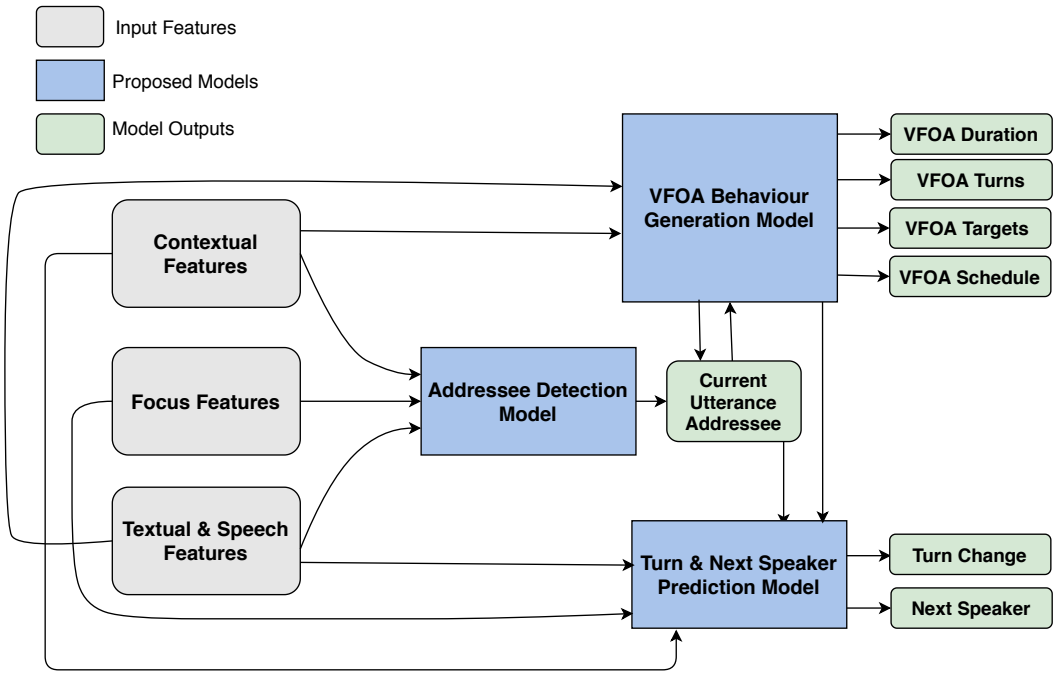

Figure III – Organigramme du processus des modèles proposés

énoncé peut être adressé à n'importe quel participant à la réunion, à tous les participants à la réunion ou à un sous-ensemble de ces participants.

Le modèle de prédiction du changement de tour et du prochain locuteur est un second modèle autonome qui effectue deux tâches : la prédiction du changement de locuteur, qui se produit lorsque le locuteur d'un énoncé est différent de celui de l'énoncé précédent, et la prédiction du participant qui prend le nouveau tour. La prédiction du changement de tour est une tâche de classification binaire avec deux sorties possibles : si un changement de tour se produit ou non après l'énoncé actuel. Au contraire, la prédiction du prochain locuteur est un problème de classification multiclasse puisque dans une interaction multipartie, n'importe lequel des participants peut prendre le tour de parole. Un modèle combiné de changement de tour et de prédiction du locuteur suivant a également été proposé, dans lequel le changement de tour prédit est inclus dans l'ensemble de caractéristiques utilisé pour entrainer le modèle de prédiction du locuteur suivant.

Le modèle de génération de comportement d'attention visuelle est un modèle hybride qui prédit le CAV final pour les agents intelligents, qu'ils soient orateur ou auditeur dans une interaction multipartie. Le modèle de génération de comportement CAV se compose de quatre sous-modèles : un prédicteur de nombre de tours d'attention visuelle, un prédicteur de durée des tours, un prédicteur de cible de l'attention, et enfin un

planificateur de CAV.

Les modèles proposés s'appuient sur un ensemble de caractéristiques qui peuvent être décomposées en trois types, à savoir les caractéristiques contextuelles, les caractéristiques textuelles et vocales et les caractéristiques d'attention. Les caractéristiques sont sélectionnées (i) sur la base de leur importance pour l'exécution de la tâche correspondante, comme indiqué dans l'analyse de l'état de l'art, et (ii) l'analyse statistique des corpus de données comme mentionné dans les sections de sélection des caractéristiques pour chaque tâche.

Dans ce travail de recherche, un pipeline standard d'apprentissage automatique a été adopté pour apprendre, tester et évaluer les performances des modèles proposés. Les modèles sont principalement basés sur des algorithmes traditionnels d'apprentissage machine supervisé.

La figure IV explique brièvement la méthodologie adoptée pour entrainer les modèles d'apprentissage machine dans ce travail.

La première étape d'entrainement d'un algorithme d'apprentissage machine supervisé consiste à collecter l'ensemble des données. La technique de validation croisée à K blocs est utilisée pour diviser les données en ensembles d'entrainement et de test et évaluer les modèles d'apprentissage machine. Dans la validation croisée à K blocs, le corpus de données est divisé en K parties. $K-1$ parties sont utilisées pour entrainer l'algorithme et la dernière partie est exploitée pour évaluer la performance de l'algorithme d'apprentissage automatique. Chacune des K parties de données est utilisée au moins une fois pour l'entrainement et une fois pour l'évaluation. Dans ces travaux, une validation croisée à 5 blocs a été utilisée pour les corpus de données MPR et AMI.

Pour le corpus de données MULTISIMO, la validation croisée à K blocs n'a pas été utilisée en raison du nombre limité d'enregistrements. Les modèles sont entraînés à l'aide d'un ensemble d'apprentissage, puis des prédictions sont effectuées sur l'ensemble de test. Les prédictions sont ensuite traitées par des approches heuristiques si nécessaire, afin d'obtenir les résultats finaux qui sont utilisés pour l'évaluation du modèle.

Détection des destinataires

La détection des destinataires est une tâche fondamentale pour la gestion fluide d'un dialogue et la prise de tour de parole dans l'interaction humain-agent. Bien que la détection du destinataire soit implicite lors d'interactions dyadiques, elle devient une tâche difficile lorsque plus de deux participants sont impliqués. Pour s'attaquer au problème de la détection du destinataire dans une interaction multipartie, des approches

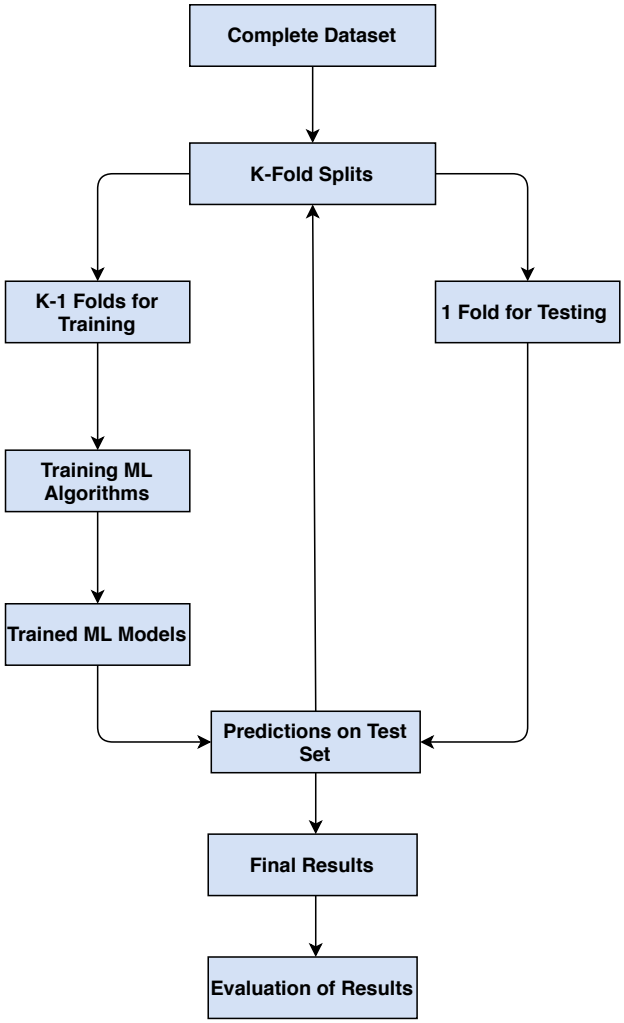

Figure IV – Processus d'apprentissage machine supervisé

heuristiques et statistiques ont été développées dans la littérature. Cependant, la plupart de ces travaux dépendent de paramètres spécifiques. La quantité limitée de données pour l'entrainement rend également difficile l'élaboration d'un modèle générique de détection des destinataires.

Dans ces travaux, nous proposons un modèle fondé sur un encodage de caractéristiques génériques qui sont utilisées pour entrainer des algorithmes d'apprentissage machine, capables de détecter des destinataires. Nous émettons l'hypothèse qu'un modèle statistique avec des caractéristiques génériques peut être performant pour la détection de destinataires dans de multiples scénarios. Les résultats obtenus sur différents corpus de données avec un nombre variable de participants confirment cette affirmation.

Travaux connexes

Le travail précurseur de [Traum et al., 2004] propose une approche basée sur des règles exploitant l'énoncé précédent, l'énoncé actuel, l'orateur précédent et l'orateur actuel pour détecter le destinataire. La précision varie entre 65 % et 100 % sur l'ensemble des données du *Mission Rehearsal Exercise* [Traum et al., 2006] selon l'acte de dialogue. Cependant, l'algorithme ne se généralise pas bien sur les autres jeux de données : appliqué au jeu de données AMI [McCowan et al., 2005], la précision tombe à 36%. [Akker and Traum, 2009] améliore ce travail en incorporant le regard comme l'un des précurseurs des règles, ce qui donne une précision de 65%. Les auteurs ont également testé le regard comme seule caractéristique pour la détection des destinataires, faisant état d'une précision de 57%. Dans ce cas, la seule règle est que si un orateur regarde un participant pendant plus de 80 % de la durée d'un énoncé, le destinataire est ce participant, sinon l'énoncé est adressé au groupe.

En ce qui concerne les approches statistiques, [Jovanovic, 2007] utilise des réseaux Bayésiens pour exploiter l'énoncé actuel et précédent, l'orateur, le regard, le sujet de discussion et d'autres méta caractéristiques sur le corpus multimodal M4 [Jovanovic et al., 2006] avec une précision de 81%. Cet algorithme est également testé par [Akker and Traum, 2009] sur le corpus AMI, avec une précision de 62% qui montre que l'algorithme ne se généralise pas bien.

[Akker and Akker, 2009] proposent un modèle statistique fondé sur des arbres de régression logistique afin de répondre à la question binaire *are you being addressed?*, avec une précision de 92% sur le corpus AMI. Les deux limites de ce travail sont, d'une part, que les auteurs ne prennent en compte que le seul point de vue du locuteur au lieu d'identifier le destinataire et, d'autre part, que le modèle dépend d'un positionnement fixe des participants.

[Baba et al., 2011] exploitent des conversations triadiques humain-humain-agent pour développer un modèle basé sur les *Support Vector Machines* qui distingue si un propos est adressé à l'humain ou à l'agent. Ils rapportent une précision de 80,28% pour cette tâche de classification binaire, en utilisant le texte, l'orientation de la tête et les caractéristiques acoustiques.

Enfin, seuls quelques travaux ont utilisé des techniques d'apprentissage profond pour s'attaquer au problème de la détection du destinataire. [Le Minh et al., 2018] propose une solution fondée sur un réseau neuronal convolutif [Krizhevsky et al., 2012] pour la détection des destinataires dans le corpus de données *GazeFollow* [Recasens et al., 2015]. L'une des principales limites de ce travail est que la détection du destinataire est effectuée par l'angle d'un tiers, avec une précision de 62,5 %.

Exigence	Description
r1	Indépendant du nombre de participants
r2	Indépendant du positionnement du participant
r3	Prédiction du destinataire (au lieu de la prédiction de si un propos est adressé à l'agent ou non)
r4	Doit être capable de faire des prévisions en conditions réelles

Table I – Exigences du modèle

Dans ce travail, nous considérons qu'un modèle avec des caractéristiques génériques peut être utilisé pour la détection des destinataires en temps réel, indépendamment du corpus de données et du nombre et du positionnement des participants. Les exigences liées à ce modèle sont mentionnées dans le Tableau I. Le raisonnement qui sous-tend ces exigences est que les participants qui ne sont pas actuellement destinataires doivent également savoir qui est adressé en temps réel (r3, r4), indépendamment de leur emplacement dans la salle (r2) et du nombre de participants (r1).

Ce travail de recherche apporte 4 améliorations par rapport au modèle de base [Akker and Traum, 2009] : (i) la sélection de nouvelles caractéristiques pour la détection des destinataires, telles qu'elles ressortent des travaux de recherche, (ii) l'exploitation de plusieurs algorithmes d'apprentissage machine qui n'ont jamais été utilisés auparavant pour la détection des destinataires, (iii) la proposition de deux approches de codage de l'attention visuelle qui améliorent la précision du modèle de référence choisi, et (iv) les changements nécessaires pour que les modèles de détection des destinataires fonctionnent en temps réel sont mis en œuvre.

Caractéristiques pour la détection des destinataires

Les caractéristiques ont été sélectionnées en raison de leur importance extraite de la littérature. Les caractéristiques sélectionnées sont destinées à être génériques et ne dépendent donc pas d'un scénario sous-jacent : le nombre de participants et leur positionnement ne sont pas pris en compte. L'analyse de certaines des caractéristiques est effectuée sur le corpus AMI [Carletta, 2007] et MULTISIMO [Koutsombogera and Vogel, 2018] puisque ce sont les deux seuls corpus de données qui contiennent toutes les caractéristiques exploitées pour la détection des destinataires dans ce travail. Les caractéristiques sélectionnées pour la détection des destinataires sont : la cible d'attention visuelle du locuteur et des auditeurs, les actes de dialogue des énoncés actuels et précédents, les locuteurs des énoncés actuels et précédents, le fait qu'un énoncé contienne ou non le mot "you", la durée et le nombre de mots de l'énoncé.

Les travaux existants montrent que l'attention visuelle est un indice important pour la détection des destinataires. Étant donné que le destinataire est prédit au niveau des actes de dialogues et que le focus visuel peut se déplacer plusieurs fois au cours d'un acte, il peut y avoir différentes façons d'encoder les focus du locuteur et des auditeurs. Cependant, à notre connaissance, aucun des travaux existants n'étudie l'impact des différents schémas de codage du focus sur la détection du destinataire.

Schémas d'encodage du focus visuel

L'attention visuelle (ou focus) des auditeurs et des orateurs est un marqueur pour la détection des destinataires. Pour mieux comprendre l'importance du focus, les corpus AMI et MULTISIMO ont été analysés. Dans AMI, une personne peut regarder 3 autres participants, un écran de projection, la table et un tableau blanc. Lorsque ni les participants ni les objets ne sont au centre de l'attention, le focus est marqué comme *autre*. Dans MULTISIMO, un participant ne peut regarder que 2 autres participants, sinon le focus est marqué comme *autre*.

Une brève description des schémas de codage du focus est donnée dans les deux sous-sections suivantes.

Encodage one-hot du focus

Au cours d'un acte de dialogue, l'attention peut se porter sur n'importe quel individu ou objet. Les étapes suivantes expliquent le processus d'annotation du focus pour chaque énoncé :

1. s'il y a plus de deux participants et/ou objets en focus, le focus est marqué comme *multiple* ;

2. si l'accent est mis sur un participant et un objet, alors

    a) si la personne est d'abord visée et ensuite l'objet, le focus est marqué comme étant la personne,

    b) sinon le focus est marqué comme multiple.

D'après la Figure V, pour un encodage one-hot, un seul des participants ou objets a une valeur de 1 alors que les autres entités ont une valeur de 0.

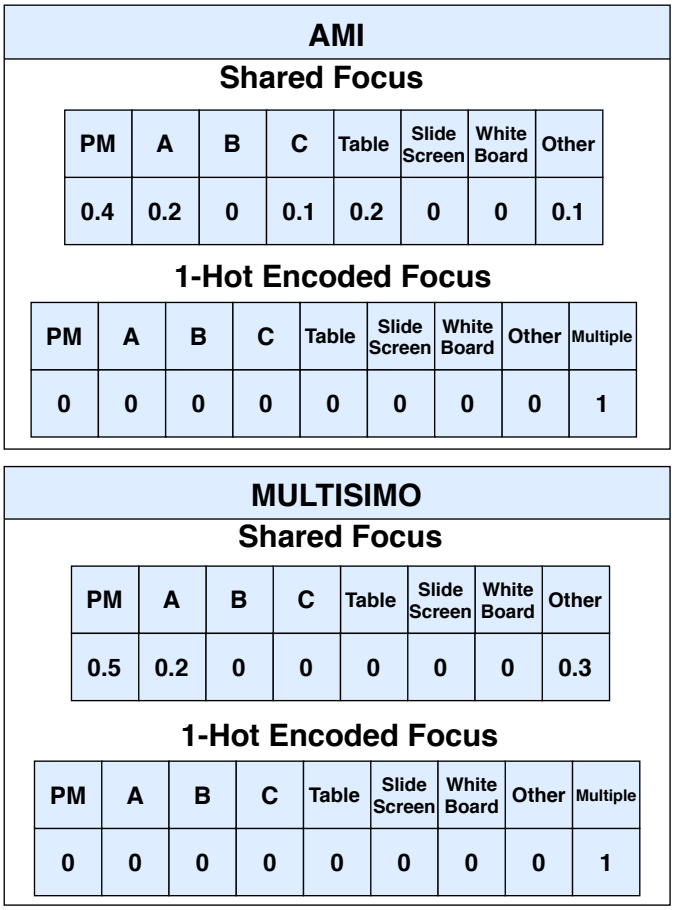

Figure V – Exemples de vecteurs de codage à focalisation unique et partagée pour les corpus de données AMI (en haut) et MULTISIMO (en bas).

Encodage proportionnel du focus

Dans l'encodage proportionnel du focus, le rapport entre la durée de focalisation sur chaque entité sur le temps total de l'acte de dialogue est utilisé. Un rapport de focalisation pour un participant ou un objet, par exemple X, est calculé comme suit :

(1)
$$\text{Ratio de concentration pour (X)} = \frac{\text{Durée globale de l'orientation vers le participant X}}{\text{Durée totale de l'énoncé}}$$

La figure V contient un exemple de focus proportionnel. Dans l'exemple donné sur AMI, la valeur pour le *Product Manager* (PM) est de 0,4, ce qui signifie que pendant l'énoncé, le participant a regardé le PM pendant 40 % de la durée totale de l'énoncé. De même, il s'est concentré respectivement sur le participant C et sur la table pendant 10 et

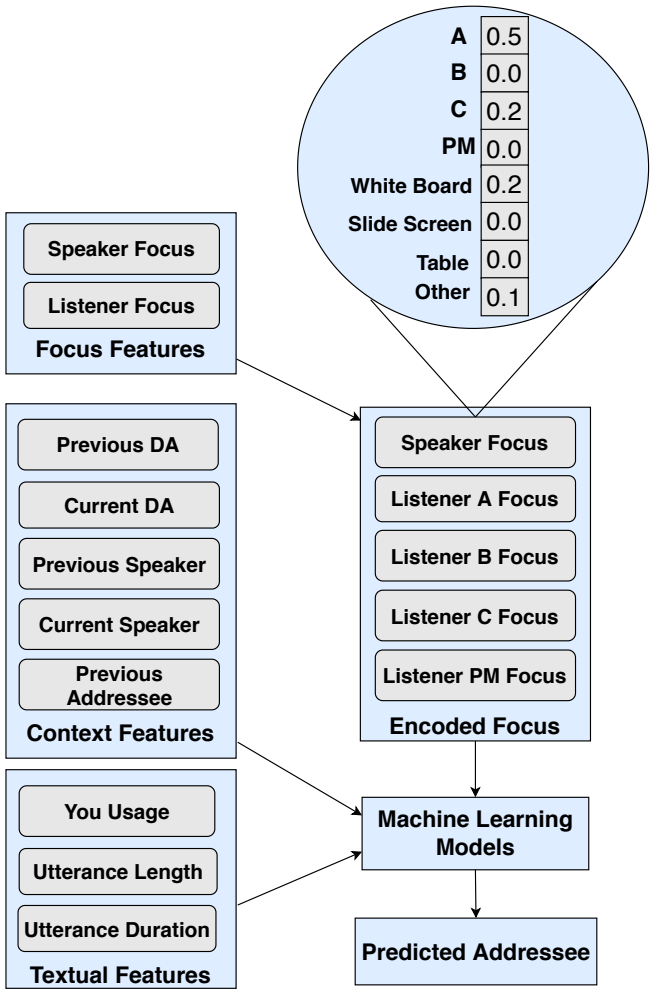

Figure VI – Modèle de détection des destinataires (exemple de codage proportionnel pour le corpus de données AMI)

20 % du temps, tandis qu'il a regardé autre chose pendant les 10 % restants.

Un point important est que l'attention des participants ne semble pas suivre une séquence particulière pendant un énoncé. Lorsqu'un participant regarde une personne ou un objet plusieurs fois, les valeurs des durées de focus correspondantes sont simplement additionnées pour calculer la valeur finale.

Les expériences sont réalisées sur les corpus AMI et MULTISIMO car ce sont les seuls corpus de données qui contiennent toutes les caractéristiques requises pour entrainer des modèles d'apprentissage machine dans le cadre de ce travail.

Modèle de détection des destinataires

La Figure VI montre les caractéristiques utilisées par le modèle de détection des destinataires. Toutes les caractéristiques, à l'exception des caractéristiques de focus, sont directement transmises aux modèles d'apprentissage. Les caractéristiques de focus sont d'abord encodées via un schéma de codage one-hot ou proportionnel, puis transmises aux modèles d'apprentissage.

Pour les prédictions en conditions réelles (r4), le modèle est étudié de deux façons : (i) le destinataire précédent n'est pas utilisé pendant l'entrainement et les tests, et (ii) la valeur réelle du destinataire précédent est exploitée pour l'entrainement et le destinataire précédent prédit est utilisé pour les tests.

Quatre hypothèses sont posées pour l'évaluation du modèle de détection des destinataires. L'hypothèse (h1) affirme que les modèles proposés améliorent la précision du modèle de référence pour AMI et MULTISIMO. L'hypothèse (h2) stipule que le modèle ayant N participants devrait obtenir des performances de classification similaires ou meilleures lorsqu'il est testé sur un ensemble de données avec un nombre de participants inférieur ou égal à N. L'hypothèse (h3-a) affirme qu'en temps réel, la performance de classification pour la détection des destinataires diminue puisque la valeur de vérité de terrain du destinataire précédent n'est pas disponible. L'hypothèse (h3-b) indique que la performance de classification est meilleure lorsque les modèles sont entraînés et testés sans destinataire précédent que lorsque les modèles sont entraînés avec un destinataire précédent issu de la vérité terrain et testés avec un destinataire précédent prédit.

Expérimentations

L'ensemble des expériences réalisées sont résumées dans le Tableau II.

Les expériences sont réalisées par le biais de neuf algorithmes de classification classiques : Multilayer Perceptron (MLP) [Kruse et al., 2013], Random Forest (RF) [Liaw et al., 2002], Logistic Regression (LR) [Hosmer Jr et al., 2013], Support Vector Machines (SVM) [Hearst et al., 1998], Naive Bayes (NB) [Rish et al., 2001], K-Nearest Neighbours (KNN) [Zhang and Zhou, 2005], XGboost (XGB) [Chen and Guestrin, 2016], Adaboost (ADB) [Hastie et al., 2009] et Extra Tree (ET) [Geurts et al., 2006]. Enfin, des tests sont également effectués avec des échantillons de 3 algorithmes d'apprentissage profond: Long Short Term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997], One-Dimensional Convolutional Neural Network (1D-CNN) [Kiranyaz et al., 2019], et Bi-Directional Long Short Term Memory Network (Bi-LSTM) [Melamud et al., 2016].

Experiments	**Description**
AA	Algorithmes entraînés et testés sur AMI
MM	Algorithmes entraînés et testés sur MULTISIMO
AM	Algorithmes entraînés sur AMI et testés sur MULTISIMO
MA	Algorithmes entraînés sur MULTISIMO et testés sur AMI
AM-PN	Algorithmes entraînés et testés sur AMI sans la vérité terrain du destinataire précédent
MM-PN	Algorithmes entraînés et testés sur MULTISIMO sans la vérité terrain du destinataire précédent
AM-PN	Algorithmes entraînés sur AMI et testés sur MULTISIMO sans la vérité terrain du destinataire précédent
MM-PN	Algorithmes entraînés sur MULTISIMO et testés sur AMI sans la vérité terrain du destinataire précédent

Table II – Résumé des expériences menées

Dans la littérature, Akker et Traum [Akker and Traum, 2009] utilisent les informations multimodales du corpus de données AMI pour atteindre une précision de 65% sur la détection des destinataires. Ce travail peut être considéré comme modèle de référence car c'est le seul travail qui respecte les exigences r1, r2 et r3.

Résultats

Les résultats des expériences AA et MM montrent que les schémas d'encodage de focus proportionnel et one-hot améliorent les résultats du modèle de référence pour AMI et MULTISIMO, validant ainsi l'hypothèse h1 qui affirme que le modèle proposé est le plus performant.

En outre, les résultats des expériences AA et MM montrent que le schéma de codage de focus proportionnel est nettement plus performant que le schéma de codage one-hot, pour AMI et MULTISIMO. Les expériences AM et MA sont réalisées pour évaluer les capacités de transfert de l'apprentissage entre corpus. Les résultats montrent que dans ce cas, l'encodage one-hot est plus performant que l'encodage proportionnel.

Les résultats des expériences AA et AM montrent qu'en cas de codage proportionnel, la précision maximale de 78,77 % est atteinte avec l'expérience AA. Cette valeur est supérieure à la précision maximale de 75,55 % avec l'encodage proportionnel pour l'expérience AM où les modèles sont entraînés sur AMI avec 4 participants et testés sur MULTISIMO avec 3 participants. De plus, avec un codage proportionnel, la précision obtenue avec l'expérience AM est supérieure à celle de l'expérience AA. Par conséquent, l'hypothèse h2 qui stipule que N devrait atteindre une performance de classification au moins similaire ou meilleure lorsqu'elle est testée sur un ensemble de données avec

un nombre de participants inférieur ou égal à N, n'est validée que pour l'encodage proportionnel.

La comparaison des résultats pour MM et MA montre que l'inverse de l'hypothèse h2 (les modèles entraînés avec N participants devraient obtenir des performances meilleures ou au moins égales lorsqu'ils sont testés avec plus de N participants) n'est pas vérifiée, ce qui signifie que pour obtenir de meilleurs résultats, les algorithmes doivent être entraînés sur des corpus de données avec plus de participants que le corpus de données testé.

Les expériences AA-PN, MM-PN, AM-PN et MA-PN montrent des résultats pour les modèles entraînés avec un destinataire précédent prédit ou sans destinataire précédent. Les résultats montrent que la précision diminue lorsque les valeurs des destinataires précédents ne sont pas utilisées, ce qui valide l'hypothèse h3-a.

Les résultats globaux montrent que les modèles proposés sont capables de détecter des destinataires (exigence r3) sur des corpus de données multiples, avec un nombre variable de participants (exigence r1), et sans dépendance de positionnement des participants (exigence r2). En outre, il est proposé une variante du modèle qui fonctionne en temps réel, mais qui ne peut utiliser la caractéristique de destinataire précédent (exigence r4).

Prédiction de changement de tour de parole et du prochain orateur

Un propos adressé à un individu ou à un groupe de destinataires, selon l'acte de dialogue et le contexte, peut ou non provoquer une réponse de la part des auditeurs. Par exemple, un orateur peut poser une question à un ou plusieurs participants qui incite l'un d'entre eux à parler et à répondre. Ce changement d'orateur est appelé changement de tour de parole et le processus de répartition des tours parmi les participants d'un dialogue est appelé gestion des tours de parole [Allwood, 2000].

Nous étudions les problèmes de changement de tour de parole et de prédiction du prochain locuteur en utilisant des algorithmes d'apprentissage automatique et une sélection pertinente de caractéristiques. Les modèles ainsi développés montrent que la performance de la prédiction du changement de tour et du prochain locuteur peut être améliorée en utilisant des caractéristiques psycholinguistiques telles que la durée de pause [O'Connell and Kowal, 2012].

Travaux connexes

[Guntakandla and Nielsen, 2015] utilise un arbre de décision J48 pour la prédiction des tours de parole. Le modèle est entrainé sur le jeu de données Switchboard [Godfrey et al., 1992]. Les auteurs font état d'une précision globale de 62,70 % pour la prédiction des changements de tour. [Meshorer and Heeman, 2016] proposent un modèle qui exploite des forêts aléatoires pour prédire les changements de tour. Le modèle est également utilisé sur Switchboard, avec une précision de 76,05 % et un score F1 de 0,74 pour la prédiction des changements de tour. [Aldeneh et al., 2018] considèrent le changement de tour comme un problème de séquence où le changement de tour se produit de manière séquentielle. Ils proposent un modèle de détection de fin de tour qui utilise un LSTM pour apprendre la fin de tour à l'aide des caractéristiques verbales. Le modèle proposé est entrainé sur le corpus Switchboard. Les auteurs rapportent un score F1 de 0,65 pour la prédiction des changements de tour.

En plus du changement de tour, plusieurs chercheurs ont proposé des modèles combinés pour le changement de tour et la prédiction du prochain locuteur. [Kawahara et al., 2012] propose un modèle de changement de tour et de prédiction du prochain locuteur basé sur l'apprentissage machine qui repose sur une combinaison de mouvements du regard, de prosodie et de mouvements de tête. Ils font état d'une précision de 70,60 % pour le changement de tour et de 69,06 % pour la prédiction du locuteur suivant en utilisant des SVM. [Ishii et al., 2013] propose un modèle de prédiction du changement de tour et du prochain locuteur fondé sur les transitions de regard des participants vers la fin d'un énoncé dans une interaction multipartie. Ils obtiennent un score F1 de 0,76 pour le changement de tour et une précision de 59,20 % pour la prédiction du prochain locuteur. [Ishii et al., 2015b] propose un modèle en deux étapes basé sur l'apprentissage machine qui prédit dans un premier temps si un changement de tour se produit ou non et qui prédit dans l'étape suivante qui est le prochain orateur. Le modèle est entrainé via un SVM sur un ensemble de données personnalisé de 4 participants. Ils atteignent une précision de 75,00 % pour la prédiction du changement de tour et de 55,20 % pour la prédiction du prochain orateur.

Caractéristiques pour le changement de tour et la prédiction du prochain locuteur

Les modèles de changement de tour et de prédiction du prochain locuteur s'appuient sur des corpus de données grâce à des techniques d'apprentissage automatique supervisé. La sélection de caractéristiques appropriées pour le changement de tour et la prédiction

du prochain locuteur améliore non seulement le temps d'apprentissage du modèle, mais aussi la précision du modèle.

La liste des caractéristiques pour le changement de tour et la tâche de prédiction du prochain locuteur est compilée sur la base de la pertinence des caractéristiques d'après la littérature. Pour confirmer l'importance des caractéristiques sélectionnées pour le changement de tour et la prédiction du prochain locuteur, une analyse de ces caractéristiques a été effectuée sur les corpus de données AMI et MPR. Les caractéristiques sélectionnées pour le changement de tour et la prédiction du prochain locuteur sont : le rôle du locuteur, le rôle du destinataire, l'instant de début et de fin de l'énoncé, l'acte de dialogue, le focus des participants et la durée de pause entre énoncés.

Modèle pour le changement de tour et la prédiction du prochain orateur

Dans une première approche, deux modèles indépendants pour le changement de tour et la prédiction du prochain locuteur sont proposés. Dans la seconde approche, un modèle connecté pour la prédiction du changement de tour et du locuteur suivant est proposé, dans lequel le changement de tour est prédit dans la première étape, puis le locuteur suivant est prédit en incluant le changement de tour dans l'ensemble des caractéristiques.

Prédictions indépendantes du changement de tour et du prochain intervenant

Compte tenu d'un ensemble de caractéristiques, la première tâche consiste à prévoir si le locuteur de l'énoncé actuel et celui de l'énoncé suivant sont différents ou non. Si le locuteur de l'énoncé actuel et celui de l'énoncé suivant sont différents, le changement de tour s'est produit. La prédiction du changement de tour est un problème de classification binaire puisqu'il n'y a que deux sorties possibles.

La deuxième tâche consiste à prédire qui est le locuteur du prochain énoncé. La prédiction du prochain locuteur est un problème de classification multiclasse puisqu'il y a plus de deux sorties possibles dans une interaction multipartie. Par exemple, dans AMI, le prochain orateur peut être n'importe lequel des trois autres participants. De même, dans le corpus MPR, l'orateur suivant peut être n'importe lequel des participants à l'interaction (A, B, C ou le robot NAO).

La Figure VII montre la méthodologie expérimentale considérant le changement de tour et la prédiction du prochain locuteur comme des modèles indépendants. Les caractéristiques d'entrée, divisées en caractéristiques contextuelles, en caractéristiques

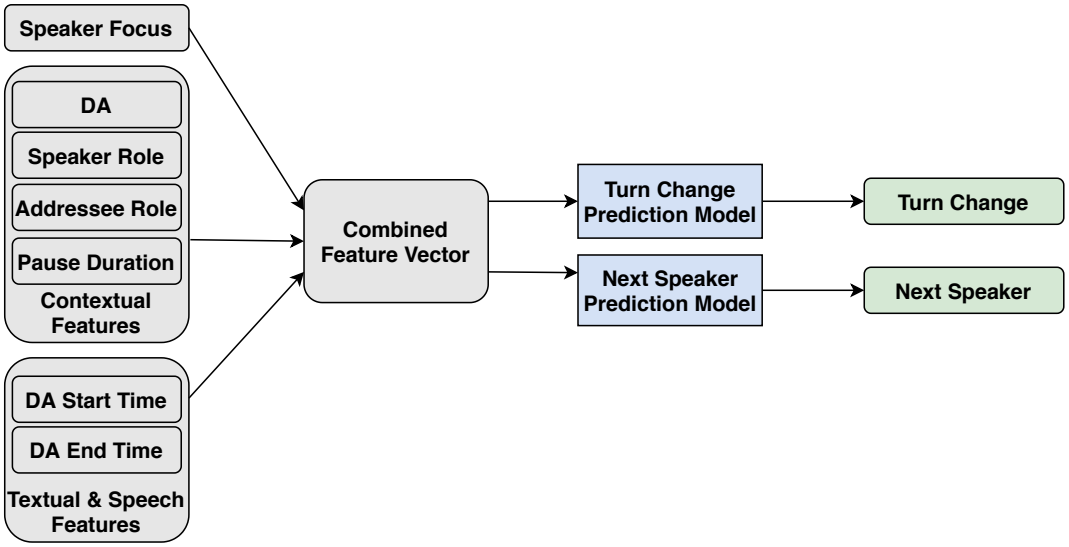

Figure VII – Modèles indépendants de changement de tour et de prédiction du prochain locuteur

de focus et en caractéristiques textuelles et vocales, sont fusionnées pour former un vecteur de caractéristiques combiné qui est utilisé pour entrainer les deux modèles.

Combinaison du changement de tour et de la prédiction du prochain orateur

Le changement de tour et la prédiction du prochain locuteur sont deux tâches liées, car la prédiction du prochain locuteur dépend du changement de tour : un changement de tour signale que le locuteur du prochain énoncé ne peut pas être le locuteur actuel et qu'une personne différente du locuteur actuel doit parler ensuite. Ainsi, la prédiction du changement de tour peut également être utilisée comme une fonction supplémentaire.

La Figure VIII illustre la méthodologie expérimentale permettant de combiner les prédictions de changement de tour et de locuteur suivant. Le prochain locuteur est prédit en deux étapes : 1) les caractéristiques d'entrée sont utilisées pour prédire le changement de tour de la même manière que dans la Figure VII ; 2) la valeur du changement de tour prédit est ensuite exploitée comme caractéristique d'entrée supplémentaire, afin de prédire le prochain orateur. Les valeurs de changement de tour réelles sont utilisées pour entrainer le modèle, tandis qu'en cours d'exécution, la valeur de changement de tour prédite est exploitée en tant que caractéristique supplémentaire.

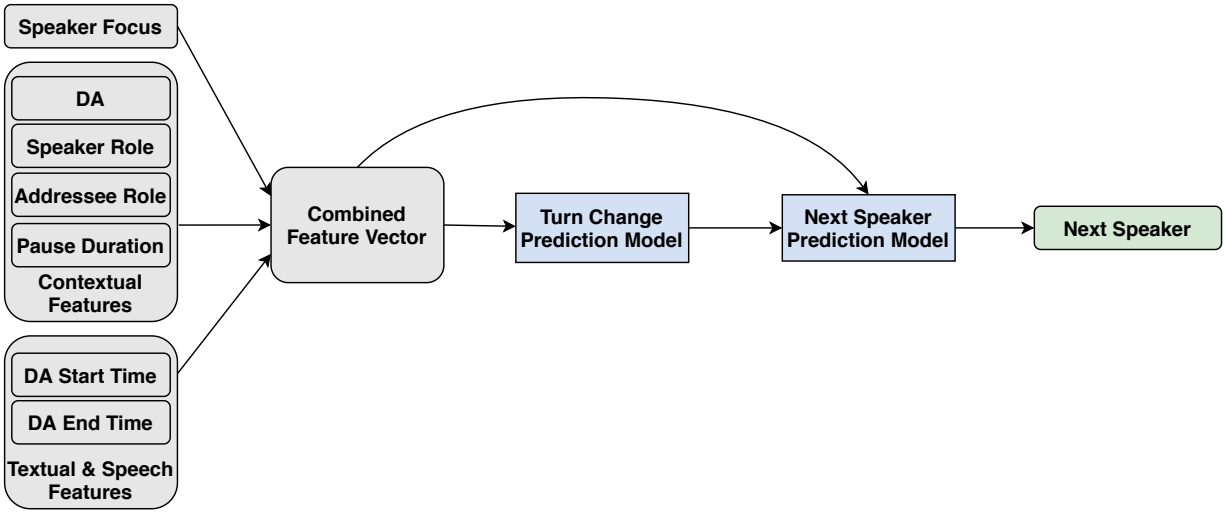

Figure VIII – Modèle combiné de changement de tour et de prédiction du prochain locuteur

Expérimentations

Trois séries d'expériences sont réalisées : (i) pour évaluer individuellement la performance des modèles de prédiction du changement de tour et du locuteur suivant (ii) pour la prédiction du locuteur suivant en utilisant le changement de tour prédit, et (iii) pour évaluer l'importance de certaines des caractéristiques.

Les corpus de données utilisés pour entrainer et tester les modèles de prédiction de changement de tour et du locuteur suivant sont AMI et MPR [Funakoshi, 2018], étant les seuls corpus qui contiennent des données annotées pour toutes les caractéristiques sélectionnées.

Un pipe-line d'apprentissage classique est suivi pour réaliser les expériences. Les caractéristiques catégorielles sont codées one-hot pour les convertir en une forme numérique. L'ensemble des caractéristiques est normalisé à l'aide d'une mise à l'échelle standard. Les ensembles d'entrainement et de test sont créés pour une validation croisée à 5 blocs.

Huit des classifieurs les plus classiques de l'apprentissage machine ont été entraînés et testés: XGboost (XGB) [Chen and Guestrin, 2016], Multilayer Perceptron (MLP) [Kruse et al., 2013], Random Forest (RF) [Liaw et al., 2002], Logistic Regression (LR) [Hosmer Jr et al., 2013], Support Vector Machines (SVM) [Hearst et al., 1998], Naive Bayes (NB) [Rish et al., 2001] et K-Nearest Neighbours (KNN) [Zhang and Zhou, 2005]. En outre, pour évaluer si les problèmes de changement de tour et de prédiction du prochain locuteur peuvent être considérés comme séquentiels, des tests sont également effectués en utilisant une combinaison de Long Short Term Memory Network [Hochreiter and Schmidhuber, 1997] et Densely Connected Neural network (DNN) dans un perceptron

multicouches.

Pour évaluer nos modèles de prédiction de changement de tour, deux modèles de référence sont sélectionnés : (i) [Meshorer and Heeman, 2016], et (ii) classe majoritaire. Le premier modèle de référence est sélectionné parce que (a) il offre les meilleures performances sur le corpus Switchboard et (b) les caractéristiques qu'il utilise sont disponibles dans la plupart des corpus de données existants et les résultats peuvent donc être reproduits. Pour évaluer le modèle de prédiction du prochain locuteur, la classe majoritaire est choisie comme modèle de référence par défaut de modèle comparable.

Dans les modèles proposés, deux nouvelles caractéristiques (durée de la pause et rôle du destinataire) sont ajoutées. À notre connaissance, ces caractéristiques n'ont pas encore été exploitées pour le changement de tour et la prédiction du prochain orateur. Des expériences sont réalisées avec et sans ces deux caractéristiques de façon à évaluer si elles améliorent de manière significative les prédictions.

Résultats

Les résultats sont divisés en trois sections : (i) résultats des expériences pour les modèles de prédiction de changement de tour individuel et du prochain locuteur, (ii) résultats pour le modèle de prédiction du prochain locuteur utilisant le changement de tour prédit, et (iii) étude de pertinance des caractéristiques.

Résultats pour les modèles individuels de changement de tour et de prédiction du prochain intervenant

Les résultats montrent que pour les corpus de données MPR et AMI, le modèle de changement de tour proposé est plus performant que les deux modèles de référence. Une précision moyenne maximale de 85,83 % est obtenue sur AMI en utilisant l'algorithme XGboost, ce qui est meilleur que les 61,57 % obtenus par le modèle de référence 1 et les 54,55 % obtenus par le modèle de référence 2. En ce qui concerne MPR, le modèle proposé atteint une précision maximale de 83,08 %, ce qui est meilleur que le modèle de référence 1 (76,89 %) et le modèle de référence 2 (74,37 %).

Pour les expériences réalisées pour la prédiction du prochain locuteur à l'aide des corpus de données AMI et MPR, les résultats montrent que pour AMI, la meilleure précision de 64,77 % est obtenue par l'algorithme XGboost, ce qui est meilleur que la précision du modèle de référence de 35,77 %. De même, pour MPR, une précision maximale de 64,73 % est obtenue grâce à l'algorithme XGBoost, qui surpasse la précision du modèle de référence de 38,75 %.

Résultats pour le modèle combinant le changement de tour et la prédiction du prochain intervenant

Les résultats montrent que pour les données d'AMI, une précision maximale de 65,53 % et une valeur F1 de 0,65 sont obtenues par l'algorithme XGBoost. Cette valeur est supérieure à la valeur obtenue (64,77% et F1= 0,64) lorsque le modèle de prédiction du prochain locuteur est considéré comme un modèle individuel.

En ce qui concerne MPR, une précision maximale de 64,46% et une valeur F1 de 0,63 sont obtenues en utilisant l'algorithme XGBoost, ce qui est similaire aux valeurs (64,73 % et F1=0,64) obtenues via le modèle indiviudel de prédiction du prochain locuteur.

Résultats de l'étude sur les caractéristiques

Les résultats montrent que pour AMI et MPR, dans le meilleur des cas (en utilisant l'algorithme XGBoost), les modèles entraînés en utilisant à la fois le rôle du destinataire et la durée de la pause surpassent à la fois les modèles entraînés sans ces caractéristiques et ceux entraînés en n'incluant qu'une de ces caractéristiques. L'importance du rôle du destinataire et de la durée de la pause a été étudiée individuellement en ce qui concerne la prédiction du changement de tour. Pour AMI et MPR, les meilleures performances obtenues par l'algorithme XGBoost en incluant le destinataire sont légèrement supérieures ou similaires aux modèles qui n'utilisent pas le rôle du destinataire.

Génération du comportement d'attention visuelle

Bien que la communication verbale soit le principal mode d'interaction entre les humains, les comportements non verbaux tels que les gestes [McNeill, 1992] et l'attention visuelle [Argyle and Ingham, 1972] peuvent améliorer et renforcer la communication verbale pour transmettre des états mentaux. Le comportement d'attention verbale (CAV) fait référence aux cibles, c'est-à-dire les participants ou les objets compagnons qui sont au centre de l'attention d'un participant pendant une interaction. Le CAV est un comportement non verbal qui peut être généré en identifiant à tout moment au cours d'une interaction le focus d'une personne.

Dans cette thèse, une approche hybride est proposée pour développer des modèles CAV dans les interactions multiparties pour les orateurs et pour les auditeurs. Les modèles proposés de génération de comportement CAV sont divisés en sous-modèles où chaque modèle résout un problème spécifique. Les modèles sont basés sur l'apprentissage machine ainsi que sur des approches heuristiques à base de règles. La fonctionnalité

combinée des modèles génère le comportement global du CAV pour un compagnon artificiel à un moment donné.

Travaux connexes

Les approches de développement des modèles d'attention visuelle peuvent être classées en trois grandes catégories : (i) les systèmes d'inspiration biologique, (ii) les systèmes fondés sur les données et (iii) les systèmes heuristiques.

Parmi les approches d'inspiration biologique, [Hoffman et al., 2006] propose un modèle d'imitation du regard et d'attention partagée qui combine un algorithme d'estimation de la direction du regard avec des cartes des points saillants des scènes visuelles pour prédire les objets regardés. [Trafton et al., 2008] proposent un modèle de génération du regard en contexte multipartie appelé ACT(R/E) qui dirige son regard sur les orateurs afin d'effectuer un suivi de la conversation. [Lee et al., 2007] propose un modèle qui intègre étroitement le regard dans le modèle cognitif sous-jacent contrôlant le raisonnement. Le modèle est implémenté dans des agents virtuels pour la génération du regard du locuteur et de l'auditeur dans des scénarios multiparties.

Parmi les modèles fondés sur les données, [Liu et al., 2012] développe une approche basée sur des règles pour le hochement de tête, l'inclinaison de la tête et la génération du regard dans une interaction multipartie. Les auteurs utilisent un ensemble de données personnalisé pour l'extraction des règles. [Pelachaud and Bilvi, 2003] propose un modèle statistique de génération du regard en interaction dyadique basé sur les réseaux bayésiens. Le modèle utilise un corpus de vidéos de 20 à 30 minutes où deux utilisateurs interagissent. [Admoni and Scassellati, 2014] développe un modèle de génération de comportement non verbal pour une application de tutorat dans des contextes dyadiques. Le modèle est entrainé à l'aide d'un corpus de données personnalisé de deux humains où l'un agit en tant qu'enseignant tandis que l'autre agit en tant qu'élève.

Dans la catégorie des systèmes heuristiques, [Mao et al., 2009] exploite le mouvement des yeux extrait à partir de l'expression faciale [Kanade et al., 2000] et des données sur les mouvements oculaires en temps réel (taux de clignement des yeux, taille de la pupille et saccade), afin de développer une approche à base de règles pour générer des comportements d'attention visuelle qui représentent diverses émotions primaires (joie, tristesse, colère, peur, dégoût et surprise) et intermédiaires (émotions qui peuvent être représentées comme le mélange de deux émotions). [Sisbot and Alami, 2012] propose un modèle de génération du comportement du regard pour un robot qui remet un objet à des humains dans un cadre multipartie. En utilisant les informations du regard de l'humain, les robots planifient le lieu de remise et transmettent ensuite ces informations

à l'humain en regardant l'endroit où l'objet sera remis. [Peters et al., 2005] développe un modèle de génération du regard qui surveille le regard des participants pour estimer leur niveau d'engagement, puis génère des mouvements du regard pour maintenir et renforcer l'engagement des participants dans la conversation.

Caractéristiques pour la génération du comportement d'attention visuelle

Comme précédemment, la liste des caractéristiques est proposée sur la base de la pertinence des caractéristiques pour le comportement de génération de CAV du locuteur et de l'auditeur comme mentionné dans la littérature, confirmée par l'analyse de ces caractéristiques sur les corpus de données AMI [Carletta, 2007] et MPR [Funakoshi, 2018]. Les caractéristiques sélectionnées pour le comportement de génération de CAV du locuteur et de l'auditeur sont les suivantes : locuteurs actuels et précédents, destinataire actuel et précédent, acte de dialogue, durée de l'énoncé et liste des participants.

Modèle pour le changement de tour et la prédiction du prochain orateur

Au cours d'une interaction mixte, impliquant des humains et des agents intelligents, l'agent peut parler et écouter d'autres agents ainsi que des locuteurs humains.

Toutes les caractéristiques sont disponibles pour la génération du comportement du locuteur et de l'auditeur lorsqu'un agent parle, si celui-ci les transmet aux autres agents. Cependant, une difficulté survient lorsqu'un humain parle car dans ce cas, la durée de l'énoncé, le temps de parole et la fin de l'énoncé ne peuvent être connus qu'à la fin de l'énoncé.

Par conséquent, le modèle de génération du CAV est divisé en trois sous-modèles : le modèle de génération du CAV du locuteur, abrégé en CAV-L, et le modèle de génération du CAV de l'auditeur, abrégé en CAV-A. Le modèle CAV-A se divise en deux sous-types : le modèle de génération de comportement CAV de l'auditeur lorsqu'un agent parle (CAV-AA) et le modèle de génération de comportement CAV de l'auditeur lorsqu'un humain parle (CAV-AH).

Génération de CAV du locuteur (CAV-L) et de l'auditeur lorsqu'un agent parle (CAV-AA)

La Figure IX décrit l'architecture du modèle proposé de génération du CAV et illustre un exemple de la manière dont les données circulent entre les éléments dans les conditions expérimentales de MPR [Funakoshi, 2018].

Le modèle comporte quatre sous-modèles principaux : un prédicteur de nombre de tours d'attention visuelle, un prédicteur de durée des tours, un prédicteur de cible de l'attention, et enfin un planificateur de CAV.

Le Prédicteur du nombre de tours d'attention visuelle est un sous-modèle basé sur l'apprentissage machine qui prédit le nombre de tours pendant l'acte de dialogue en cours, sous forme de problème de régression. Par exemple, dans la figure IX, la sortie du prédicteur du nombre de tours d'attention visuelle est 2,8, puis arrondie au nombre entier le plus proche, c'est-à-dire 3.

Le prédicteur de cible de l'attention prédit les personnes et/ou les objets dans le focus visuel de l'agent. De même que le prédicteur de durée des tours, le prédicteur de cible de l'attention produit un vecteur de taille 5 dans les conditions MPR et 6 dans les conditions AMI. La figure IX illustre un exemple de vecteur de sortie du prédicteur de cible de l'attention. Le vecteur de sortie est un vecteur codé en one-hot où 1 correspond à la cible CAV prédite. Les cibles CAV prédites dans la Figure IX sont B, NAO et autres.

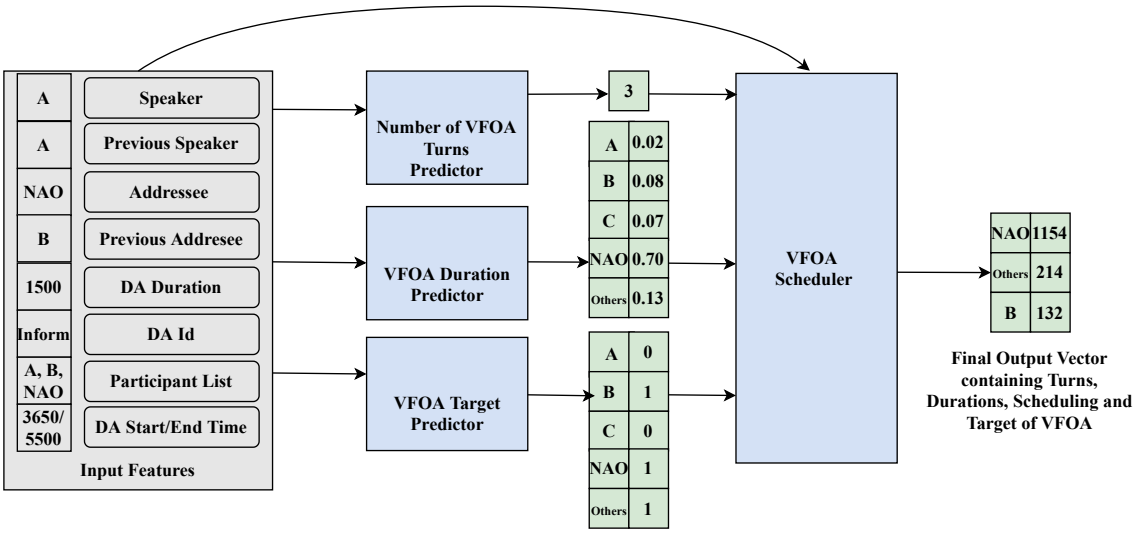

Figure IX – Modèle de génération de comportement CAV-L et CAV-AA (valeurs tirées de MPR)) [Funakoshi, 2018]

Le prédicteur de durée des tours est également un sous-modèle basé sur l'apprentissage machine qui prédit la durée totale du CAV par cible pendant un acte de dialogue. Par

exemple dans MPR, la cible du CAV peut être un participant parmi A, B, C, NAO ou autre, donc la longueur du vecteur de sortie des probabilités est de 5. La Figure IX montre un exemple de vecteur de sortie du prédicteur de durée des tours. Ce vecteur indique que la durée de focus dirigée vers A est de 2%, celle dirigée vers NAO est de 70% et ainsi de suite.

Les tours de focus visuel ne semblent pas suivre de schéma spécifique. Par exemple, dans MPR, il y a un total de 10 214 énoncés qui contiennent plus d'un tour de CAV. Pour ces 10 214 énoncés, le nombre total de séquences de focus uniques est de 470. Cela montre que la planification des tours d'attention visuelle est très difficile et que les approches d'apprentissage machine ne sont pas adaptées pour effectuer cette planification. Une approche heuristique est donc menée, afin de proposer des modèles spécifiques à l'orateur et aux auditeurs.

Génération du comportement de l'auditeur lorsqu'un humain parle (CAV-AH)

Le modèle de génération de comportement CAV-AH se compose également de quatre sous-modèles, cependant seul le prédicteur de cible de l'attention se base sur l'apprentissage automatique. La Figure X décrit l'architecture du modèle de génération de comportement CAV-AH avec un exemple de la manière dont les données circulent entre les éléments dans les conditions expérimentales de MPR [Funakoshi, 2018].

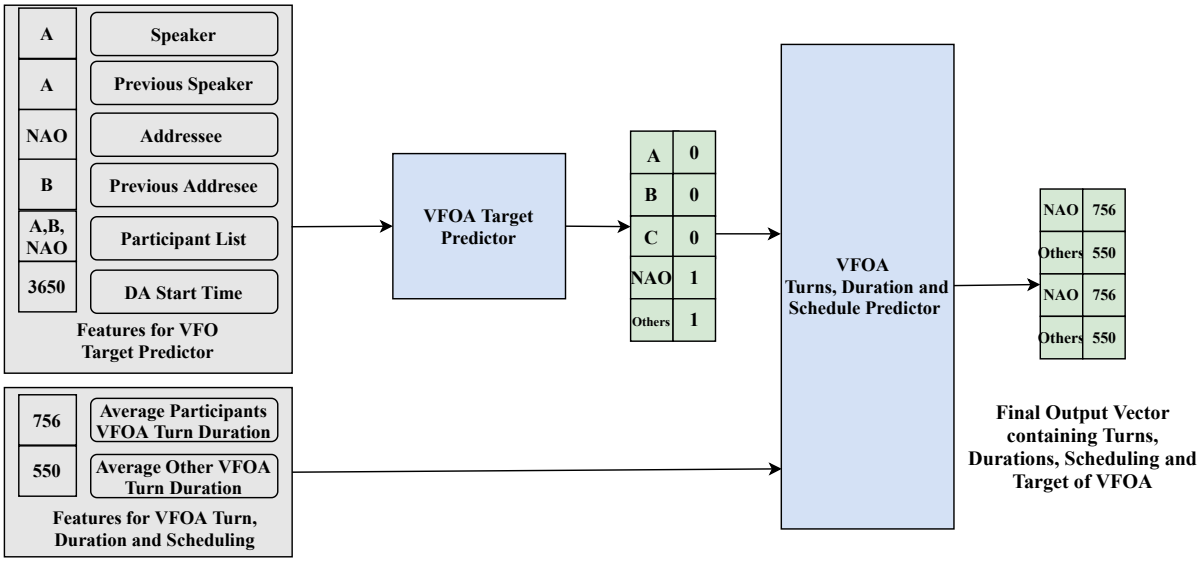

Figure X – Modèle de génération de comportement CAV-AH

Le prédicteur de cible de l'attention prédit une liste de participants ou d'objets qu'un participant peut regarder pendant un énoncé, de façon similaire aux modèles CAV-L et CAV-AA. La seule différence est que le CAV-AH n'exploite pas la durée de l'énoncé,

l'heure de fin de l'acte de dialogue et sa durée. La figure X illustre un exemple de vecteur de sortie du prédicteur de cible de l'attention. Le vecteur de sortie est un vecteur encodé one-hot où 1 correspond aux cibles prédites.

Pour le nombre de tours, la durée par tour et la planification des tours, des modèles heuristiques sont proposés car si la durée de l'énoncé n'est pas connue avant sa fin, alors la prédiction d'une valeur fixe pour le nombre de tours peut entrainer des durées de tour très longues ou très courtes. Par exemple, si le nombre de tours est prédit à 1 et que le locuteur humain continue de parler pendant une très longue durée, le focus sur une seule cible devient déraisonnable. Par conséquent, au lieu d'utiliser des modèles d'apprentissage machine qui produisent une valeur fixe, une approche heuristique est proposée qui incrémente dynamiquement le nombre de tours d'attention visuelle. Pour la durée totale par cible dans un énoncé, les valeurs moyennes de durée du tour par cible potentielle sont utilisées à partir des corpus de données correspondants.

Expérimentations

Le modèle de génération de CAV du locuteur et du CAV de l'auditeur ainsi que les modèles de base suivants sont développés et évalués au cours d'une expérimentation permettant à des spectateurs d'observer et évaluer plusieurs scènes de dialogues multiparties, générées à l'aide de deux modèles de référence, de notre modèle, et des données réelles.

Modèles de référence

Modèle de référence 1 : Baseline-Random Dans le premier modèle de référence, le comportement d'attention visuelle de l'orateur est généré de manière aléatoire. Le nombre de tours par énoncé est choisi au hasard en fonction de la probabilité globale du nombre de tours par énoncé dans les corpus de données correspondants. La durée du focus par tour est calculée en divisant le nombre de tours par énoncé par la durée totale de l'énoncé. La cible de chaque tour est sélectionnée de manière aléatoire dans la liste des auditeurs et des objets, selon la distribution par défaut des directions du focus dans le corpus de données correspondant.

Modèle de référence 2 : Baseline-rule-based Dans le second modèle de référence, le CAV est généré sur la base d'un ensemble de règles :

1. Le nombre de tours est généré comme une fonction linéaire de la durée de l'énoncé en utilisant un algorithme de régression linéaire sur le corpus étudié.

2. La durée du focus par tour est calculée en divisant le nombre de tours par la durée totale de l'énoncé.

3. **Cible des tours pour les orateurs** : (i) si le destinataire du tour est un groupe, le participant ou l'objet pour la cible de l'attention est choisi au hasard dans la liste des destinataires et des objets de manière circulaire, (ii) si le destinataire est un individu, la cible pour le premier tour est ce participant. Pour les tours restants, les participants ou les objets sont sélectionnés de manière aléatoire dans la liste des destinataires et des objets.

   Cible des tours pour l'auditeur : (i) si un agent est destinataire, le locuteur est défini comme cible pour l'agent, (ii) si le destinataire du message est un groupe ou si un agent est adressé indirectement, le locuteur est défini comme cible pour le premier tour, puis pour les tours restants, la cible est choisie aléatoirement.

Hypothèses

Les hypothèses de ces travaux de recherche sont présentées dans le Tableau III.

Questionnaires d'enquête

Les utilisateurs qui évaluent le modèle de génération de comportement d'attention visuelle par rapport aux modèles aléatoires et à base de règles, ainsi qu'au comportement réel sont présentés avec des paires de vidéos. Dans chaque paire, une vidéo présente un comportement généré par le modèle CAV, tandis que le comportement CAV de l'autre vidéo est généré par un des trois autres modèles.

Pour chaque paire de vidéos, quatre questions sont posées. Les questions visent à valider les hypothèses proposées dans le Tableau III.

Une échelle de Likert à sept points (1-7) [Harpe, 2015] a été utilisée pour enregistrer les réponses à chaque questionn de *completely unnatural* à *fully natural*.

Mise en œuvre du modèle et génération des vidéos

L'organisation de l'expérimentation nécessite le développement de vidéos d'interaction contenant les différents comportements d'attention visuelle : le modèle proposé, les deux modèles de référence, et les valeurs réelles issues du corpus. Dans chaque vidéo d'interaction, des agents intelligents jouent le rôle des participants réels dans des scénarios du corpus AMI. Le processus global de mise en œuvre du modèle et de génération de vidéos est illustré dans la figure XI.

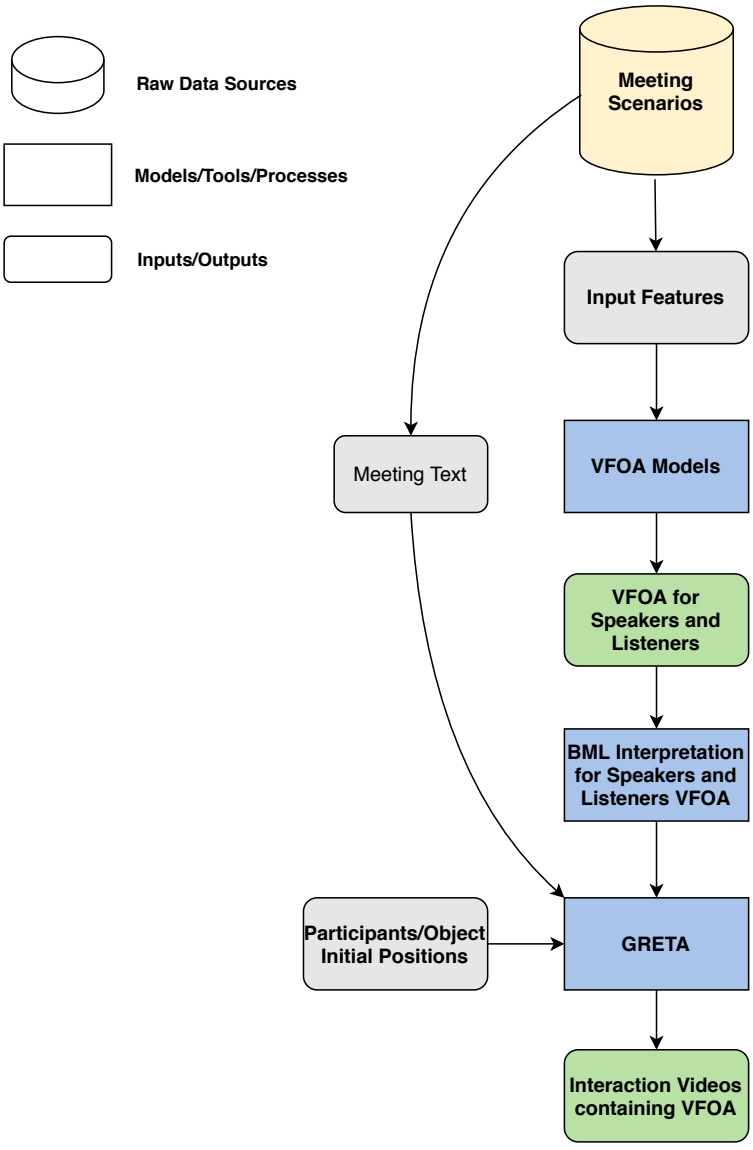

Figure XI – Processus de génération du CAV pour les expériences

Hypothèses	Description
h1(s)	Pour les locuteurs, le nombre de tours prédit par le modèle proposé est perçu comme plus naturel que le nombre de tours des deux modèles de base aléatoire, et aussi naturel que le comportement réel d'attention visuelle
h1(l)	Pour les auditeurs, le nombre de tours prédit par le modèle proposé est perçu comme plus naturel que le nombre de tours des deux modèles de base aléatoire, et aussi naturel que le comportement réel d'attention visuelle
h2(s)	Les durées prédites par le modèle sont perçues comme plus naturelles que celles prévues par les 2 modèles de référence, et aussi naturelles que le comportement réel pendant que l'agent parle
h2(l)	Les durées prédites par le modèle sont perçues comme plus naturelles que celles prévues par les 2 modèles de référence, et aussi naturelles que le comportement réel pendant que l'agent est à l'écoute
h3(s)	Les cibles prédites par le modèle sont perçues comme plus naturelles que les cibles issues des modèles aléatoires et basées sur des règles, et aussi naturelles que le comportement réel d'attention visuelle, pendant que l'agent parle
h3(l)	Les cibles prédites par le modèle sont perçues comme plus naturelles que les cibles issues des modèles aléatoires et basées sur des règles, et aussi naturelles que le comportement réel d'attention visuelle, pendant que l'agent écoute
h4(s)	Pour les locuteurs, le comportement global d'attention visuelle est perçu comme plus naturel que les comportement aléatoires et ceux basés sur les règles, et aussi naturel que le comportement réel
h4(l)	Pour les auditeurs, le comportement global d'attention visuelle est perçu comme plus naturel que les comportement aléatoires et ceux basés sur les règles, et aussi naturel que le comportement réel

Table III – Hypothèses pour comparer l'approche proposée avec les modèles de référence

Les scénarios d'interaction contiennent des données brutes qui sont utilisées pour entrainer ou calibrer les modèles. Les modèles CAV exploitent les caractéristiques d'entrée, extraites des données du corpus, pour prédire le comportement d'attention visuelle des participants. Ces comportements sont convertis en instructions BML (Behaviour Markup Language) pour les agents intelligents dans GRETA [Niewiadomski et al., 2009].

Les entrées de GRETA sont ainsi le texte des scénarios, les instructions BML pour la génération du comportement et la position initiale des participants et des objets dans l'interaction. Quatre agents virtuels sont créés pour ces expériences dans chaque vidéo, correspondant aux quatre participants du corpus AMI. Des ordinateurs portables sont utilisés comme objets pour chaque participant. Une valeur seuil de 700 millisecondes est utilisée pour la durée des tours car les tours de regard pour des durées inférieures à 700 millisecondes entrainent des changements très brusques de l'attention qui ne sont pas

S.No	Question	Validation des hypothèses
1	How natural does the overall visual focus of attention look for the speakers and Listeners in the VIDEO-A and VIDEO-B? (Visual Focus of attention refers to the head and eye movements and the places where the participants look at during an utterance)	h4(s,l)
2	For speakers and listeners in VIDEO-A and VIDEO-B, how natural does the target of the visual focus of attention looks? (Visual focus of attention target is the participants or objects that the speakers and listeners look at during interaction)	h3(s,l)
3	For speakers and listeners in the VIDEO-A and VIDEO-B, how natural does the number of changes in the visual focus of attention look? (Visual focus of attention changes refers to head shifts and gaze shifts (change in target) during an utterance)	h1(s,l)
4	For speakers and listeners in the VIDEO-A and VIDEO-B, how natural does the duration for the visual focus of attention looks? (Visual focus of attention duration refers to the duration for meeting participants when they look at other participants or objects without changing VFOA target.)	h2(s,l)

Table IV – Questions et hypothèses connexes

naturels pour l'œil humain.

Résultats

Les résultats montrent que tant pour les orateurs et les auditeurs, le modèle de génération de l'attention visuelle proposé est perçu comme plus naturel que les deux modèles de base (aléatoire et fondé sur des règles), et aussi naturel que le comportement réel du CAV pour : (i) la prédiction du nombre de tours, qui valide l'hypothèse h1(s,l), (ii) la prédiction de la durée des tours, qui vérifie l'hypothèse h2(s,l), (iii) la prédiction des cibles de l'attention, qui confirme l'hypothèse h3(s,l), et (iv) la génération globale de CAV qui satisfait l'hypothèse h4(s,l).

Concernant le comportement des locuteurs, le modèle proposé obtient une note de 5 à 6 pour le CAV global, les cibles du CAV, les changements de tour et la durée des focus, ce qui montre que, selon les utilisateurs, le CAV généré par notre modèle est perçu comme naturel à très naturel. Ces résultats sont similaires aux notes attribuées par les utilisateurs pour les valeurs réelles du CAV. Pour les auditeurs, le modèle proposé obtient des notes moyennes de 4 à 5, ce qui correspond à un niveau acceptable selon

l'échelle utilisée. Les résultats sont similaires aux vidéos générées à l'aide des valeurs réelles de CAV pour les auditeurs. Il n'y a pas de différence statistique significative de performance entre le modèle CAV proposé et le comportement réel pour les orateurs et les auditeurs, ce qui montre que le modèle de CAV proposé peut être utilisé pour générer des comportements qui sont perçus comme aussi naturels que le comportement réel.

Les résultats montrent en outre que, tant pour les orateurs que pour les auditeurs, les évaluations des utilisateurs pour les sous-tâches de génération de CAV sont très similaires. Cela peut s'expliquer par le fait que toutes les sous-tâches du CAV contribuent de manière égale à la génération du comportement global.

Conclusion et Perspective

Dans cette thèse, nous étudions et améliorons trois aspects d'un agent intelligent interagissant dans des scénarios multiparties et multimodaux : (i) la capacité perceptive de l'agent à identifier le destinataire d'un énoncé, (ii) la capacité décisionnelle de l'agent à détecter le changement de tour et le prochain locuteur, et (iii) la capacité de génération du comportement de l'agent à générer un comportement d'attention visuelle (CAV). Les modèles sont basés sur la sélection de caractéristiques pertinentes et utilisent des approches d'apprentissage machine, parfois couplées à des approches à base de règles.

Bien que les modèles individuels soient plus performants que les modèles de référence, l'intégration des trois modèles pour développer un système unifié capable de prédire le destinataire, de détecter le changement de tour et le prochain locuteur, puis de générer le comportement d'attention visuelle des agents est une tâche difficile. Une autre limite importante est la propagation des erreurs en temps réel. Dans certains cas, les modèles proposés dépendent de la valeur prédite précédemment.

La nature dynamique des données est un second challenge majeur lors de l'interaction en temps réel. Par exemple, les modèles de changement de tour et de prochain locuteur utilisent la durée de pause qui est le temps entre deux prononciations successives. La durée de la pause est une valeur dynamique puisque la durée ne cesse d'augmenter jusqu'à ce que le prochain énoncé commence.

L'indisponibilité de corpus de données contenant toutes les caractéristiques pertinentes est une dernière difficulté. Enfin, en raison des conditions sanitaires imposées par COVID 19, certaines des expériences, telles que la génération du comportement des auditeurs en cas de parole humaine dans des interactions en temps réel, n'ont pas pu être réalisées.

CHAPTER 1

INTRODUCTION

Advancements in artificial intelligence techniques have led to a growing interest in systems where humans and artificial agents interact with each other to achieve a common goal. Such systems are commonly known as human-agent interaction systems. Human-agent interaction is a research area that deals with the understanding, design, development and evaluation of intelligent agents capable of interacting with humans. Human-agent interaction systems are employed in various domains to achieve different tasks e.g. personal assistants, training and tutoring systems, chat bots, service robots, robot nurses and doctors, etc. [Kleinerman et al., 2018; Rosenfeld et al., 2017, 2015; Salem et al., 2015; Sheh, 2017; Traum et al., 2003].

Embodied Conversational Agents (ECA) are one of the most common types of intelligent agents involved in human-agent interaction. ECAs in the form of robots and virtual agents have gained popularity in the last decade. Airbus's Tim and Toshiba's Yoku are common examples of virtual agents. In addition, ECAs have also been implemented in the form of physical robots such as Sophia [Weller, 2017], Nao [Jokinen and Wilcock, 2014] and Ocean One [Khatib et al., 2016]. The ultimate goal of human-agent interaction is to enable intelligent agents to propose natural interaction with humans and other agents using multiple modalities.

In this thesis, we aim at improving human-agent communication as a whole, by proposing solutions for several sub-tasks in multimodal, multiparty human-agent interaction.

1.1 Context and Motivation

1.1.1 Multimodal & Multiparty Human-Agent Interaction

In the context of human–agent interaction, *a modality is the classification of a single independent channel of sensory input/output between a computer and a human* [Karray et al., 2008]. A modality can thus be defined as a mode of communication between humans and agents. From a human perspective, modalities include sense of touch, sight, hearing, taste, and smell. From an agent perspective, modalities include input and output devices that mimic human senses. For example, a microphone is used to capture human speech, a camera captures vision and replicates human sight, a haptic sensor mimics touch inputs, etc. In addition, some input devices do not match any human sense such as mouse, keyboard, writing tablet, etc. Multimodal human-agent interaction can therefore be defined as interaction between humans and intelligent agents involving multiple modalities.

Figure 1.1 contains an example of a typical cycle of multimodal human-agent interaction. The figure shows that human expresses its intention, attention or emotion with the help of actions such as speech, gesture, etc. An agent perceives human actions via input devices such as microphones, video camera, etc. To generate a response, input data received from humans and surrounding environment is processed to make a final decision regarding the action that has to be taken in response to the input. To express an output or action, agents use devices such as text to speech converter, etc. The actions of an agent are perceived by humans via senses such as sight, hearing, etc. This process continues throughout the interaction between humans and agents.

Multimodal human-agent interaction lies at the crossroad of different research areas including artificial intelligence, computer vision, psychology, etc. Owing to the rise in pervasive and ubiquitous computing, computers have become integrated into our daily lives. In many applications, humans need to interact with agents in a way similar to how they interact with each other. For example, in human-human interactions, on average, 65% of the meaning of communicated message is inferred via non-verbal behaviours [Foley and Gentile, 2010]. Therefore to ensure natural interaction between humans and agents, agents should be capable to handle multimodal inputs and exhibit multimodal behaviour.

Moreover, humans-agent interactions can be dyadic, involving two participants, or multiparty, where more than two participants interact as illustrated in Figure 1.2.

Interaction tasks are simpler in dyadic interaction compared to multiparty interaction. For instance in dyadic interaction, the addressee of an utterance is known since there is

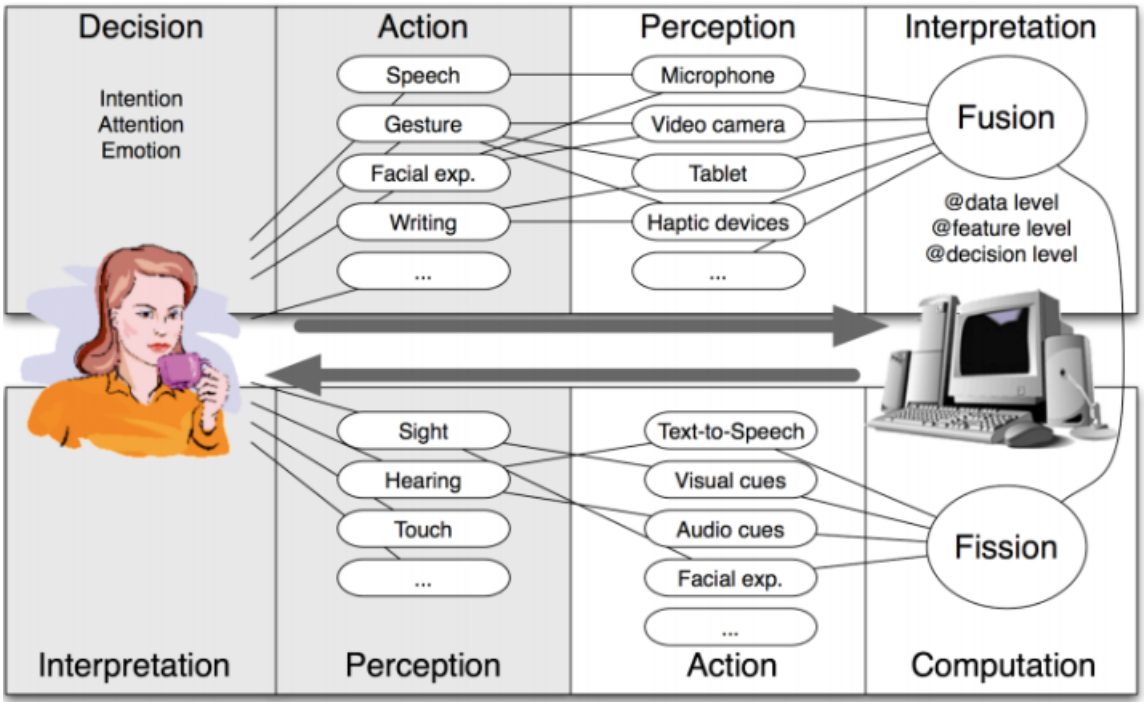

Figure 1.1 – Multimodal Human-Agent Interaction.

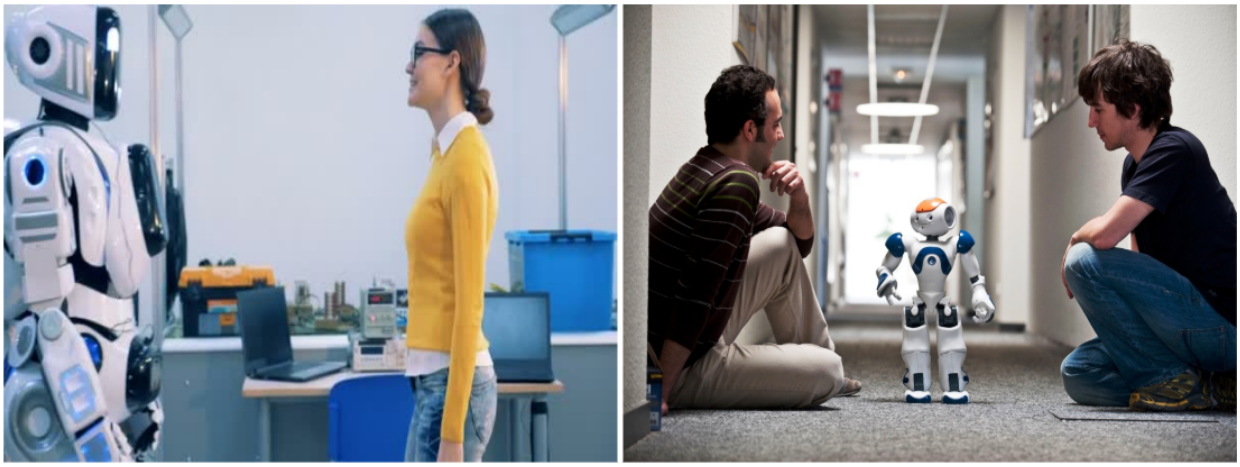

Figure 1.2 – Dyadic Interaction (Left) vs Multiparty Interaction (Right)

only one listener. Similarly, the focus of attention of the participants is relatively straight forward as in addition to objects, there is only one participant to look at. On the contrary, decision making for interaction tasks is relatively complex in multiparty interaction. For example predicting the speaker of the next utterance after the current speaker finishes is a complex task owing to the presence of multiple participants.

Though rule-based techniques can be used for human-agent interaction, the availability of data and improvements in computing power has tilted the balance towards

using of machine learning techniques for human-agent interaction.

1.1.2 Machine Learning in Human-Agent Interaction

Machine learning is a sub-field of artificial intelligence whose main goal is to program computers to use example data or past experience to solve a given problem [Alpaydin, 2020]. Machine learning algorithms learn patterns from data. A new data is classified or clustered based on patterns identified by machine learning algorithms. Machine learning techniques are widely exploited by researchers to solve various tasks in human-agent interaction, for instance, dialogue act annotation [Li et al., 2011], head gesture recognition [Ferstl et al., 2019; Ishi et al., 2018], emotion recognition, turn management, addressee detection, gaze generation [Trafton et al., 2008], etc.

Machine learning techniques allow intelligent agents to learn interaction behaviour from human-human interactions. To do so, corpus collected from human-human and human-agent interactions can be exploited to learn how humans interact. The interaction characteristics learned from human-human interactions can then be integrated into intelligent agents. In this regard, machine learning in human-agent interaction intends to answer three questions [Kaiser et al., 1997]:
Q1: What and how the intelligent agents learn from humans?
Q2: What and how intelligent agents learn from each other?
Q3: What and how intelligent agents can learn without external guidance?

This thesis focuses on the first question where we propose machine learning and heuristic models that learn from human-human and human-agent interactions to improve baselines for different tasks in human-agent interaction.

1.1.3 Motivation

Owing to the complexity of multiparty interaction, existing research works that propose solutions to various tasks in multiparty interaction, such as addressee detection, turn change and next speaker prediction, dialogue act generation and behaviour generation, show room for improvement. Furthermore, existing approaches make a very limited use of multimodal features and the context in which the interaction takes place. The basic motivation behind this research work is that, with the selection of relevant multimodal features and efficient machine learning algorithms, the improved solutions for various human-agent interaction tasks can be proposed.

1.2 Contributions

In human-agent interaction, an agent perceives the interaction environment, makes decisions based these on perceptions and may generates behaviours based on these decision. In this research we propose to tackle a perception problem, a decision making problem and a behaviour generation problem. We propose models for three related tasks in multiparty human-agent interaction: addressee detection, turn change and next speaker prediction, and visual focus of attention generation. Our proposed models are based on machine learning techniques that learn from multimodal nature of human interactions in multiparty interaction. In some cases, heuristic approaches are used to further process the results of the machine learning models.

1.2.1 Addressee Detection

The main contributions concerning the addressee detection model are:
C1: Selecting the most relevant features for addressee detection in multimodal, multiparty settings.
C2: Proposing models that outperform baselines for next speaker addressee detection.
C3: Comparing the performance of addressee detection models via different focus encoding techniques.

1.2.2 Turn Change and Next Speaker Prediction

The main contributions to the turn change and next speaker prediction problem are:
C4: Selecting the most relevant features for turn change and next speaker prediction models.
C5: Proposing models that outperform baselines for turn change prediction problem.
C6: Proposing models that outperform baselines for next speaker prediction problem.

1.2.3 Visual Focus of Attention Generation

The main contributions of visual focus of attention (VFOA) generation model are:
C7: Selecting the most relevant features for speaker and listener VFOA generation.
C8: Proposing speaker and VFOA behaviour generation models that outperform baselines.

1.3 Organization of the Manuscript

The dissertation is divided into 8 chapters:

Chapter 2: Overview of Multimodal Human-Agent Interaction. Chapter 2 presents an overview of multimodal human-agent interaction including brief history, the main approaches, applications and common tasks in multimodal human-agent interaction. The chapter also presents some commonly used multimodal human-agent interaction datasets and features.

Chapter 3: System Work Flow, Methodology and Evaluation Criteria. The third chapter concerns the overall system work flow along with the methodologies adopted for the development of the models proposed in this research work. The evaluation criteria for the models is discussed.

Chapter 4: Addressee Detection. Chapter 4 describes the addressee detection model where the first two contributions of this research work are discussed. The related work, methodology and feature selection, along with the experiments and obtained results for addressee detection model are presented in this chapter.

Chapter 5: Turn Change & Next Speaker Prediction. The turn change and next speaker prediction model (contributions 3-5), is explained in Chapter 5. The chapter contains related work, methodology, feature selection, experiments and results for turn change and next speaker prediction models.

Chapter 6: Visual Focus of Attention (VFOA) Generation. Visual focus of attention (VFOA) generation system concerns the 8-10 contributions of this research work. The research work, selected features, methodology and experiments and results for VFOA generation system are explained in the Chapter 6.

Chapter 7: Evaluation of the Visual Focus of Attention Generation Model. The VFOA generation system is implemented in the form of multiple virtual agents interacting with each other. The interaction videos are presented to the end users for evaluation. The protocol, organization and the process adopted for the analysis of the results is explained in Chapter 7.

Chapter 8: Conclusion. Chapter 8 contains the summary of the research work and provides perspective on various limitations of the research work.

1.4 Publications

The proposed research work resulted in the following publications:

1.4.1 Journal

- U. Malik, M. Barange, J. Saunier, and A. Pauchet, *"Addressee Detection in Multiparty Interactions via Machine Learning Algorithms using Different Focus Encoding Schemes,"* Journal on Multimodal User Interfaces, Special Issue on Intelligent Virtual Agents. 12 pages, Submitted on 29th October, 2019, **(under review)**.

1.4.2 International Conferences

- U. Malik, M. Bouabdelli, J. Saunier, K. Funakoshi and A. Pauchet, *"A Hybrid Model of Visual Focus of Attention for Artificial Companions in Multiparty Interaction,"* International Conference on Intelligent Virtual Agents (IVA, 2020), Glasgow, UK, 8 pages, Submitted on 9th August, 2020 (**under review**).

- U. Malik, J. Saunier, K. Funakoshi and A. Pauchet, *"Who Speaks Next? Turn Change and Next Speaker Prediction in Multimodal Multiparty Interaction,"* IEEE 32th International Conference on Tools with Artificial Intelligence (ICTAI, 2020), Baltimore, USA, 8 pages, Submitted on: July 8th, 2020 (**Accepted as a short paper**).

- U. Malik, M. Barange, J. Saunier and A. Pauchet, *"A Generic Machine Learning based Approach for Addressee Detection in Multiparty Interaction,"* International Conference on Intelligent Virtual Agents (IVA), Paris, pp. 119-126, 2019.

- U. Malik, M. Barange, J. Saunier and A. Pauchet, *"Using Multimodal Information to Enhance Addressing Detection in Multiparty Interaction,"* International Conference on Agents and Artificial Intelligence (ICAART), Prague, Czech Republic, pp. 267-274, 2019.

- U. Malik, M. Barange, J. Saunier and A. Pauchet, *"Performance Comparison of Machine Learning Models Trained on Manual vs ASR Transcriptions for Dialogue Act Annotation,"* 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), Volos, Greece, pp. 1013-1017, 2018.

1.4.3 Posters

- U. Malik, M. Barange, J. Saunier and A. Pauchet, *Towards Generic Multimodal Interaction Systems based on machine learning and context awareness,"* Workshop Affects, Artificial Companions and Interaction (WACAI), France, 9p. (poster), 2018.

CHAPTER 2

OVERVIEW OF MULTIMODAL HUMAN-AGENT INTERACTION

This chapter discusses brief history of human-agent interaction, and then reviews the main approaches and applications along with common tasks in multimodal human-agent interaction. The chapter also details some commonly exploited multimodal human-agent interaction datasets and features.

The literature review related to the specific problems solved in this research work i.e. addressee detection, turn change and next speaker prediction, and VFOA behaviour generation, are further detailed in Chapters 4,5 and 6.

This chapter is divided into 8 sections. Section 2.1 presents a brief history of human-agent interaction. Chapter 2 explains the difference between types of human-agent interactions. Sections 2.4 and 2.5 respectively discusses application areas and tasks in multimodal human-agent interaction, while the major challenges of multimodal human-agent interaction are presented in section 2.5. The sections 2.6 and 2.7 respectively presents common datasets and features used for multimodal human-agent interaction. Finally section 2.8 contains discussion of the chapter.

2.1 History of Human-Agent Interaction

History of human-agent interaction dates back to ancient times. The idea of *golem*, who was an artificial being endowed with life has been around for centuries and is discussed by Wiener in his book [Wiener, 1964]. Ancient Chinese legends also refer to robot-like

creations such as Yanshi which looked so real that it had to be dismantled in order to be proved that it was an artificial creation [Goodrich and Schultz, 2008].

In modern times, Nicolas Tesla was the first to develop an artificial companion in the form of a radio-controlled boat which he referred to as a *a borrowed mind* described in his patent "Methods and Apparatus for Controlling Mechanism of Moving Vessels" [Nikola, 1898]. Tesla states in the patent that *"you see there the first of a race of robots, mechanical men which will do the laborious work of the human race."*

The field of human-agent interaction formally emerged in the mid 1990s and early 2000s when researchers from various scientific domains such as natural language, psychology, cognitive sciences, robotics and human computer interaction start to organize events where human-agent interaction emerged as a specialized domain. In this regard, the IEEE International Symposium on Robot and Human Interactive Communication (RoMan) is one of the earliest scientific meeting that first occurred in 1992 and continues annually till date. Since early 1990 until recently, many international conferences and symposiums are annually conducted to further research in the domain of human-agent interaction including, International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), IEEE Conference on Robotics and Automation (ICRA), International Conference on Intelligent Virtual Agents (IVA), Human Factors and Ergonomics Society, Advancement of Artificial Intelligence (AAAI) Symposium, International Joint Conferences on Artificial Intelligence Organization (IJCAI).

Recently, human-agent interaction has become ubiquitous and industries have invested huge resources in developing advanced human-agent systems for example Tutoring Assistants, Child Care Assistants, Robot Soldiers, and Patients Care taker robots, etc. Human can interact with an intelligent-agent in various ways as discussed in the next section,

2.2 Unimodal vs Multimodal Human-Agent Interaction

Human-agent interaction can be broadly divided into two types: unimodal interaction & multimodal interaction.

2.2.1 Unimodal Human-Agent Interaction

As defined in Chapter 1, *a modality is the classification of a single independent channel of sensory input/output between a computer and a human* [Karray et al., 2008]. A human-

agent interaction system that allows users to interact with agents via a single modality such as with voice, hand gestures or any other single channel of communication is called unimodal human-agent interaction system [Newell and Card, 1985], [Karray et al., 2008] [Adkar, 2013]. Examples of unimodal human-agent interaction systems include graphical user interfaces, touch and pointed-based interfaces, text-based interfaces and so on.

Though unimodal human-agent interaction systems are relatively e easier to build compared to human-agent interaction involving multiple participants, they have their own limitations. Some of these limitations are listed by Adkar [Adkar, 2013] as:
- unnatural way of human interaction,
- designed for average user,
- difficult to be used by disabled, illiterate and untrained people,
- unable to provide a universal interface resistant to errors.

Owing to these limitations, researchers offer solutions that combine two or more modalities to develop more robust and less error prone systems. Such systems are referred as multimodal human-agent interactions systems [Karray et al., 2008], [Adkar, 2013], [Gupta, 2012].

2.2.2 Multimodal Human-Agent Interaction

Multimodal human-agent interaction systems provide multiple channels (modalities) of communication between human and intelligent agents. Various communication modalities in multimodal human-agent interaction on system can be redundant or may complement each other depending the system requirements. If one of the modality is not available or cannot be comprehended, communication between human and agents may continue via other modalities. Some of the advantages of multimodal human-agent interaction as proposed by Adker [Adkar, 2013] are as follows:
- accommodate wide variety of users e.g. disable, illiterate and untrained users,
- more natural way of human interaction,
- less error prone because one modality can compensate the errors in other modalities

In the remainder of the thesis, we only be deal multimodal human-agent interaction where multiple modalities are involved for human-agent interaction.

2.3 Application Areas of Multimodal Human-Agent Interaction

Multimodal human-agent interaction systems have a wide range of applications from personal assistants to virtual story teller and patient health care assistants to personalized tutoring and training system.This section reviews some of these applications areas.

2.3.1 Virtual Assistants

Virtual assistants help users schedule their tasks and answer common questions. Amazon Alexa and Apple Siri are two of the most common examples of virtual assistants [Kepuska and Bohouta, 2018]. Recent research has tried to incorporate multimodality in virtual assistants. To this end, [Johnston et al., 2014] propose a model for a virtual assistant that allow users to plan an outing via an interactive multimodal dialogue through a mobile device. Users can employ speech and gesture inputs to issue commands to the virtual assistant which in turn suggest best outing plans and ideas. Furthermore, [Makula et al., 2015] develop Smart Robotic Assistant (SRA) that help human beings in reducing the manual efforts and the risk factor in hazardous situations. The SRA is capable of performing different types of tasks such as picking up an object from one location and dropping it at another location. The body of the SRA is controlled via tilt-gesture of a smart-phone. The arms and claws of the SRA are controlled through human voice commands.

2.3.2 Training and Tutoring

Intelligent agents have many applications in training and tutoring various skills to human users. In this regard, [Bradbury et al., 2003] propose a virtual agent named *eyeCook* which helps novice cooks to cook different types of meal by issuing instructions based on the combination of eye-gaze and speech. [Park, 2018] develop a multimodal pedagogical agent in the form of a talking head that delivers multimedia contents to students through its appearance and voice narration for English teaching and learning.

2.3.3 Healthcare

Multimodal human-agent interaction systems have found another application in healthcare systems as well. To this end, [Jacob et al., 2012] develop robotic scrub nurse *Gestonurse* that interacts with human surgeon by passing surgical equipment as a response

to gesture and speech commands by the human surgeon. [Ronzhin and Karpov, 2005] propose a multimodal human-agent interaction systems that enable people with hand disability to interact with computers with the help of voice and head movements.

2.3.4 Autonomous Vehicles

Self-driving cars and autonomous vehicles such as UAVs (Unmanned Aerial Vehicles) make extensive use of multimodal human-agent interaction. [Manawadu et al., 2017] propose an automated vehicle control system where user perform basic tasks such as lane-changing, overtaking and parking by issuing instructions via touchscreen, hand-gesture and haptic sensors. [Gutierrez et al., 2016] develop a framework for the control of UAV (Unmanned Aerial Vehicles) using text, speech and keyboard/mouse input via a web based graphical interface. The interface is deployed on cloud which can be accessed from anywhere with an Internet powered device.

2.3.5 Games and Entertainment

Multimodal human-agent interaction has found its way in gaming and entertainment industry as well. More and more virtual reality applications are now allowing users to interact with virtual characters via multimodal inputs. [Gutierrez et al., 2016] design an augmented reality interface that allow users to control the verbal and non-verbal behaviours in telepresence robot applications. A single user can control both the verbal and non-verbal behaviour or two users can individually control verbal and non-verbal behaviours. [Link et al., 2016] develop mixed reality tabletop role-playing game where users can interact with virtual agents via speech and gesture commands.

2.3.6 Arts and Culture

Multimodal human-agent interaction systems are being used to play music, generate avatars and interact with people in museums. In this regard, [Wicaksono and Paradiso, 2017] propose *FabricKeyboard* which is a deformable keyboard interface based on a multimodal fabric sensate surface. User can interact and play music by pressing keys and via hand gestures like hovering and waving towards and against the keyboard. [Capurro et al., 2015] propose a multimodal interface for storytelling about different artifacts and monuments in museums. Users can interact via speech, touch and gestures with the interface to get various information about different museum artifacts.

2.3.7 Military and Police

Virtual Agents are increasingly being employed to complete *dull, dirty and dangerous tasks*. To this end, military and police make most use of such virtual agents. Intelligent agents in the form of robots are often employed for bomb disposal tasks. These robots are frequently used to reach and inspect suspicious packages and dispose them off. The robots are controlled remotely by human operators via multimodal inputs [Wells and Deguire, 2005], [Scholtz et al., 2006]. Robots are also being utilized to serve soldiers and officers during war and peace times where they receive instructions through multimodal channels such as speech and gestures [Schneider, 2007].

Multimodal human-agent interaction is a vast field and is applicable to any research area where humans and artificial companions interact to perform common tasks. In addition to broad application domains, there are several human-agent interaction tasks where multimodal nature of interaction can help improve interaction and achieve mutual goal.

2.4 Common Tasks in Multimodal Human-Agent Interaction

A fully functional multimodal human-agent interaction system, relies on multiple interdependent tasks e.g. natural language processing, dialogue act annotation, participant gaze annotation, addressee detection, next speaker identification, etc.

Owing to a large number of interdependent tasks, developing an end-to-end multimodal human-agent system is a huge task. Hence, researchers have tried to propose solutions for the development of sub-tasks involved in human-agent interaction. This section reviews related works that propose solutions to some of the common tasks for multimodal human-agent interaction.

2.4.1 Dialogue Act Annotation

A dialogue act (DA) is defined as the meaning of an utterance at the level of illocutionary force [Searle, 1969] or as the function of a user's utterance [Stolcke et al., 2000]. For instance, when a user says "*Could you please lend me this book*", she is actually performing a *request* function. Similarly, when a user says "*I will leave tomorrow*", she is committing herself to doing something, hence performing a *commit* act. DA annotation of speech utterances is one of the most important tasks in order to correctly interpret user utterance in human-agent interaction.

[Carpenter and Fujioka, 2011] achieve an accuracy of 90% for the DA annotation of IRC Chat messages [Werry, 1996] into 43 predefined DA categories. They employ string matching technique in combination with a complex set of rules. However, they state that the high accuracy is partly due to the highly constrained nature of the corpus. [Tran et al., 2017] propose a hierarchical recurrent neural network combined with attention mechanism to learn sequence of DAs on Switchboard [Godfrey et al., 1992] and MapTask corpora [Anderson et al., 1991]. The input to the proposed model is a sequence of DAs while the output is the corresponding sequence of DAs. [Ortega and Vu, 2017] explore recurrent neural network coupled with attention mechanism for DA annotation at different context levels. The proposed model achieves an accuracy of 84.3% on the MRDA dataset [Janin et al., 2003] and 73.8% on the Switchboard dataset.

2.4.2 Addressee Detection

Addressee detection refers to identifying the addressee of the speaker's utterance. In dyadic interactions, involving a human and an agent, addressee detection is fairly straight forward since there is only one listener for the current utterance. However, for multiparty interactions where more than two participants are involved and at least one of the participants is an agent, addressee detection is a fairly complex task. Several works exist that use multimodal information such as human gaze, DA, hand gesture, context, and content of the utterance to predict the addressee of the current utterance [Akker and Traum, 2009; Jovanovic, 2007].

[Traum et al., 2004] propose a rule-based approach exploiting current utterance, previous utterance, current and previous speakers, and previous addressee to the detect the addressee of the current utterance. They achieve an accuracy between 65% to 100% on Mission Rehearsal dataset [Traum et al., 2006], depending upon the DA type. However, the algorithm does not generalize well on other datasets. For example the accuracy of the proposed approach decreases to 36% on the AMI dataset.

[Akker and Traum, 2009] improve the work by [Traum et al., 2004] by incorporating gaze as additional feature resulting in improved overall accuracy of 65% on the AMI dataset. They also test the gaze as the only feature for addressee detection. The only rule they use for this work is that the speaker of an utterance looks at a participant for more than 80% of the time duration of an utterance, the corresponding participant is chosen as the addressee of the utterance. Otherwise, the utterance is marked as being addressed to the group. This approach results in the overall accuracy of 57%.

2.4.3 Turn Change & Next Speaker Prediction

Identification of the participant that should speak next is a fundamental task in multi-party interaction. The identification of next speaker not only enables the agent to decide when it has to speak but also allows the agent to exhibit certain behaviour such as gaze movement towards the next potential speaker. Multimodal signals such as speech, human gaze, respiratory behaviour, lip movements, and posture shifts are identified as common markers for turn change and next speaker prediction [Ishii et al., 2015a; Petukhova and Bunt, 2009]

[Petukhova and Bunt, 2009] studied the importance of various multimodal signals such as gaze directions, verbal signals, lip movements and posture shifts for the identification of next speaker. The approach simply finds the correlation between the various multimodal features and the speaker change.

[Ishii et al., 2015a] uses human gaze and respiratory behaviour for turn change and prediction of next speaker. They propose individual models based on gaze and respiratory behaviour and a combined model which fuses gaze and respiration for the identification of the next speaker. The result show that the model based on the fusion of respiration and gaze behaviour result in a better performance (0.52 F1) compared to the models based individually on gaze (0.456 F1) or respiration (0.47 F1).

[Ishii et al., 2015b] propose a two-step model that initially predicts whether or not the turn change will occur and in the next step predicts who will be the next speaker based on head movements such as the amplitude and frequency of the movement of head position of the speaker as well as for the listeners. They achieve an accuracy of 75% for turn change prediction and an accuracy of 55.2% for next speaker prediction.

2.4.4 Emotional Attitude Recognition

Multimodal emotional attitude recognition refers to the automatic detection of human emotions using multiple modalities such as speech, the content of the utterance, hand and head gesture, gaze movements, etc [Barros et al., 2015; Ranganathan et al., 2016]. Emotional attitude recognition is one of the most important tasks in multimodal human-agent interaction. By determining emotion attributes, robots can identify important variables of human behavior and use these to communicate in a more human-like fashion and thereby extend the interaction possibilities.

[Barros et al., 2015] propose spontaneous emotional attitude recognition model based on hierarchical feature representation. The proposed model learns to integrate multiple modalities for non-verbal emotion recognition using hierarchical convolutional neural

network. The results show an accuracy of 91.30% for multimodal emotion recognition.

[Ranganathan et al., 2016] use deep belief neural networks for the generation of features that can be efficiently used for emotion recognition in an unsupervised manner. The results show that DBN achieves an accuracy of 96.3% and outperforms the state-of-the-art emotional attitude recognition architecture on emoF-BVP multimodal dataset. They further propose convolutional deep belief neural network (CDBN) that further improves the results for the multimodal emotional attitude recognition task. They achieve an accuracy of 97.3% for multimodal emotional attitude recognition via CDBN.

[Choi et al., 2018] propose a convolutional attention neural network based model for emotional attitude recognition that learns joint hidden relationships between speech and text. The result shows that compared to shallow representation that involves simple concatenation of text and speech feature vectors, the proposed joint attention model yields better result for emotion classification using text and speech data. They report an overall accuracy of 88.89% for all emotions in the CMU-MOSEI dataset [Zadeh et al., 2018].

2.4.5 Engagement Detection

Engagement detection techniques are used to identify the level of engagement or interest of interaction participants towards the proceedings of a meeting. Engagement detection is crucial in human-agent interaction and can help a robot to perform actions that can grab user attention or increase engagement. Multimodal information can be used to detect how engaged a person is in a meeting.

[Kim et al., 2016] propose a multimodal engagement detection model trained via a custom database of audiovisual interaction of children. The model is trained via Ranking SVM algorithms using turn taking and body movements as features. The results show that Ranking SVM algorithm significantly outperforms the baseline SVM algorithm at ($p < .0001$).

[Fedotov et al., 2018] exploit logistic regression to predict engagement and disengagement using highly imbalanced dataset. The feature set consists of audio signals, eyes, lips, body and facial expressions. They report an unweighted average recall value of 0.715.

2.4.6 Hand Gesture Generation

In addition to detection tasks such as emotional attitude recognition and engagement detection, multimodal human-agent interaction involves exhibiting multimodal behaviour.

Hand gestures are commonly coupled with speech to convey various signal such as emotional state of participants and addressee detection [Barros et al., 2015; Petukhova and Bunt, 2009]. Hand gestures commonly occur in daily dialogue interactions, and have important functions in communication. Creating a human-like embodied conversational agent requires an agent to express verbal as well as non verbal behaviour. Hand gestures are one of the most common ways of non-verbal communications.

To this end, [Chiu and Marsella, 2014] propose a machine learning based hand gesture generation model that uses speech content contents and the timing of the speech as input features. The proposed model consists of two sub-modules. The first module learns the mapping between speech and gestural annotations. The second models learns the mapping between gestural annotations to gestural motion. The combined model learns to synthesize natural gesture animations from audio speech. They report an accuracy of 74%.

[Ishi et al., 2018] analyse human-human interactions to study the correlation between hand gestures and DA. They also conduct cluster analysis on speech content and gesture emotion to study relationship between speech content and hand gesture. Based on the analysis they propose a speech-driven gesture generation method based on DA, utterance text and prosody. The model is implemented in an android robot. T-tests from real-time experiments reveal that hand gesture generated by robot during interaction were judged to be relatively natural.

2.4.7 Head Gesture Generation

Head gesture generation is another area that has attracted large research interest from multimodal human-agent interaction community. Head gestures are also commonly used in interaction to express non-verbal behaviour which can be used for dialogue act annotation, addressee detection, engagement and emotion detection, etc. To this end, [Sargin et al., 2008] analysis head gesture and prosody patterns from human interactions to propose a model for data driven head gesture animation. [Jia et al., 2014] propose head movements and facial gestures generation model for a talking avatar using the three dimensional pleasure-displeasure, arousal-nonarousal and dominance-submissiveness (PAD) descriptors of semantic expressivity. In addition, [Aly and Tapus, 2012] use coupled hidden markov models for mapping speech to a sequence of head gestures. They report a similarity score of 62% between the original sequence of gestures and predicted gestures.

2.4.8 Visual Focus of Attention Behaviour Generation

Visual Focus of Attention (VFOA) is one of the most important non-verbal cues for seamless interaction. In our research we identified that VFOA is a marker can be exploited as a feature to predict next addressee. Owing to the importance of VFOA behaviour generation for human-agent interaction, several researchers have tried to propose gaze generation approaches in dyadic as well as multiparty settings [Andrist et al., 2014; Liu et al., 2012].

[Peters et al., 2005] develop a VFOA behaviour generation model that monitors participant gaze to assess their level of engagement and then generates gaze movements to maintain and enhance participant engagement in the conversation. The model consists of a set of rules for both speaker and listener VFOA behaviour generation in multiparty settings.

Liu *et al.* [Liu et al., 2012] develop a rule based approach for nodding, head tilting and VFOA behaviour generation in multiparty interaction. Experiments are performed with humans interacting with two robots who play the role of receptionists at an information desk. The users rate the naturalness of the robots.

2.5 Major Challenges in Multimodal Human-Agent Interaction

While multimodal human-agent interaction fosters naturalness, flexibility and error avoidance, designing a multimodal human-agent interaction system is a challenging endeavour.

According to [Baltrušaitis et al., 2017], major challenges in multimodal interaction can be divided into 5 categories i.e. input data representation, multimodal translation, alignment of multimodal inputs, multimodal input fusion, and co-learning. [Baltrušaitis et al., 2017] propose these challenges with reference to machine learning based solutions for multimodal interaction, however some of these challenges are generic enough and can be handled by heuristic solutions as well. In addition, several research works have shown that capturing the context of interaction is another challenge in multimodal human-agent interaction. Therefore a total of 6 categories of challenges are discussed in this section.

2.5.1 Input Data Representation

Input data representation refers to the representation of raw data in a format that statistical or rule based models can work with. Data representation is one of the most important task in multimodal human-agent interaction for both rule-based and heuristic approaches. Rule-based approaches may avoid data representation step since rule-based approaches can directly work with the raw data and rules can be based on raw data values. On the other hand, in case of machine learning based system, data representation, which is also known as feature is one of the most important multimodal human-agent interaction task [Hinton et al., 2012; Krizhevsky et al., 2012].

Several researchers have tried to propose models for input data representation. To this end, [Bengio et al., 2013] identified that natural clustering, spatial and temporal coherence, scarcity and smoothness are some of the important properties for efficient feature representation. [Mroueh et al., 2015] use a neural network based approach for feature representation. In their work, each modality is represented by a dedicated input layer neurons. The hidden layers project the input modalities into a joint representation which can be either passed to further hidden layers or used to make predictions. With this approach, both feature representation and predictions can be learned at the same time. A downside to this approach is that it requires large amount of supervised training data. [Chen and Jin, 2015] employ recurrent neural network for emotion recognition from multimodal input. Their research work focuses on long short-term neural network (LSTM) [Hochreiter and Schmidhuber, 1997] for modeling multimodal temporal information. However, their research work was limited to the detection of emotion from multiple input modalities.

The proposed research work employs data from multiple modalities. i.e. speech, context, gaze, etc. which have to be represented in a form that the statistical or rule based approaches can work with. Therefore, data representation is an important step in the context of this thesis.

2.5.2 Multimodal Translation

Multimodal translation is concerned with mapping or converting data from one modality to another. For instance, given an image, the task can be to convert the image to text or vice versa. Multimodal translation is one of major challenges in multimodal human-agent interaction, particularly if data contains missing values for one or more modalities.

There are several reasons for having missing data in one of the modalities in the dataset. For instance, for a particular part of interaction, data is not recorded due to

technical reasons, or sometimes users are not willing to let the system use some of the components of video, speech or texts. In such cases, either one can remove the parts where the data for all the modalities is missing or you can fill in missing values via different techniques. Multimodal translation is one such technique.

With multimodal translation, missing data from one modality can be filled based on the information from other modalities. For instance, if the answer to a question is not available in speech, but the head gesture, which signals positive response, is available, the text can be filled with a *yes.*. Rule-based, as well as statistical approaches have been developed for multimodal translation. On the other hand, rule-based approaches use a dictionary that contains mapping from one modality to another. On the other hand machine learning approaches learn mapping using data corpus where the training set contains labelled translations from one modality to the other.

To this end, [Gupta et al., 2012] used supervised ML nearest neighbor algorithm to predict the description of an image. [Mansimov et al., 2015] propose a reversed approach and predicted text description using an image input, based on recurrent neural network (RNN) [Wang et al., 2019]. The application draws images on the canvas based on the words in the text.

2.5.3 Alignment of Multimodal Inputs

Once mapping between modalities is achieved, the next problem is to align individual segments of the multimodal inputs. Multimodal input alignment refers to synchronizing sub-components of the instances of information received from multimodal input [Baltrušaitis et al., 2017]. For instance, for a given video, alignment maps areas within a video frame that correspond to a particular word or group of words in the sentence. A simplest example of alignment is a video greeting. When a user says *"Hello, how are you?"* the alignment task involves mapping parts of the video frame that correspond to the words *"hello"* and *"how"*, etc.

In rule-based approaches, multimodal alignment is governed by a set of heuristics. For instance, one of the heuristics can be simply time alignment of multimodal inputs where multimodal inputs having common time frame are aligned together. Machine learning approaches on the other hand learn the alignment process using statistical algorithms trained on multimodal datasets.

Several research works exist that propose model for multimodal input alignment. In this regard, [Malmaud et al., 2015] propose a method to align sequences of instructions with the task performed. The method is implemented in cooking domain where a video recipe is aligned with the automatically generated text instruction. The work involves

two stages. In the first stage, hidden markov models [Ghahramani, 2001] are used to align the recipe steps with text instructions. In the next step, a visual food detector is trained using neural networks that further refines the alignment process.

[Zhu et al., 2015] developed a method to align visual contents from a movie to the text in their book counterpart. The proposed method takes text from the books and the video content of the their movie releases as input features and uses convolutional neural network [Ji et al., 2013] to predict the semantic textual information by aligning movie scenes to the information in the book.

In this research work, multiple modalities are manually aligned at the level of utterance. For instance, the VFOA behaviour during an utterance and text of the utterance are aligned together.

2.5.4 Multimodal Input Fusion

In the proposed research work, different types of features have to be fused together in order to train machine learning models. Input fusion refers to integrating inputs received from multiple sources (speech, video, text) to predict an output class such as user action. An input fusion technique is required to fuse or merge the features together. While rule-based approaches do not necessarily require to fuse multimodal inputs since rules can be based on individual modalities, the research area concerning multimodal input benefits greatly from machine learning approaches [Yuhas et al., 1989].

Multimodal input fusion is broadly categorized into three types: early, late and hybrid fusion [Zeng et al., 2009].

Early fusion refers to the process of integrating features extracted from different modalities into a single feature vector before training the model. One associated benefit of early fusion is that only one model has to be trained. [De Mulder et al., 2015]. In case of late fusion, individual models are trained for features extracted from different modalities and the output is fused using criteria such as averaging [Shutova et al., 2016], signal variance [Evangelopoulos et al., 2013], voting scheme [Morvant et al., 2014] or some learned model [Glodek et al., 2011]. Hybrid fusion involves a combination of both early and late fusion.

In the proposed research work, early fusion techniques are exploited for fusing multimodal inputs in order to study the mutual impact of all the multimodal features during training and testing the machine learning models.

2.5.5 Co-Learning

Co-learning is another important multimodal human-agent interaction challenge as per [Baltrušaitis et al., 2017] who defines co-learning as *"aiding the modeling of a resource poor modality by exploiting knowledge from other modalities"*. Co-learning like multimodal translation can help fill missing values in one of the modalities. Rule-based approaches are not concerned with co-learning. In rule-based approaches missing values can be replaced on the basis of rules such as by mean, median or mode in case of numerical data and by most frequently occurring category. Machine learning based approaches have to make up for missing or poorly represented data. [Moon et al., 2014] show how to transfer information from a speech recognition neural network (based on audio) to a lip-reading one (based on images), leading to a better visual representation, and a model that can be used for lip-reading without need for audio information during test time. Similarly, [Frome et al., 2013] improved image classification by using text to improve visual image representation using word2vec approach [Mikolov et al., 2013].

2.5.6 Context in Interaction

Human-agent interaction occurs within a particular context. For instance, the current focus of the speaker depends upon several factors i.e. who was the previous speaker, what was the dialogue act of the previous utterance. If the previous utterance was a question to the current speaker, the focus of the current speaker will be most probably towards the last speaker. All in all, any information that is not directly part of any modality i.e. speech, audio, text, facial expression, etc, but has a certain impact on interaction, can be defined as context information. Context information must be taken into account when predicting the next move in multimodal human-agent interaction.

Different researchers have tried to define the term context. [Dey et al., 2001] has discussed the strength and weaknesses of these definitions. Dey claims that context is something that cannot be defined as a specific set of entities. Rather context is all about the overall situation related to user and application. In different settings or tasks, one aspect of the context is more important than the others. For instance, physical environment can be a more important aspect of context in informal discussions, while topic of discussion can be more important in formal discussions. [Dey et al., 1999] define context as: *"any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."*

According to Dey, four major categories of context can be identified: identity, location,

status and time. Identity refers to any information specific to persons, places or things involved in the interaction. For instance, the name, age and qualifications of a person; the quality of the hardware involved in interaction, etc. Location refers to the geographical position of a person, the coordinates of a person in a room, etc. Status refers to the mood of a person. For a place, status can be noise level, ambient light or temperature. Time is another dimension of context. It can be used to retrieve historical information regarding interaction, person, place and objects involved in the interaction. A survey of different works in context categorization has been presented by [Perera et al., 2014].

In the proposed research all the models, i.e. addressee detection, turn change and turn change and next speaker prediction and gaze generation, use contextual features for making predictions. The proposed research work relies on Dey's definition of context for selecting context features as it is flexible enough to accommodate all the feature entities related to interaction such as names, places, persons and objects.

2.6 Datasets

The importance of data is highlighted by the fact that not only statistical algorithms such as machine learning models learn from data, but rule-based approaches also analyse data for the identification of relevant rules. The availability of dataset with all the relevant annotations is vital to the successful development of machine learning as well as rule based models. A multimodal human-agent interaction dataset is a type of dataset where humans and agent interact using at least two modalities.

There are two options to obtain datasets for experimentation: the dataset can either be developed from scratch through experiments. Developing datasets from scratch can help to obtain customized annotations in desired scenarios. However, a major downside to developing customized datasets is the time and effort required to create the dataset. The other option is to explore existing datasets that satisfy the proposed requirements.

In the domain of multimodal human-agent interactions, several datasets have been proposed depending upon the goal of the research. A typical multimodal human-agent dataset, in case of dyadic interaction consists of two humans or a single human user and an agent. While in multiparty interaction, datasets consist of more than two participants.

The main goal of multimodal human-agent interaction models is to ensure that the interaction between humans and agents is as similar as possible to human-human interaction. Therefore, in order to model human-agent interaction, two types of datasets can be used: (i) datasets of human-human interactions, and (ii) datasets of human-agent interactions.

In addition to the type and scenario of the interactions, the datasets differ in terms of annotations, number of interactions, duration of each interaction, number of participant per interaction, task i.e. whether they are open ended or task-oriented. Not every dataset can be used to solve every problem involving multimodal human-agent interaction.

2.6.1 Dyadic Datasets

Dyadic datasets involve interaction scenario between two participants.

2.6.1.1 RECOLA Dataset

[Ringeval et al., 2013] has developed Remote Collaborative and Affective Interactions (RECOLA) database which aimed at measuring emotions (arousal and valence) and social signals in a dyadic interaction setting. The Interaction participants are asked to perform widely known Mission Survival task which requires significant mutual collaboration. The dataset is annotated in two sets: self-assessment and other-assessment. In self-assessment, the participants are asked to fillin a questionnaire regarding to their emotional state during different task phases. For other-assessment, six external observers annotate different scenes of the recorded data along arousal and valence axis of emotion. They have also collected annotations with respect to five social signals, namely: agreement, engagement, dominance, performance and rapport based on a Likert scale rating [Harpe, 2015].

2.6.1.2 JOKER Dataset

JOKER dataset is collected via a robotic platform developed by [Devillers et al., 2015] for the analysis of laughter elicitation. Interactions involve a human and a humorous robot. The job of the robot is to interact with the human via multiple modalities in order to make him laugh. The goal of the autonomous platform was to recognise the emotions of the participants based on audio cues (para-linguistic features), in order to endow the robot with a comprehension of the user's receptiveness to humour before producing an action.

2.6.2 Multiparty Datasets

Multiparty datasets concerns interaction datasets containing more than two participants.

Figure 2.1 – Three Meeting Rooms for the AMI Dataset. Image taken from [Carletta et al., 2005]

2.6.2.1 The AMI Dataset

The AMI dataset is a multiparty dataset consisting of 100 hours of human-human interactions. The interactions are divided into two types: (i) natural real-time interactions, and (ii) scenario driven meetings where participants play roles of employees in an electronic company. The task is to propose a novel design for a TV remote control.

In the scenario driven meetings, there are four types of participants: The Project Manager (PM), who moderates the meetings and ensures that the project is completed within the time and cost constraints. A User Interface Designer (UI) is responsible for technical specifications of the remote control. The Marketing Expert (ME) is responsible for evaluating market trends and determining user requirements. Finally, the industrial designer is responsible for designing how the remote control works along with the production scenarios. The data is recorded in three different types of rooms as shown in Figure 2.1.

Various annotations are available for the dataset including speech transcriptions, word timings, speaker turn boundaries, named entities and references to people, time, numbers and artifacts. The annotations also include the DA, topic segments and extractive summaries. Among non-verbal behaviours, the annotations are available for head and hand gestures, focus of attention, etc. The AMI dataset contains manual as well as ASR transcriptions. A comparison of some of the utterances transcribed manually and via ASR is given in Table 2.1

The AMI dataset follows a custom DA annotation scheme. Utterances in the AMI corpus have been categorized into 15 DAs as shown in Table 2.2.

Though the AMI dataset contains a high number of annotations, the number of interactions involving all the annotations is limited. For instance, accross all interactions, 117,000 utterances have been annotated with DA. Out of 117,000 utterances, only

Manual Transcriptions	**ASR Transcriptions**
Um , monkeys have attitude	Um What's had a cheap
but uh cats are one of my favourite animals	but uh Cats are one of my favourite animals
they are Very independent	there Very independent
they're snotty as hell at the best of times	sloppy as hell are the best of times

Table 2.1 – Some manually and ASR transcribed utterances from AMI database.

Dialogue Act	**Symbol**	**Example (In capital)**
Backchannel	bck	YEAH, YUP, OKAY, HMMM
Stall	stl	SO | then we have sample sensor.
Fragment	fra	I THINK THE | The remote shouldn't be too heavy.
Inform	inf	DEADLINE FOR THIS PROJECT IS AT THE END OF TODAY
Elicit-Inform	el.inf	ARE THERE ANY CONSTRAINTS FROM THE MARKETING SIDE
Suggest	sug	MAYBE JOHN CAN FILL IN THE NOTES WHILE MARK IS GOING ON
Offer	off	THEN I'M GOING TO TALK ABOUT UH THE PROJECT MANAGEMENT
Elicit-Suggestion	el.sug	OR HAVE WE MISSED ANYTHING
Assess	ass	so if you find out from the technology background, | THAT WOULD BE GOOD.
Elicit-Assessment	el.ass	DO WE ACTUALLY WANT TO INCORPORATE ALL OF THEM | or have wee missed anything?
Elicit Comment About Understanding	el.und	B is describing an idea about the remote control functions B | like three mental states, yeah, you know what I mean, we can just make it uh [other participants acknowledging] B controlled by a brain, | HUH?
Be-Positive	be.pos	THANKS
Be-Negative	be.neg	I DO NOT LIKE THE IDEA
Comment About Understanding	und	Alright, okay.
Other	oth	All the other type of utterances.

Table 2.2 – Summary of DAs in AMI corpus (examples taken from the official documentation).

9,071 utterances have been annotated with VFOA information for meeting participants, whereas addressee information is only available for 8,874 utterances. Only 5,628 utterances contain annotations for VFOA, DA and addressee, simultaneously. The complete annotation details for all the meetings is available at the official documentation link [Carletta, 2005].

2.6.2.2 The MPR Dataset

The MPR dataset contains 30 trios of Japanese individuals that are friends or family and range in age from their 20s to 60s. The genders of the participants are balanced.

Each trio participated in two 25-minute interaction sessions with a robot in which they repeatedly engaged in a conversational game. The robot spoke English only, as it was explained that the robot was designed for English learning purposes. The participants were allowed to speak either English or Japanese. The robot was controlled by a human operator in a wizard of Oz setting. A director was present in the scenario to give instructions to the participants. Each participant could enter or leave the interaction field depending upon these instructions. A view of the interaction in MPR 2012 dataset is shown in Figure 2.2.

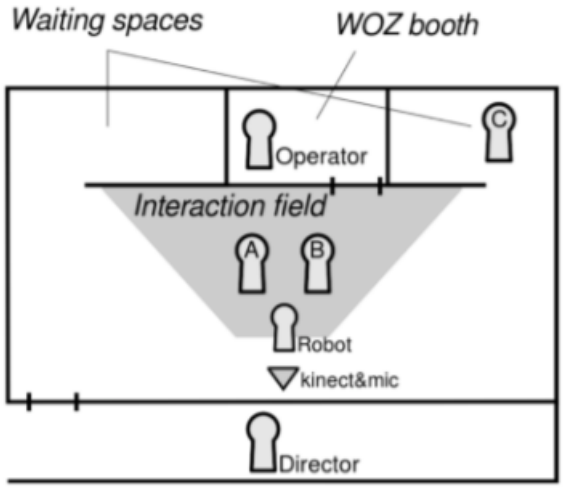

Figure 2.2 – A view of interaction in MPR 2012. Figure from [Funakoshi, 2018]

The data consists of two sessions. In the first session, each trio was engaged in a 20 Questions game. In each question game, the robot secretly chose a target. The trio had to guess in 20 questions what the robot had chosen. In the second session, participants played a gesture mimicking game. First, each participant was taught gestures corresponding to various English words. When more than one participant were ready in the field, the robot started a mimicking speed competition in which it said a word and participants had to make the corresponding gesture as fast as they could. There are 31 meetings for each session.

The MPR datasets contain annotations for participant status which can have three possible labels: participating (maintaining interaction with the robot), passing (leaving or entering the interaction field, or passing behind the participants), and observing (inside

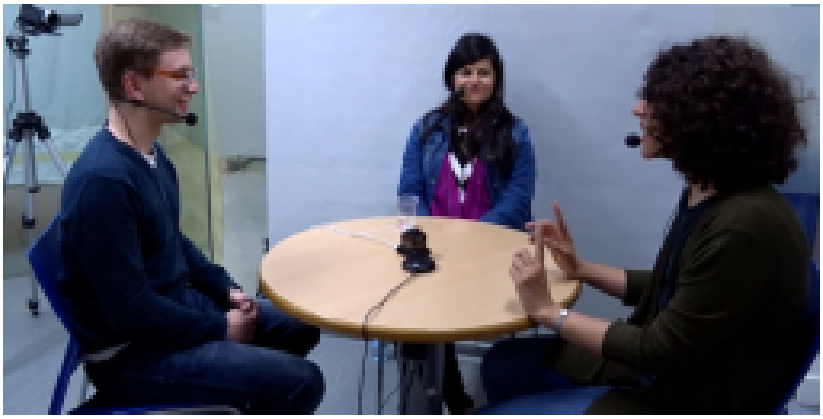

Figure 2.3 – Meeting view from the MULTISIMO dataset ([Koutsombogera and Vogel, 2018])

the interaction field but not interacting with the robot). Furthermore, the annotations are available concerning human participants VFOA information, addressee of the utterance, utterance segment, transcript of the utterance (mix of japanese and english) and dialog acts. The DA have been annotated based on DIT ++ Taxonomy [Bunt, 2009].

2.6.2.3 The MULTISIMO Dataset

The MULTISIMO dataset [Koutsombogera and Vogel, 2018] is a multiparty multimodal corpus that contains 23 meetings of 3 participants. The average duration of the meetings is 10mn ($min = 6$, $max = 16$), for a total time of 4 hours.

The average age of the participants is 30 years where the youngest participant is 19 years old while the oldest participant is 44 years old. There are 25 female and 24 male participants. 16 are native while 33 are non-native. Speaker familiarity between the participants is totally random.

The corpus is task oriented where one of the 3 participants plays the role of a facilitator while the other 2 participants have the role of players (see Figure 2.3). The facilitator asks a question for which there are three best answers. The participants have to find the answers and then rank them based on their popularity.

Different meetings in the corpus have been annotated for speech, acoustic, visual, lexical, perceptual and demographic information. However, only two meetings (S02 and S18) in the MULTISIMO corpus are annotated with gaze information.

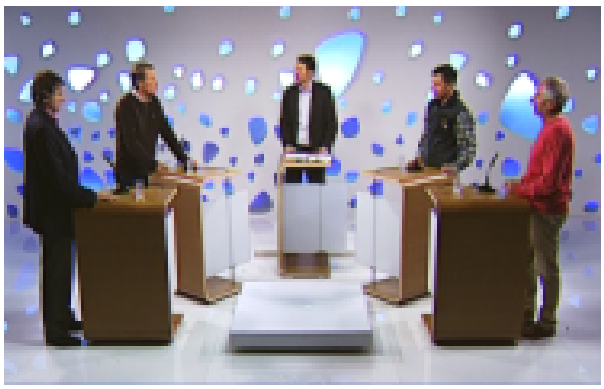

Figure 2.4 – Meeting view from the Canal9 Dataset ([Vinciarelli et al., 2009])

2.6.2.4 Freetalk Corpus

Freetalk corpus is a multimodal multiparty video and audio corpus recorded over three days during three 90 minute sessions [Advanced Telecom Research Labs, 2005]. The meetings involve 4 to 5 participants. The participants are originally from Australia, UK, Finland, Belgium, and Japan. The language of interaction is English and there is no restriction on the topic of discussion. The records are made in Japan and hence there are several references to Japanese topics triggering the use of Japanese words at times. A small 360-degree lens attached to a Pointgrey Flea2 camera is used to record videos. The data is annotated manually for speech transcriptions, the topic of discussion and topic changes, the mood of the participants, speaker of the utterance and participant head movements.

2.6.2.5 Canal9 Dataset

The Canal9 dataset [Vinciarelli et al., 2009] is another multiparty multimodal dataset consisting of 70 interactions involving political debates between multiple participants. Total recording time for 70 interactions is 4 hours and 10 minutes. Each debate revolves around a single question e.g *"Are you in favour of a new law on scientific research?"*. The interactions involve: moderator and 4 participants. Moderator is expected to make sure that all the participants speak for an equal amount of time. Two of the participants answer yes or argue in the favour of the statement of the question whereas the remaining two oppose the question statement. A view of one of the meetings in Canal9 dataset is depicted in Figure 2.4.

The Canal9 dataset is manually annotated for speech transcriptions and participant role and has been automatically annotated for speaker segmentation.

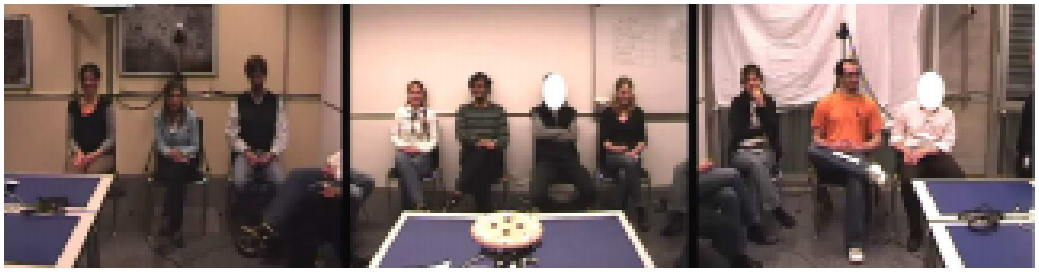

Figure 2.5 – Meeting view from the IDIAP Wolf Dataset ([Hung and Chittaranjan, 2010])

2.6.2.6 The IDIAP Wolf Dataset

The IDIAP Wolf dataset [Hung and Chittaranjan, 2010] is another multiparty, multimodal human-human interaction dataset consisting of 4 groups involving 8-12 people. The complete corpus consists of 7 hours of audio-video corpus. In each interaction is a session of the role playing game *Are you a Warewolf?*. The game consists of two alternating phases. (i) The night-phase in which the villagers are asleep (have their eyes closed) and the werewolves kills one of the villager of their choice (they make their decision in a discrete manner, so that it is not revealed to anybody except for the narrator). (ii) The day-phase in which the narrator reveals who was killed by the werewolves. In the second phase, all players still alive can discuss freely and decide whom they believe to be a werewolf. Figure 2.5 shows different meeting views in the IDIAP wolf dataset.

The IDIAP wolf dataset is annotated with speech transcriptions, participant roles, start and end time of each game and speaker segmentation.

2.6.2.7 Vernissage Dataset

Vernissage is another multimodal dataset involving interactions between two humans and a robot [Jayagopi et al., 2013]. The interaction scenario involves a robot that serves as art guide, introducing paintings to the participants followed by quizzing the participants in art and culture. Robot gaze, head nodes and dialogues are controlled via a wizard-of-oz setup. The interactions are annotated manually with a set of verbal and nonverbal cues including speech utterances, visual focus of attention, and 2D head location.

2.6.2.8 ELEA Dataset

The Emergent Leadership (ELEA) dataset [Sanchez-Cortes et al., 2011] is one of the most commonly used multimodal databases for the analysis of the Big Five personality traits, i.e. openness, conscientiousness, extraversion, agreeableness, and neuroticism, and other social traits including dominance, leadership, likeness and competence. It was initially

built to study leadership in small groups of 3-4 participants, and its relationship with participants' personality traits. Each group of participants is asked to perform winter survival task. After the task, the participants are asked to fill a questionnaire where they have to score other group participants in terms of dominance, likeness, leadership, competence and likeness.

2.6.3 Discussion

The Table 2.3 summarizes some of the previously presented multimodal datasets along with their characteristics.

The review of existing multimodal human-human and human-agent interaction datasets reveal some interesting facts. There is no all-purpose that can be used for proposing solutions to all the challenges and sub-tasks of multimodal human-agent interaction. The existing datasets are task specific and have not been fully annotated.

For instance, the Canal9 dataset [Vinciarelli et al., 2009] has been only annotated for speech transcriptions, video shot segmentation, participant roles etc. It has no information about the addressee of the utterance or the focus of attention. Similarly, the ELEA dataset has been developed in order to test the leadership, competence and dominance in human-agent interaction, hence it does not contain information about the DAs, addressee of the utterance, etc. Similarly, the JOKER [Devillers et al., 2015] dataset has been annotated for sense of humour and self-assessed personality.

To train models in the proposed research work, multimodal, multiparty human-agent interaction datasets with various annotations are required. The annotations required to train models in the proposed work are mentioned in Table 2.4.

The annotations are the gaze information of the meeting participants in multiparty interactions, the DA of utterances, the speaker and addressee information of each utterance, the text and duration of each utterance and the start and end time of each utterance. In addition to these, the propose models require that the datasets are multiparty i.e contain more than 2 participants. Among the datasets presented in Section 2.6, only the AMI and MPR contains annotations all the necessary annotations mentioned in Table 2.4. MULTISIMO contains all the annotations except the addressee information. The addressee information for the MULTISIMO dataset has been therefore manually annotated in this research work. Hence, to implement addressee detection, turn change and next speaker prediction, and participant gaze generation models, the AMI, MPR and the MULTISIMO datasets are used.

The next section presents some of the most common characteristics used to develop human-agent interaction models.

Dataset	Number of Participants	Number of Sessions	Recording Duration	Modalities	Annotations
AMI	4	40	100 hrs	audio, video	DAs, focus of attention, disfluency, extractive summaries, head, hand and leg movements, automatic & manual speech transcriptions, topic segmentation, you usage, speaker and addressee information, named entities
MPR	3	60	4 hrs	audio, video	participant status, focus of attention, speaker and addressee information, DAs, speech segmentation
MULTI-SIMO	3	25	4 hrs	audio, video	speech utterances, focus of attention, lexical, perceptual and demographic information, DAs
Freetalk	4-5	3	4 hrs 30 mins	audio, video	speech transcriptions, topic of discussion, topic changes, mood of the participants, speaker information and head movements
Canal9	5	70	4 hrs 10 mins	audio, video	participant roles, speach transcriptions, speaker information, shot segmentation
IDIAP Wolf	8-10	50	7 hrs	audio, video	speaker segmentation, speech transcriptions, start and end time of the game, participant roles
RECOLA	2	23	3.8 hrs	audio, video, ECG, EDA	valence, arousal, agreement, dominance, performance, rapport, engagement, utterance
Vernissage	3	13	2 hrs 23 mins	audio, video, motion capture	speech utterances, head-location, nodding, focus of attention, speech transcription
ELEA	3-4	40	10 hrs	audio, video	self-assessed personality, power, dominance, leadership, perceived leadership, dominance, competence, likeness
JOKER	2	111	8 hrs	audio, video, depth	self-assessed personality, sense of humour

Table 2.3 – Summary of some of the multimodal datasets

Annotation	**Addressee Detection**	**Turn Change and Next Speaker Prediction**	**VFOA Behaviour Generation**
Participant VFOA Information	✓	-	✓
Dialogue Act	✓	✓	✓
Speaker Information	✓	✓	✓
Addressee Information	✓	✓	✓
Utterance Text	✓	-	-
Utterance Duration	✓	✓	✓
Start and End time of Utterance	✓	✓	✓

Table 2.4 – Annotations and modules which require the corresponding annotations

2.7 Characteristics & Features for Human-Agent Interaction Models

This section presents some of the most commonly used interaction characteristics for developing rule based as well as statistical human-agent interaction models. The characteristics are called features when used for training statistical models. The analysis is based on existing literature. Analysis of some of the datasets is also performed to study the importance of features for different tasks.

2.7.1 Dialogue Acts

Dialogue act (DA) is one of the most important features for various multimodal human-agent interaction tasks such as addressee detection, turn change and next speaker prediction, and gaze generation. For instance, a DA involving a question often prompts turn change and provokes the addressee to respond to the question. Similarly, a DA such as stalling tells the listeners that the current speaker has not yet finished talking and that the turn is not available.

The importance of DA for different human-agent interaction tasks is highlighted in existing works [Jovanovic and op den Akker, 2004], [Petukhova and Bunt, 2009], [De Kok and Heylen, 2009], [Guntakandla and Nielsen, 2015] and [Meshorer and Heeman, 2016]. Furthermore, past works have shown that DA and eye gaze are strongly related [Poggi et al., 2000], [Andrist et al., 2014] thus DAs can help in gaze generation as well.

We perform the analysis of some of the existing datasets which also reveal the importance of DA for human-agent interaction. For instance, analysis of the AMI dataset reveals that 79% of the utterances having DA elicit info, i.e. a question, prompt a turn

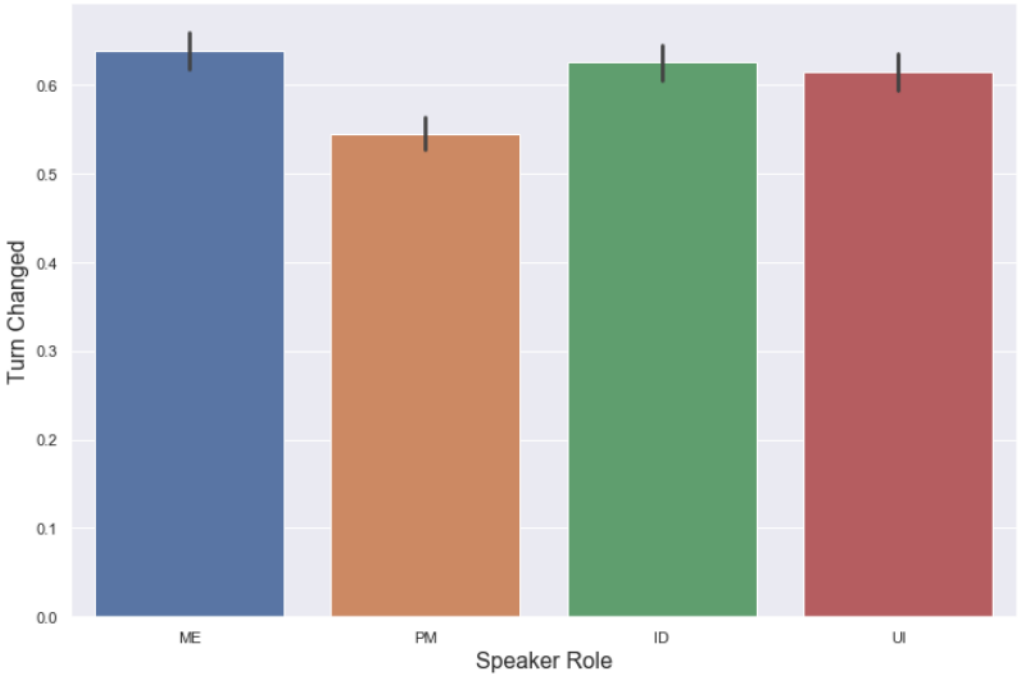

Figure 2.6 – Speaker Role vs Frequency of Turn Change in the AMI Dataset

change. The ratio is 72% and 68% for the elicit suggestions, and elicit assessment DA, that are also DA involving a question. If the previous DA is *elicit-info* and the previous speaker is any participant X and the previous addressee is participant Y, if the current speaker is participant Y, then 93% of the time, current addressee is participant X. The data analysis thus shows that, at least some of the DAs are actually important indicators for addressee detection.

2.7.2 Speaker Information

Speaker information refers to any information related to the speaker of the utterance such as name or role used to uniquely identify the speaker of the utterance. Since every participant in a multiparty interaction can speak, therefore every participant has a unique speaker role. For example, in AMI, the speakers are identified by their job roles i.e. Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI), and Industrial designer (ID). In the MPR corpus, speakers are identified by IDs with no semantic meaning e.g. A, B, C and NAO where A, B, C are the human participants and NAO is the name of a robot.

Speakers of previous and current utterances have also been used as features for various human-agent interaction tasks [Jovanovic, 2007], [Traum et al., 2004]. Work

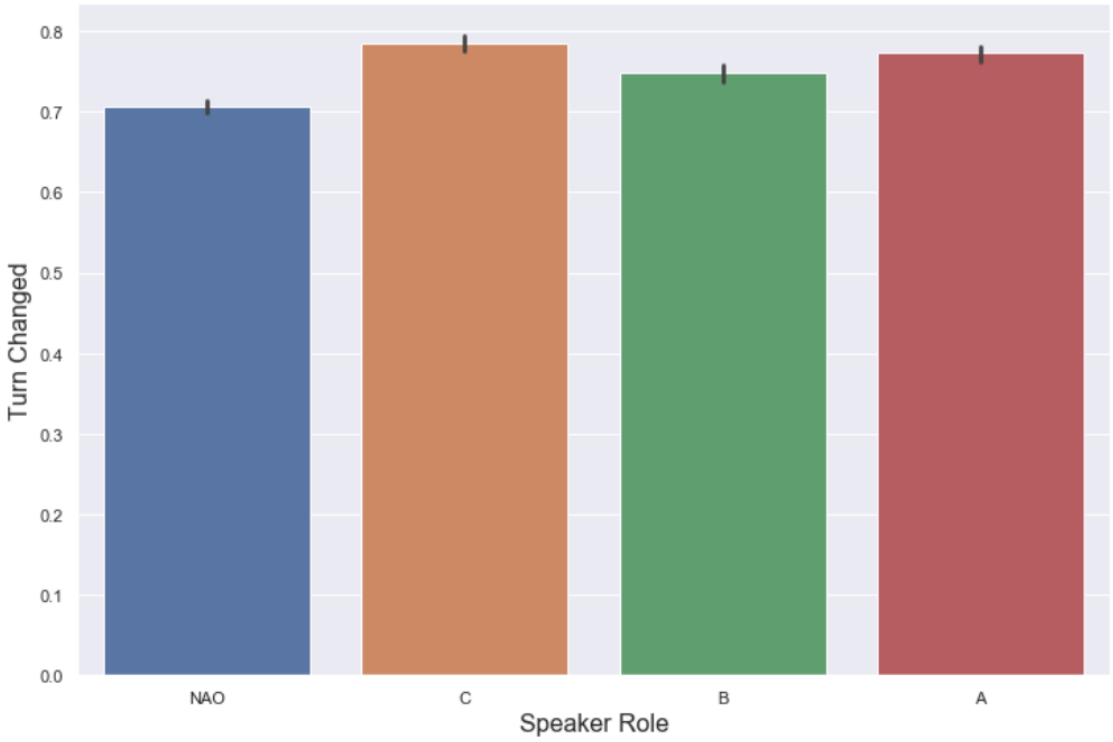

Figure 2.7 – Speaker Role vs Frequency of Turn Change in the MPR Dataset

from [Guntakandla and Nielsen, 2015] highlights the importance of current speaker for turn change prediction. Even intuitively, some speakers tend to keep talking whereas others more easily release dialogue turns. Our analysis of the AMI dataset confirms this claim. The frequency of turn change after each utterance for different speakers in the AMI dataset is depicted by Figure 2.6. The Figure shows that frequencies of turn change vary with the speaker role. For instance, PM has lowest frequency of turn change after the current utterance which shows that PM likes to keep the turn or interaction conditions forces PM to keep the turn compared to other meeting participants. Similar trend is observed for NAO robot in case of MPR dataset, as shown in Figure 2.7.

Study of the eye gaze behaviour in the last century shows that human visual focus of attention (VFOA) depends upon three factors: the task, the stimuli, and the observer [Buswell, 1935; Yarbus, 1967]. The observer refers to meeting participants including speakers, addressees and bystanders. [Yarbus, 1967] further claimed that *"it may be concluded that individual observers differ in the way they think and, therefore differ also to some extent in the way they look at things"*, which highlights that the VFOA behaviour is dependent on individual observers. The works from [Guy et al., 2019] show that VFOA behaviour varies with individuals.

In addition, analysis of the AMI and MPR datasets shows that on average the number

of VFOA turns vary by speaker. Some speakers have a longer utterance duration but less VFOA direction changes while others have shorter utterances but a higher number of gaze direction changes, which highlights the role of speaker information for gaze generation in human-agent interaction.

2.7.3 Addressee Information

[Goffman, 1981] defines addressee as: *"Those ratified participants oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants"*. Like speaker role, addressee role refers to the name or identity that is used to identify the meeting participant being addressed. Previous addressee is commonly used for current addressee detection in multiparty interaction [Akker and Traum, 2009].

The analysis of the AMI dataset further reveals that utterances addressed to individual are more likely to cause turn change as compared to utterances addressed to groups as shown in Figure 2.8 and Figure 2.9. Thus, addressee information can also be used for turn change prediction. Furthermore, the analysis shows that more than half of the utterances are addressed to the whole group rather than individual participants. The addressee count is higher for PM among individuals because the PM has the highest frequency of speaking, and there is a higher chance that people reply to her.

2.7.4 Pause Duration

Pause duration refers to the duration between two consecutive utterances where none of the interaction participants speak. When two or more people converse, there are three possibilities for how utterances follow each other as shown in Figure 2.10: (i) there can be partial overlap between two consecutive utterances where the second utterance may start before the end of the first utterance. In this case, the pause is negative since second utterance starts before the first utterance ends. (ii) there can be full overlap where the first utterance starts before the second utterance and ends after the second utterance. Finally, (iii) there can be no overlap between utterances. In this case the pause duration is positive.

Psycho-linguistic evidence shows a strong correlation between pause duration and turn change [O'Connell and Kowal, 2012]. A comparison of quantitative analysis of speaker changes in empirical and conversational settings reports that in empirical settings, 91% of the speaker changes occur with a pause between the utterances whereas 8% of the turn changes occur with no pause, and while only 1% of the utterance are

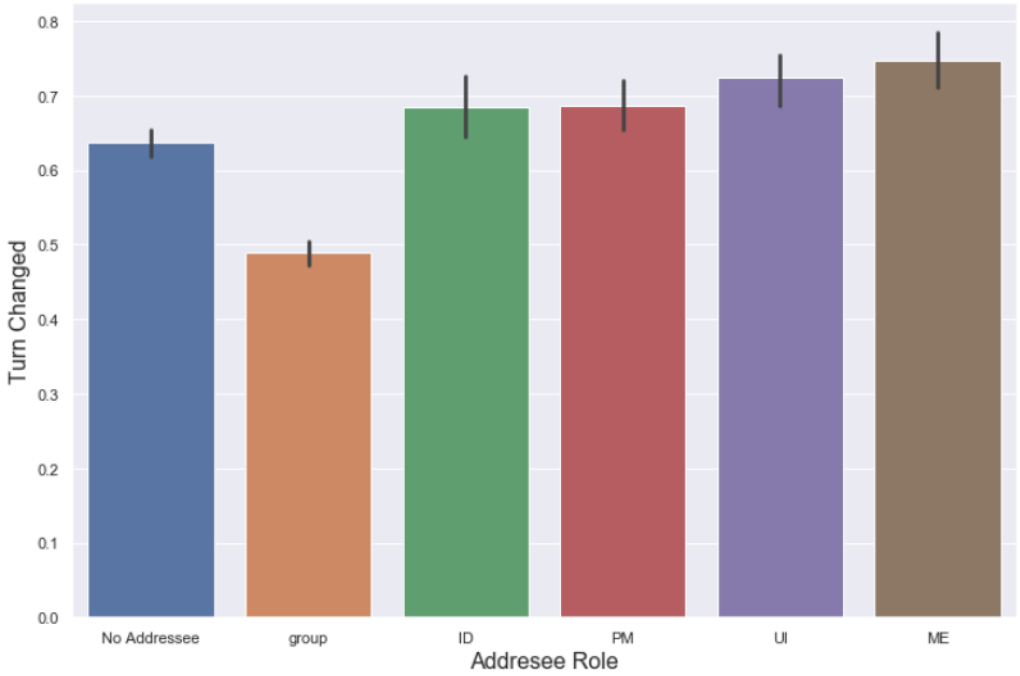

Figure 2.8 – Addressee Role vs Frequency of Turn Change in AMI Dataset

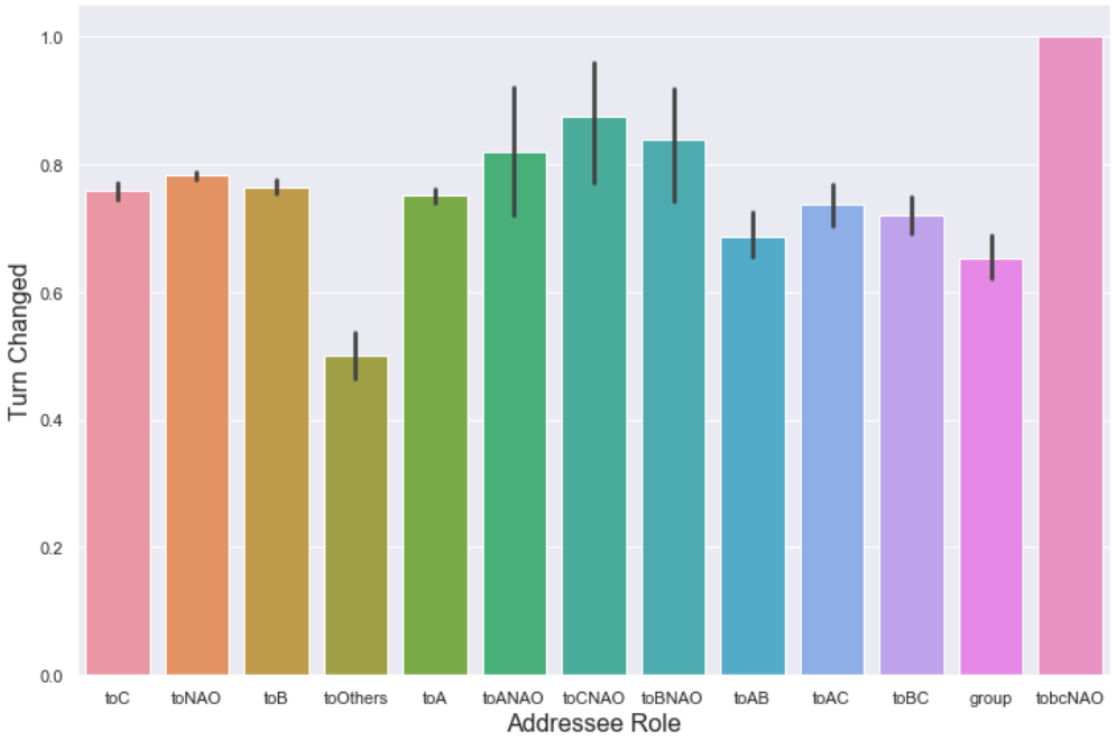

Figure 2.9 – Addressee Role vs Frequency of Turn Change in MPR Dataset

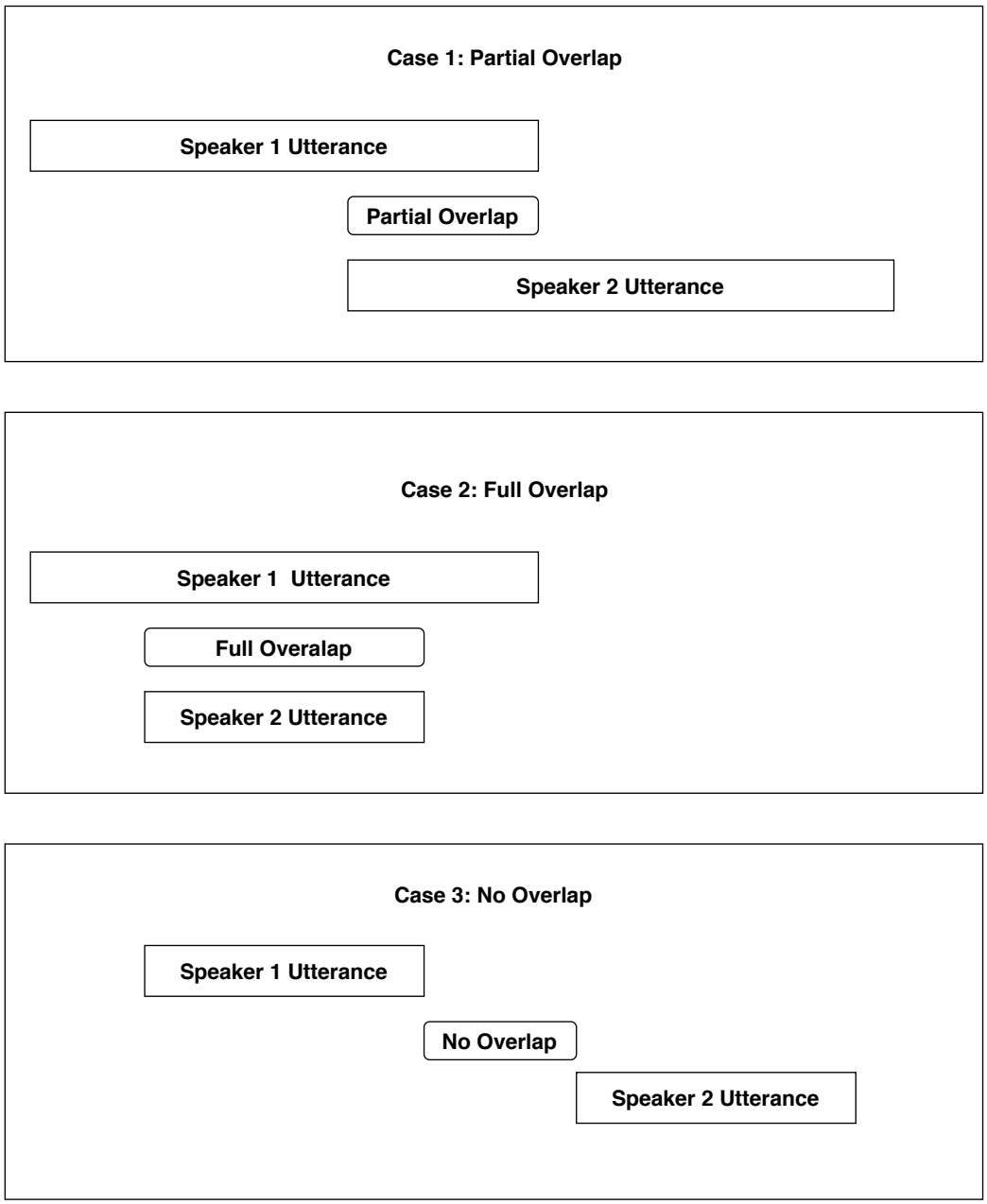

Figure 2.10 – Pause duration between two consecutive utterances

overlapped by another one. For conversational settings 92% of the speaker changes occur with a pause between two consecutive utterances. The authors further report that the frequency of turn changes is higher in empirical settings as compared to conversation settings. This trend can also be noted in the analysis of the AMI and MPR datasets. Since AMI dataset involves discussion and brainstorming, the frequency of turn change is lower (59.20%) compared to MPR (74.06%) which mainly involves question answering.

[Hilbrink et al., 2015; Ten Bosch et al., 2004] further suggest that pause duration is an important indicator for speaker turn change. [Hilbrink et al., 2015] explain the turn taking behaviour between a mother and an infant. They report that mothers use longer pause duration when they want the infant to say something, compared to the situations where they want to keep the turn [Takeuchi et al., 2003] show that pause duration can be employed to generate the response time for a robot in dyadic interactions.

2.7.5 Adjacency Pairs

In linguistics, an adjacency pair is an example of conversational turn-taking. An adjacency pair is composed of two utterances by two speakers, one after the other. Literature work has shown that adjacency pairs is a marker for addressee detection [Galley et al., 2004]. Intuitively, a response in an adjacency pair is addressed to the speaker of the first utterance in the adjacency pair.

To see if adjacency pairs actually play a role in addressee detection, the percentage of utterances where the previous speaker is the current addressee is computed using the AMI and MPR datasets. The result shows that in AMI, of all the utterances addressed to individual participants, only 32% utterances has the current addressee as the previous speaker, whereas 31% of the utterances are addressed to the whole group. Similarly, for MPR in 45% utterances the current speaker was the previous addressee. These results show that adjacency pairs alone are not a good indicator of addressee.

2.7.6 Important Keywords

Identification of certain keywords in speaker's utterance can help to perform different human-agent interaction tasks. For example, [Gupta et al., 2007] show that utterances that contain "*you*" are usually addressed to an individual user and thus the word "*you*" can be used for addressee detection.

Analysis of AMI reveals that of all the utterances where the word "you" is used, the utterance is addressed to individuals and the focus of attention is also an individual, only 42.44% of utterances are addressed to the focused individual. However, this number increases up to 78% when the group is addressed and multiple objects are in focus. This indicates that the "you" usage can be exploited to distinguish between individual and group addressees.

2.7.7 Utterance Duration and Number of Words

The duration of utterance and the number of words in an utterance are indicators of some important characteristics of human-agent interaction. [Meshorer and Heeman, 2016] shows that the utterance duration and lengths are an important indicator for turn change.

Furthermore, the duration of an utterance is directly related to the number of VFOA target changes. Longer utterances tend to have higher number of gaze direction changes compared to shorter DA. This relation is confirmed by the analysis of AMI and MPR datasets where a linear trend is observed between the number of VFOA target changes per utterance and average duration of the utterance.

2.7.8 Focus of Attention

Focus features containing the gaze target of meeting participants can be used to identify various aspects of multimodal interaction. For instance [Vertegaal, 1998], [Akker and Traum, 2009], [Le Minh et al., 2018], used focus information for addressee detection. This claim has been evaluated on the AMI as well. Table 2.5 shows the percentage of addressee against the focus of the speaker. The table depicts that when the focus is on an individual during an utterance, the individual is the addressee almost half of the times. The values of 0.48, 0.52, 0.50 and 0.47 for ID, ME, PM and UI substantiates this argument. Furthermore, if the individual under focus is not the addressee, the utterance is normally addressed to the whole group. Only in rare cases does the speaker look at one individual to then address another individual.

Focus	ID	ME	PM	UI	Group
ID	**0.48**	0.04	0.02	0.02	0.42
ME	0.03	**0.52**	0.03	0.02	0.38
PM	0.01	0.02	**0.50**	0.03	0.44
UI	0.03	0.01	0.02	**0.47**	0.44
Multiple	0.05	0.05	0.07	0.07	**0.74**
no	0.10	0.08	0.15	0.08	**0.57**
Slide Screen	0.08	0.05	0.27	0.06	**0.52**
Table	0.07	0.12	0.14	0.10	**0.55**
Whiteboard	0.05	0.14	0.11	0.07	**0.60**

Table 2.5 – Frequency of Focus vs Addressee in AMI dataset (values in percentage) ID: Industrial Designer, ME: Marketing Executive, PM: Project Manager, UI: User Interface Expert

Similarly, if the speaker is looking at multiple users, 74% of the time the utterance is

addressed to the whole group. If the Slide screen is the focus of the speaker, she normally addresses the group as shown in the table. The results show that the speaker focus is actually crucial for addressee detection in this corpus.

In addition, the importance of gaze or focus of attention as a marker for turn change and next speaker prediction has been investigated by several researchers [Petukhova and Bunt, 2009], [De Kok and Heylen, 2009], [Kawahara et al., 2012], [Ishii et al., 2013], [Ishii et al., 2015a]. The results from these research works show that focus of attention is an important feature for turn change and next speaker prediction. Particularly, the participant in focus near the end of the utterance of the current speaker is often the next speaker.

2.7.9 Start and End Time of Utterance

The start and end time of an utterance can also play an important role for making various decision in human agent interaction [Meshorer and Heeman, 2016]. Start and end time of an utterance determine the duration of utterance which is an important indicator of turn change as explained in the previous section. In addition, the analysis of the AMI and MPR datasets show that the utterances at the beginning of a conversation are less likely to have turn change compared to utterances at later stages of a meeting which shows that start and end time of utterance can also be used as a separate feature, in addition to calculating utterance duration.

2.7.10 Hand and Head Gestures

Gesture information such as head and hand gestures is also used by agents for decision making in human-agent interaction. The works from [Petukhova and Bunt, 2009], [Kawahara et al., 2012], [Ishii et al., 2015b], [Hossain and Muhammad, 2019], show that head gaze and hand gesture movements can be used for addressee detection, emotion detection, turn change and next speaker prediction.

2.7.11 Speech

Speech refers to the features related to human speech such as prosody, intensity and pitch of voice. Existing works show that speech features are commonly used in multimodal human-agent interaction for emotion detection, turn change and next speaker prediction [Aldeneh et al., 2018], [Ranganathan et al., 2016], [Kawahara et al., 2012].

2.8 Discussion

Seamless working of end-to-end multimodal human-agent interaction system such as an embodied conversational agent entails execution of various tasks. In addition, interactions involving more than 2 participants further complicates the communication process. Several challenges emerge when multiple modalities have to be simultaneously processed in order to decide the next step in the interaction. Owing to these challenges, most of the existing works in multimodal human-agent interaction either attempt to solve one of the challenges (see section 2.5 or propose solutions for common tasks in multimodal human-agent interaction (see section 2.4).

Though several researchers have proposed solutions to various challenges and tasks in multimodal human-agent interaction, existing solutions to some of the tasks are not flexible enough to be implemented in multiparty scenarios. Furthermore, in addition to various multimodal features, exploitation of the context features can also help improve existing multimodal human-agent interaction systems. For instance, in multimodal human-agent interaction, it is important to know who was the previous speaker or previous addressee of an utterance, what topic is being discussed, and what is the role of different participants, in addition to individual features such as speech, head gestures and other multimodal signals of the participants.

This research work proposes machine learning based approaches for multiparty multimodal human-agent interaction. The proposed approaches exploit multimodal and context features in a multiparty interaction that have not been previously exploited in order to design three interdependent modules multiparty human-agent interaction. This research work improves: (i)addressee detection, (ii) turn change and next speaker prediction, and (iii) VFOA behaviour generation in multiparty, multimodal human agent interaction.

There are three main reasons for choosing these problems (i) the problems are generic enough to be implemented in any multimodal, multiparty human-agent interaction system, (ii) the problems are quite under-explored, specially in multiparty settings and, (iii) there is a room for improvement in the performance of the existing systems.

The proposed research not only outperform the baseline models but also adds flexibility to the existing modules so that they can be used with varying number of participants in different multiparty scenario.

The next chapter presents the overall system work flow along with the methodologies adopted for the development of the models proposed in this research work. The evaluation criteria for the models is discussed.

CHAPTER 3

System Work Flow, Methodology, and Evaluation Criteria

This chapter describes the overall workflow for the proposed models including dependencies among them, followed by the explanation of methodology used to develop the models, and to conduct the experiments. The chapter also contain details about the performance metrics used to evaluate the models presented in this research work.

The chapter is divided into 4 sections. Section 3.1 explains process flow chart. Section 3.2 presents methodology and evaluation metrics. A brief study comparing manual and ASR transcriptions for dialogue act annotation is presented in section 3.3. Section 3.4 contains discussion.

3.1 Process Flowchart

The overall workflow for the various tasks performed by the proposed models is depicted in Figure 3.1. At the center of the flowchart, lay three related models: addressee detection, turn change and next speaker prediction and Visual focus of attention behaviour (VFOA) generation models. The three models rely on supervised machine learning and heuristic approaches.

As explained in chapter 2, machine learning approaches learn from data that is represented in the form of feature set. The goal of supervised machine learning approaches is to find a relationship between input feature set and the ground truth output label. The learned relationship is then exploited by the model to make predictions on

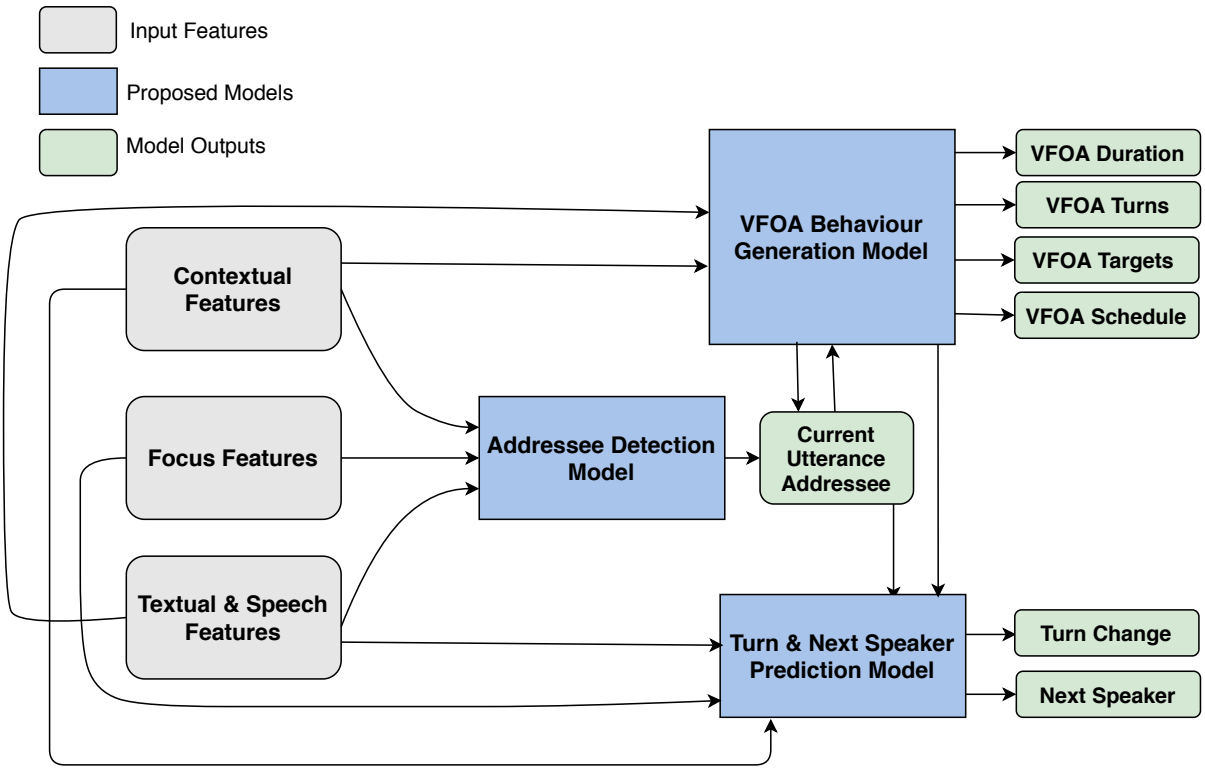

Figure 3.1 – Process flowchart of the proposed models

unseen data. In supervised machine learning, a ground truth output label is the true or real value for the output contained in the dataset. The output value predicted by a trained machine learning model for the unseen data is referred as predicted output label. Heuristic approaches can be used individually or can be exploited to refine machine learning based models in order to further improve the performance of the system. In addition, heuristic approaches can be independently exploited to formulate sets of rules to perform various tasks.

The input to the proposed models is a feature set that can be decomposed into three types of features i.e. contextual features, textual and speech features and focus features. The features are selected (i) on the basis of their importance in order to perform the corresponding task as found in the literature review (see section 2.7), and (ii) the statistical analysis of the features from the datasets as mentioned in the feature selection sections of Chapter 4,5 and 6.

To identify contextual features, the definition from [Dey et al., 1999] is considered: *"any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."*. Textual and speech

features are the features related to the content of an utterance such as the text of the utterance, the duration of the utterance, the number of words in the utterance, etc. Focus features, also named as Visual Focus of Attention (VFOA) concerns the targets i.e. persons or objects, that an individual can look at during an interaction. In multiparty interaction, the visual focus of attention of the speaker can be a key clue to identify who is the addressee of the utterance.

In real-time interactions, the VFOA behaviour generation model depends on the addressee detection model since the VFOA model takes current addressee as one of the inputs. Vice versa, the addressee detection model is also dependent on the VFOA generation model since addressee detection requires focus to predict addressee. Error in prediction of one of the model propagates to the other. Therefore, in order to get the best possible results, instead of using predicted addressee from the addressee detection model, for VFOA behaviour generation the ground truth value of current addressee from the dataset is used for training and testing. Similarly, for addressee prediction, machine learning models are trained using ground truth focus features instead of relying on the predictions from the VFOA behaviour generation model. The real-time mutual dependency between the VFOA behaviour generation and addressee detection models implies that improving performance of one of the models can also improve performance of the other model for real-time VFOA and addressee detection. Similarly, in real-time, the turn change and next speaker prediction model also relies on addressee and VFOA since turn change and next speaker prediction model requires current addressee and VFOA as features. However to avoid error propagation during training, the ground truth values of the current addressee, and speaker focus are used as input for the turn change and next speaker prediction model.

The following sub-sections briefly review the addressee detection, turn change and next speaker prediction, and VFOA behaviour generation models.

3.1.1 Addressee Detection

The goal of the addressee detection model is to detect the addressee(s) of the current utterance. In the proposed research work, addressee detection has been framed as a supervised multiclass machine learning problem since in multiparty interaction, an utterance can be addressed to any of the meeting participants, all the meeting participants or a subset of the meeting participants.

The overall flowchart for the addressee detection model is depicted in Figure 3.2. The addressee detection model consists of a machine learning model that takes as input focus, contextual and textual features to predict the addressee of the current utterance.

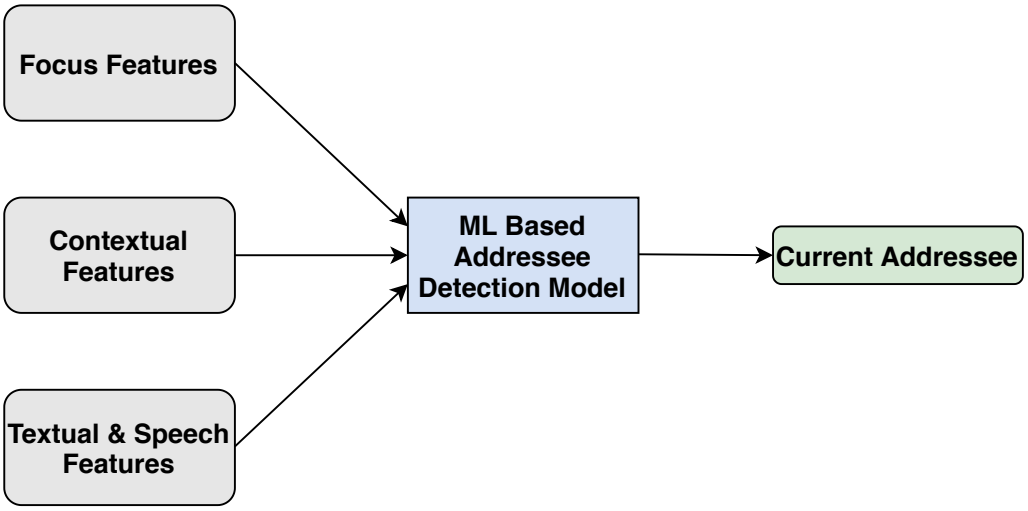

Figure 3.2 – Addressee Detection Model

Among the focus features, the VFOA of the speaker and listeners are used to train the model. Contextual features consists of the DAs of the previous and current utterances, the speaker of the previous and current utterances and the addressee of the previous utterances. Finally, textual features include the length of the utterance such as the number of words, the total duration of the current utterance, and whether or not the utterance contains the word *you*.

3.1.2 Turn Change and Next Speaker Prediction

The turn change and next speaker prediction model is also a stand-alone model that performs two tasks: turn change prediction, and next speaker prediction. The turn change occurs when the speaker of an utterance is different from the speaker of the previous utterance. The turn change prediction is a binary classification task with two possible outputs: whether or not a turn change occurs after the current utterance. On the contrary, next speaker prediction is a multiclass classification problem since in multiparty interaction, any of the participants can take a turn and become the speaker. Combined turn-change and next speaker prediction model has also been proposed where predicted turn change is included in the feature set used to train the next speaker prediction model.

Figure 3.3 shows the flowchart of the turn change and next speaker prediction model. The flowchart is similar to addressee detection model flowchart. Since there are two models (turn change and next speaker prediction) combined feature vector is used as input to both models.

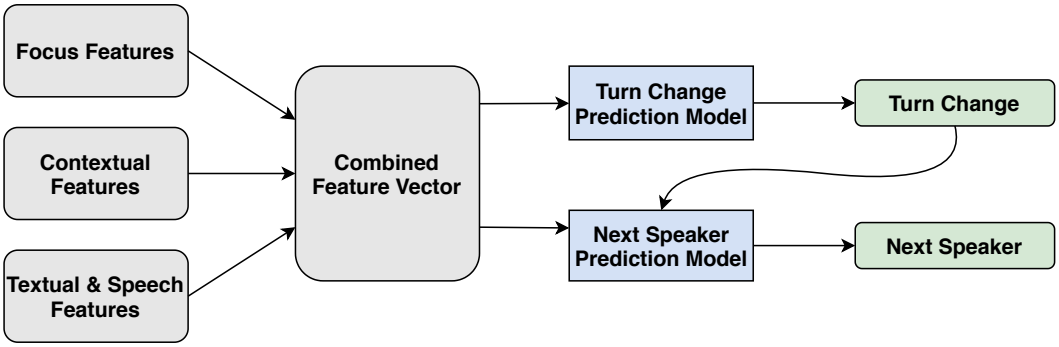

Figure 3.3 – Turn Change and Next Speaker Prediction model

The turn change and next speaker prediction model is trained using contextual, textual and focus features. From contextual features, the turn change and next speaker prediction model exploits the DA, speaker, and addressee of the current utterance, along with the pause between the current utterance and the next utterance and the start and end time of an utterance. Among focus features, the focus of the speaker is used to train the turn change and next speaker prediction model. From textual and speech features, the models exploit start and end time of the utterance.

3.1.3 Visual Focus of Attention Behaviour Generation Generation

The VFOA behaviour generation model is a hybrid model that predicts the final VFOA for intelligent agents while speaking and listening in multiparty interaction. The VFOA behaviour generation model consists of four sub-models: VFOA turns predictor, VFOA duration predictor, VFOA target predictor, and finally VFOA scheduler. Figure 3.4 shows the flowchart for the VFOA behaviour generation model.

The input to the VFOA turn predictor, VFOA duration predictor, VFOA target predictor sub-models is a set of contextual features that include the speaker of the current and previous utterances, the addressee of the current and previous addressee, the DA and its duration, and the information about the meeting participants that are not in the meeting at that time.

The number of VFOA turn predictor model decides how many turns there are in the participant VFOA during an utterance. The VFOA target predictor indicates the targets (persons or objects) that each participant should look at during an utterance. There can be multiple targets per utterance since there can be multiple VFOA turns. The VFOA duration predictor model predicts the duration of VFOA for each target in a turn. Finally,

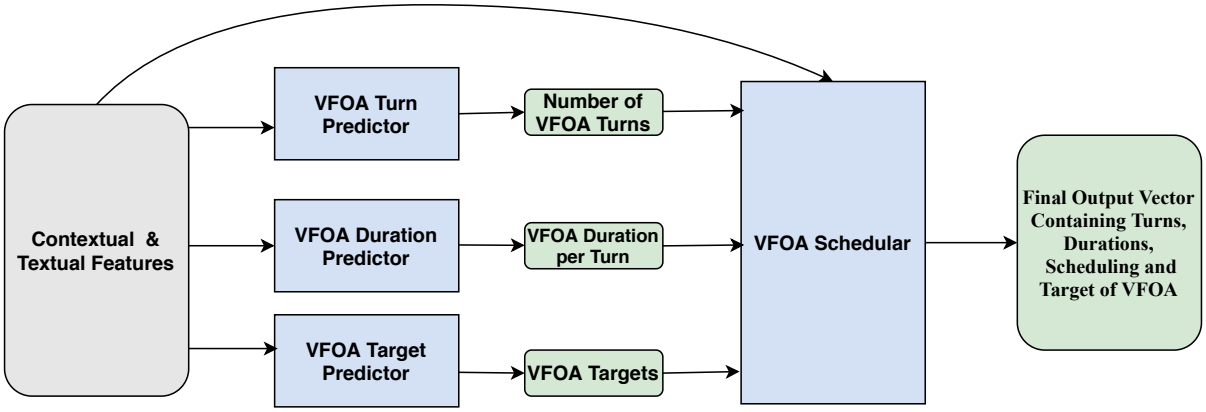

Figure 3.4 – VFOA Generation Model

the VFOA scheduling predictor decides the sequence in which participants look at other participants or objects during an utterance.

The VFOA turns predictor model is a regression model where the output is a continuous value i.e. the number of VFOA turns per utterance. The VFOA duration predictor is a multi-output regression problem where the outputs represents the VFOA distribution for all the targets (participants and objects) involved in the interaction. The VFOA target predictor model tackles a multiclass classification problem where the output is one or multiple meeting participants or objects that an agent should look at while speaking or listening.

During an utterance, the speaker and the listeners of the utterance can look at one or multiple participants in a sequence. The sequence in which the meeting participants change VFOA from one object or participant to another is referred to as VFOA schedule in this thesis. The outputs from the VFOA turn predictor model, the VFOA target prediction model, and the VFOA duration prediction model are transmitted as inputs to the VFOA scheduler model that schedules the VFOA.

The next section discusses the methodology followed to develop the proposed model, and the evaluation process and metrics.

3.2 Methodology and Metrics

The addressee detection model, and the turn change and next speaker prediction models are based on machine learning algorithms. In addition, for VFOA behaviour generation, three of the four models proposed in this research work i.e. VFOA turns predictor, VFOA duration predictor and VFOA target preditor, are also based on machine learning models. On the other hand, the VFOA scheduler combines machine learning and heuristic

approaches. This section briefly reviews machine learning methodology and evaluation criteria for performance

3.2.1 Methodology

In this research work, a standard machine learning pipeline has been adopted to learn, test and evaluate performance of the proposed models. The models are mostly based on traditional supervised machine learning algorithms.

In supervised machine learning, the actual output or the ground truth label is provided for each input record in the dataset. The task of machine learning algorithms is to find the relationship between the input features and the output labels.

Figure 3.5 briefly explains the methodology adopted to train the machine learning models in this proposed research work.

The first step in training a supervised machine learning algorithm is to collect the dataset. K Fold cross validation technique is used to divide the data into training and test sets and evaluate the machine learning models. In K fold cross validation, the dataset is divided into K parts. K-1 parts are used to train the algorithm and the K^{th} part is exploited test set to evaluate the performance of the machine learning algorithm. Each of the K data parts is used at least once for training and once for the evaluation. In the proposed research, 5 fold cross validation has been used for MPR and AMI datasets. For MULTISIMO dataset, K Fold cross validation has not been used owing to the limited number of records. The models are trained using training set and then predictions are performed on the test set. The predictions are further processed via heuristic approaches when necessary, to get the final results which are used for model evaluation.

To select the best parameters for some of the machine learning based models, a grid search algorithm has been used. Grid search algorithm looks for the best possible combination of parameters that yield the highest performance on the training set. The machine learning algorithms trained via the training set are employed to make predictions on the unseen test. In some models, the default parameter values as mentioned in the [Pedregosa et al., 2011] library, are used since training the model with default parameters yield better results than the baseline models. Therefore, owing to time constraints no further parameter tuning is performed using grid search. In order to evaluate different models, various evaluation metrics are used that are detailed in the next section.

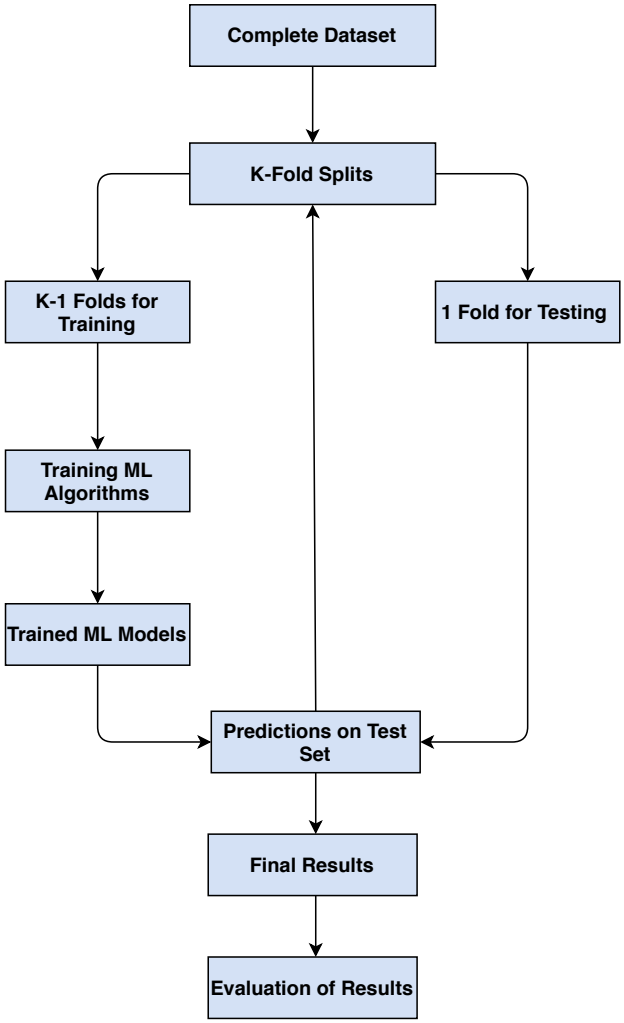

Figure 3.5 – Supervised Machine Learning Process

3.2.2 Evaluation Metrics

To check how well a model is performing, various approaches can be exploited depending on several factors such as the type of model, baseline results, the type of problem i.e. regression or classification, the distribution of output class labels in the dataset, the number of output labels, etc.

Most of the supervised machine learning problems can be further divided into classification and regression problems. Different performance metrics can therefore be used.

3.2.2.1 Evaluation Metrics for Classification models

Several classical evaluation metrics exist to evaluate the performance of a classification task. They are based on the confusion matrix, which is a 2-dimensional structure that

shows the predicted negatives and positives, and true or actual negatives and positives for each class in the form of a table. Figure 3.6 shows an example of a confusion matrix. True negative outputs are actually negative and also predicted as negative. Similarly true positive outputs are both actually and predicted as positive. On the other hand, false negative outputs are the ones that are actually positive but predicted as negative. Similarly, false positive outputs are those that are actually negative but predicted as positive.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Figure 3.6 – Confusion Matrix

The most commonly used metrics are accuracy, precision and recall, and F1 measures:

$$Accuracy = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Positives} + \text{All Negatives}} \tag{3.1}$$

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.2}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.3}$$

$$F_1 = 2 \text{ x } \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

Accuracy is suitable for the evaluation of algorithms trained on balanced dataset (dataset containing records with similar distribution for all output classes) whereas F1 measure necessary concerning the performance of models trained on unbalanced data.

For binary and multiclass classification problems, accuracy, precision, recall and F1 measures are good indicators of performance. However, these metrics are not suitable to evaluate multi-label classification as for example, accuracy considers a prediction as correct when all the labels in the output are 100% correct. In multi-label classification

problems, a prediction can be partially correct depending upon the number of correctly predicted labels. Micro-average F1 is one of the most common metrics to evaluate the performance of a multi-label classification algorithm. The equations to calculate micro-precision, micro-recall, and micro-f1 are as follows:

$$Precision^{micro} = \frac{\sum_{ci \in C} TruePositives(c_i)}{\sum_{ci \in C} TruePositives(c_i) + FalsePositives(c_i)} \tag{3.5}$$

$$Recall^{micro} = \frac{\sum_{ci \in C} TruePositives(c_i)}{\sum_{ci \in C} TruePositives(c_i) + FalseNegatives(c_i)} \tag{3.6}$$

$$F1^{micro} = 2 \text{ x } \frac{Precision^{micro} x Recall^{micro}}{Precision^{micro} + Recall^{micro}} \tag{3.7}$$

3.2.2.2 Evaluation Metrics for Regression models

A regression problem is a type of supervised machine learning problem where output is a continuous value. For example, the duration of VFOA per participant and object or the number of VFOA turns during an utterance are continuous values. The number of VFOA turns can be considered a discrete value with one output class per turn. However, technically there can be unlimited VFOA turns, hence VFOA turn prediction cannot be limited to a fixed set of classes.

Mean Absolute errors (MAE), Mean Squared error (MAE), and Root Mean Squared Errors (RMSE) are the most commonly used metrics for evaluating regression problem. Depending upon the problem, any of these metrics can be chosen. In this research work, MAE is chosen owing to its simplicity and explainability. The MAE can be calculated according to equation 3.8.

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} \left| y_{(i)} - y_{pred(i)} \right| \tag{3.8}$$

In equation 3.8, $y_{(i)}$ refers to the actual output value and $y_{pred(i)}$ corresponds to the predicted output value.

3.2.3 User Evaluation

For evaluation of a model where a users rates different aspects in comparison with baselines, none of the machine learning metrics can be used. In such cases, statistical

significance tests such as T-Tests [Krzywinski and Altman, 2013] can be performed to find if the perception difference between the model and baselines is significant. P-values measure the statistical significance. In this research, a user evaluation of the VFOA behaviour generation model in comparison with baselines is conducted via T-Tests which results in p-values.

In addition to exploring the impact of ground truth and predicted output labels when used to train machine learning models, the impact of difference in training and testing data quality on the performance of machine learning algorithms is also studied. In this context, in the next section we investigate the particular problem of DA annotation using machine learning models trained via manual and automatic speech recognition (ASR) transcriptions.

3.3 Dialogue Act Annotation using Manual vs ASR Trained ML Models

In a typical DA annotation process, speech utterances are manually transcribed and annotated with DAs. These utterances are then exploited to train statistical models that can predict the DA of new utterances. As transcriptions using ASR usually contain errors, manual transcriptions are often preferred over ASR transcriptions during the learning of DA annotation model (*e.g.* [Amanova et al., 2016; Can et al., 2015]). However, in common human-agent interaction situations, models used to predict DAs are exploited on utterances transcribed via ASR, which increase the annotation error rate as learning and exploitation conditions differ.

To confirm that models trained and tested via ASR transcriptions effectively return better results compared to models trained via manual and tested using ASR transcriptions, we perform an experiment using the methodology depicted in Figure 3.7.

The results (Appendix A) show that the algorithms trained on ASR transcriptions perform better or equivalent to models trained on manual transcriptions. For the models trained on ASR transcriptions, an accuracy of 66.87% is achieved in the best case, through random forest algorithm, using n-grams between 2 and 7 with top 25,000 features. This is significantly greater than the best accuracy of 57.71% achieved by random forest trained on manual transcriptions.

The most obvious reason for performance difference is that features generated using ASR trained transcriptions are more similar to ASR test transcriptions since transcription source is the same.

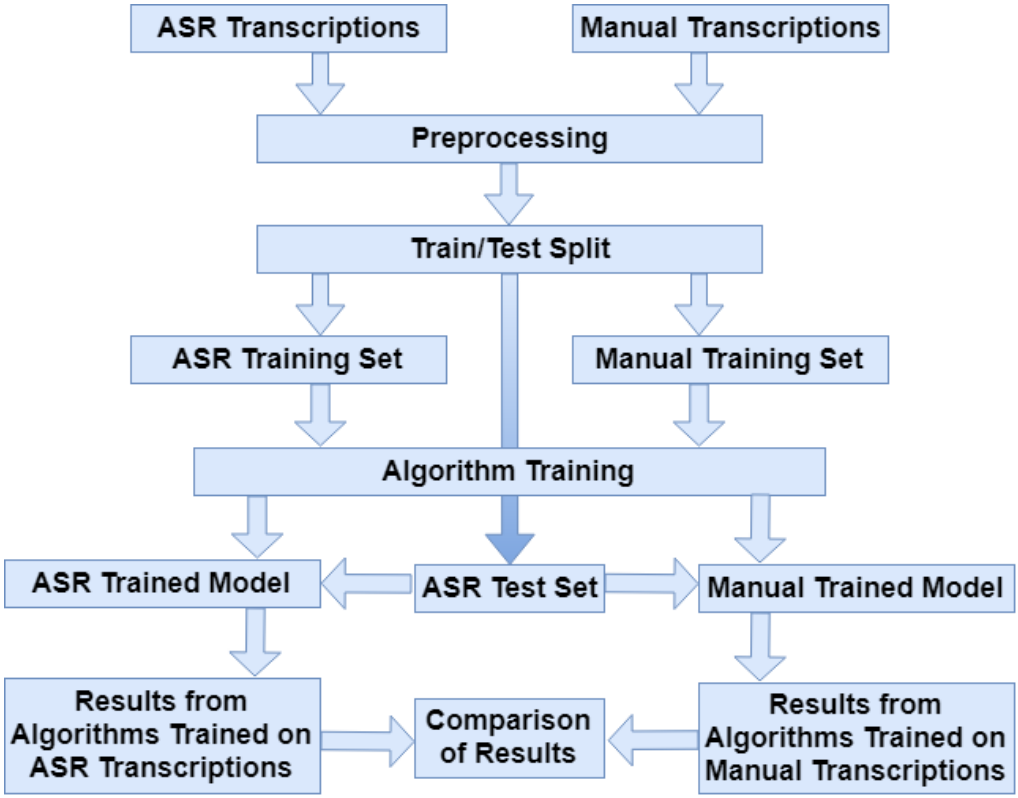

Figure 3.7 – Methodology for the Comparison of ASR vs Manual Transcription for DA Annotation

The results show that ground-truth data for various features may not be available during real interactions. Therefore, to use a machine learning model in real conditions, the algorithms should be trained on dataset that is more representative of real conditions rather than the most-accurate training sets.

3.4 Discussion

This chapter has presented a high level overview of the models. The proposed research work involves the development of three interrelated human-agent interaction models: addressee detection, turn change and next speaker prediction, and VFOA behaviour generation models.

The VFOA behaviour generation and the turn change and next speaker prediction model depend on the output of addressee detection model. However, in order to get the best results and to avoid error propagation, the proposed work uses ground truth values for addressee detection for training and testing in turn and next speaker prediction and the VFOA behaviour generation model.

For models based on machine learning, a typical machine learning pipeline has been adopted where data is divided into test and training sets via K-fold cross validation scheme. The models are learned via the training set and are evaluated via the test sets. Heuristic approaches can be further used to process the results of machine learning models or can be independently exploited to make predictions. In the proposed research work, the VFOA scheduler is a heuristic model that processes the results from the VFOA turns predictor, VFOA duration predictor and VFOA target predictor to make predictions regarding VFOA scheduling.

Model	**Task**	**Problem**	**Evaluation Metrics**
Addressee Detection	Addressee Prediction	Multiclass Classification	Accuracy and F1
turn change and next speaker prediction	Turn Change Prediction	Binary Classification	Accuracy and F1
turn change and next speaker prediction	Next Speaker Prediction	Multiclass Classification	Accuracy and F1
VFOA Behaviour Generation	Number of VFOA Turn Prediction	Regression with single output	MAE
VFOA Behaviour Generation	VFOA Duration Prediction	Regression with multiple outputs	MAE
VFOA Behaviour Generation	VFOA Target Prediction	Multi-label Classification	Micro-Average F1
VFOA Behaviour Generation	VFOA Scheduling	Classification	Accuracy
VFOA Behaviour Generation	Real Time Experiments	User Evaluation	T Tests and P Values

Table 3.1 – Evaluation metrics for evaluating different tasks

In the context of this thesis, the performance evaluation of addressee detection, and turn change and next speaker prediction models is straight forward since they are classification problems where correct output labels are available. On the other hand, for instance, the evaluation of VFOA behaviour generation model is a complex tasks because there can be multiple correct answers. For instance, a speaker can look at any participant during a multiparty interaction. In VFOA behaviour generation model, the performance of the individual models can be evaluated via common machine learning metrics. However for evaluation and comparison with the baseline models during real interactions, traditional classification and regression metrics cannot be used. To evaluate the performance of the overall VFOA behaviour generation model, user

surveys are conducted and the results are evaluated via statistical significance between the performance of proposed and baseline models.

Table 3.1 summarizes the evaluation metrics used to evaluate the tasks performed by various models in the proposed research work.

A major challenge for machine learning based systems is the unavailability of ground-truth values of some features during real interactions. A system consisting of multiple machine learning models rely on each other. During real interactions, values predicted by machine learning algorithms are not always correct. Hence error is propagated from one machine learning model to another, which further affects the performance of the models that rely on other machine learning models.

Furthermore, the values of some of the features are consistently changing during real interaction. For instance, duration of an utterance consistently updates during real interaction. Machine learning algorithms on the other hand are trained on a fixed set of features. Therefore, a mechanism is needed to deal with the consistently changing feature values.

In the next three chapters, the addressee detection (chapter 4), turn change and next speaker prediction (chapter 5), and Visual Focus of Attention generation (chapter 6) models are detailed. The corresponding chapters include detailed problem formalization, a comprehensive comparison of related works, the methodology adopted for experimentation, and the results obtained.

CHAPTER 4

ADDRESSEE DETECTION

Addressee detection is a fundamental task for seamless dialogue management and turn taking in human-agent interaction. Though addressee detection is implicit in dyadic interaction, it becomes a challenging task when more than two participants are involved.

This chapter proposes addressee detection models based on smart feature selection and focus encoding schemes. The models are trained using different machine learning algorithms. This research work improves existing baseline accuracies for addressee prediction on two datasets. In addition, we explore the impact of different focus encoding schemes in several addressee detection cases.

The chapter is divided into seven sections. The first section contains a brief discussion of addressee detection mechanism in conversations. Section 4.2 presents the related work specifically on addressee detection. The process of feature selection is explained in section 4.3. Section 4.4 explains the focus encoding techniques. Section 4.5 formalizes the problem while experiments and results are explained in section 4.6. The chapter ends with a discussion in section 4.7.

4.1 Addressee Detection Mechanisms in Dialogues

In the domain of Human Agent Interaction (HAI), a virtual agent is a system able to interact with humans and other virtual agents in order to accomplish a specific task. An intelligent agent has to perform several key tasks during a dialogue, such as speaker identification, intent classification or addressee detection. In dyadic interactions, involving two participants, addressee detection is straightforward, whereas addressee

detection in a multiparty context is a complex task. A speaker can address any specific participant, a subset of participants or all the participants. To generate an appropriate response, an agent needs to know when it is being addressed. In addition, if an agent is not being addressed, it is still important to understand who else is being addressed, in order to better contribute to the interaction.

In multiparty interaction, addressee detection involves multimodal information, such as speech utterances, hand gestures, facial expressions and focus of attention. Textual features among which the content of the utterance, can also be exploited to identify who is being addressed when the speaker explicitly mentions an interlocutor [Akker and Traum, 2009]. Furthermore, contextual features such as the previous speaker, previous addressee(s), and current and previous dialogue acts also play a crucial role in the identification of the addressee [Akker and Akker, 2009]. During an interaction, a speaker expresses an intention in the form of a dialogue act (DA). DA can be defined as the meaning of an utterance at the level of illocutionary force [Searle, 1969]. As per [Goffman, 1981], a DA can be addressed to three types of addressees: *over-hearers* who are not concerned by the interaction and whose dialogue states are thus not changed; *participants* whose dialogue states are changed by the speaker utterances but are not addressed by the speaker, and finally the direct *addressees* of the DA. This chapter focuses on the detection of direct addressee and henceforth the term "*addressee*" refers to the direct addressee(s) of an utterance. Direct addressees are defined as "*those ratified participants oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants*" [Goffman, 1981].

To tackle the problem of addressee detection in multiparty interaction, both heuristic and statistical approaches have been developed in the literature. However, most of these works depend on specific settings. Limited amount of training data also makes it difficult to develop a generic addressee detection model. In this research work, we propose a model which revolves around the encoding of generic features that are used to train machine learning algorithms, capable of addressee detection. We hypothesize that a statistical model with generic features could perform well on the addressee detection task in multiple scenarios. The obtained results on different datasets with varying number of participants substantiate this claim.

4.2 Related Works

This section presents an overview of related addressee detection approaches and features used for this specific task.

4.2.1 Approaches for Addressee Detection

The seminal work by [Traum et al., 2004] proposes a rule based approach exploiting previous utterance, current utterance, previous speaker and current speaker to detect the addressee. The accuracy varies between 65% and 100% on Mission Rehearsal dataset [Traum et al., 2006] depending upon the DA. However, the algorithm does not generalize well on other datasets: applied to the AMI dataset [McCowan et al., 2005], the accuracy drops to 36%. [Akker and Traum, 2009] improve this work by incorporating gaze as one of the foundations of the rules, resulting in an accuracy of 65%. The authors also test the gaze as the only feature for addressee detection, reporting an accuracy of 57%. In this case, the only rule is that if a speaker looks at a participant for more than 80% of the time duration of an utterance, the addressee is that participant, otherwise the utterance is addressed to the group.

Concerning the statistical approaches, [Jovanovic, 2007] use Baysian Networks to exploit current and previous utterance, speaker, gaze, topic of discussion and other meta features on the M4 multimodal corpus [Jovanovic et al., 2006] with an accuracy of 81%. This algorithm is also tested by [Akker and Traum, 2009] on the AMI corpus, with an accuracy of 62% that illustrates that the algorithm does not generalize well. [Akker and Akker, 2009] propose a statistical model based on logistic regression trees in order to answer the binary question *are you being addressed*, with a best case accuracy of 92% on AMI. The two limits of this work are firstly that the authors only consider a single speaker point of view instead of identifying who is being addressed, and secondly that the model depends upon a fixed positioning of the participants. [Baba et al., 2011] exploit human-human-agent triadic conversations to develop a SVM based model that distinguishes whether an utterance is addressed to the human or the agent. They report an accuracy of 80.28% for this binary classification task, using text, head orientation and acoustic features.

Finally, only few works have used deep learning techniques to tackle the addressee detection problem. [Le Minh et al., 2018] propose a convolutional neural network [Krizhevsky et al., 2012] based solution for addressee detection in the *GazeFollow* dataset [Recasens et al., 2015]. One major limitation of this work is that the addressee detection is performed through third party angle, with an accuracy of 62.5%.

4.2.2 Features Exploited in Existing Works for Addressee Detection

A natural starting point of feature selection for creating addressee detection models is the study of human behavior in human-human-interactions.

[Sacks et al., 1974] explore the importance of adjacency-pairs in addressee detection. An adjacency-pair consists of two consecutive utterances that can take form of a question-answer, statement-agreement and so on. The information related to the first utterance of the pair, e.g previous speaker, previous addressee, etc., are useful for predicting the next addressee [Galley et al., 2004].

[Vertegaal, 1998] explores the impact of gaze for addressee detection and reports that in 77% of the utterances, the person whom the speaker is looking at is actually the addressee of the utterance. However, this finding cannot be generalized in cases where a virtual agent is part of the meeting. [Bakx et al., 2003] show that in triadic communication where one of the participant is a multimodal agent, the user looks at the agent 94% of the time. However, while addressing the other user, 57% of the time the gaze remains on the multimodal agent. This shows that the presence of a virtual agent can monopolize the speaker's focus.

DA can also be an indicator of addressee as reported by [Jovanovic and op den Akker, 2004]. DA conveys the intent of the speaker e.g. seeking information, providing information, acknowledging something, etc. An utterance can contain multiple DAs, addressed to different participants, for instance, the utterance *Are you feeling hungry?* (addressed to an individual), followed by *May be we should take lunch now* (addressed to the group).

4.2.3 Summary and Discussion

A detailed summary of the existing addressee detection approaches in multiparty interaction is presented in table 4.1.

These approaches either solve binary classification problem such as differentiating between an agent or a human [Baba et al., 2011], or depend upon the fixed positioning of the participants [Akker and Akker, 2009; Jovanovic, 2007]. Furthermore, most of the systems have low accuracy and do not scale well over different numbers of participants [Akker and Traum, 2009; Le Minh et al., 2018; Traum et al., 2004].

In this research work, we claim that a model with generic features can be used for real-time addressee detection irrespective of the dataset and of the number and positioning of participants. The requirements for are mentioned in table 4.2. The rationale behind

References	Approaches	Evaluation Datasets	Salient Features	Accuracies / Accuracies on AMI (%)
[Traum et al., 2004]	Rule Based	Mission Rehearsal Exercise [Traum et al., 2006], AMI [Carletta, 2007]	Current & previous utterance and speaker	65-100 / 36
[Akker and Traum, 2009]	Rule Based	AMI [Carletta, 2007]	Gaze, current & previous utterance, speaker and addressee	65 / 65
[Jovanovic, 2007]	Bayesian Network	M4 [Jovanovic et al., 2006], AMI [Carletta, 2007]	Current & previous speaker and utterance, topic of discussion, gaze, etc.	81 / 62
[Akker and Akker, 2009]	Logistic Model Trees	AMI [Carletta, 2007]	Current & previous utterance, current & previous speaker, topic of discussion, gaze, etc	92 / 92
[Baba et al., 2011]	SVM	Home made	Head orientation, acoustic features	80.28 / NA
[Le Minh et al., 2018]	CNN, LSTM	GazeFollow [Recasens et al., 2015]	Utterance and gaze information	62 / NA

Table 4.1 – Existing addressee detection approaches for multiparty interaction

Requirement	**Description**
r1	Independent of the number of participants
r2	Independent of the participant positioning
r3	Prediction of the addressee instead of prediction if an utterance is addressed to the agent or not
r4	Should be able to make predictions in real conditions

Table 4.2 – Model requirements

these requirements is that the participants who are actually not being addressed should also be aware of who is being addressed in real-time (r3, r4), independently of how they are located in the room (r2) and of how many participants there are (r1).

This research work makes 4 improvements over the baseline model [Akker and Traum, 2009]: (i) Novel features for addressee detection as evident from the research work are selected (ii) Multiple machine learning algorithms, that are never used for addressee detection before are exploited (iii) we propose two focus encoding approaches

that improves existing baseline accuracy for addressee prediction. The impact of alternate focus encoding schemes on the performance of addressee detection models is studied (iv) the changes needed for addressee detection models to work at real time are implemented.

4.3 Feature Selection

Features have been selected owing to their importance as mentioned in the literature review. Selected features are intended to be generic and thus do not depend upon an underlying scenario: the number of participants and their positioning are not taken into account. Feature analysis of some of the features is further performed on the AMI [Carletta, 2007] and MULTISIMO corpus [Koutsombogera and Vogel, 2018] since these are two datasets that contain all the features exploited for addressee detection in this research work.

4.3.1 Speaker and Listener Focus of Attention

Works from [Akker and Traum, 2009; Le Minh et al., 2018; Vertegaal, 1998] show that focus is an in important feature for addressee identification. The analysis of AMI dataset further highlight the importance of focus for addressee detection. The analysis show that the focus is on an individual during an utterance, the individual is the addressee almost half of the times. Furthermore, if the individual under focus is not the addressee, the utterance is normally addressed to the whole group. Only in rare cases does the speaker look at one individual to then address another individual. Similarly, if the speaker is looking at multiple users, 74% of the time the utterance is addressed to the whole group. If the Slide screen is the focus of the speaker in the AMI dataset, the speaker normally addresses the group. The results illustrate that the speaker focus is actually crucial for addressee detection in this corpus.

4.3.2 Current and Previous Dialogue Acts

DAs play an important role in conversational tasks such as addressee detection [Jovanovic and op den Akker, 2004]. If a DA is a question to an individual, the response is normally addressed to the speaker of the question.. Furthermore, the statistical analysis of AMI corpus reveals that if the DA of the previous utterance contains a question, then 80% of the utterances are addressed to previous speaker. This percentage rises up to 93% in MULTISIMO. This illustrates the importance of DAs as a feature.

4.3.3 Current and Previous Speakers

Speakers of previous and current utterances have also been used as features for various human-agent interaction tasks. [Jovanovic, 2007; Traum et al., 2004]. The analysis of the AMI and MULTISIMO datasets show that when the current and immediate previous speakers are different, the current addressee is the immediate previous speaker (62 % in AMI corpus 63% in MULTISIMO).

4.3.4 You Usage

Usage of the word *"you"* in an utterance is a key indicator of the addressee. [Gupta et al., 2007] show that utterances that contain "*you*" are usually addressed to an individual user. Therefore the usage of the word "*you*" in an utterance is an important indicator for addressee detection.

Analysis of the MULTISIMO and MPR datasets also reveal some interesting facts in this regard. In MULTISIMO, when an utterance contains the word "you" and the focus is an individual, 70% of the time, that individual is an addressee. In AMI, this number is 42.22%. In cases where "you" is used in the sentence and the focus is multiple people, 85% of the time, the addressee is the group. This percentage is 78% in AMI. The difference may again be attributed to the difference in number of participants and objects that a person can look at in the MULTISIMO and AMI corpora.

4.3.5 Utterance Duration and Length

Utterance duration and length are the textual features as they depend upon the text of the DA. Utterance length refers to the number of words in an utterance whereas utterance time corresponds to the time taken produce the utterance.

[Webb et al., 2005] show an increase of 4% in overall accuracy for DA annotation task when multiple models are trained on utterances of uniform lengths compared to single model trained on utterances of varying lengths.

Longer utterances have longer durations and lengths. Utterance refers to a single DA in this research work. The analysis of AMI dataset reveals that utterances shorter in duration tend to have lesser number of VFOA turn changes. The analysis further reveales that for utterances with only one VFOA turn and where VFOA target is a participant, that participant is often the addressee of the utterance.

4.3.6 Discussion

Existing research works have explored both machine learning and heuristic techniques for addressee detection in multiparty interaction using various interaction features. However, none of the existing approach uses all the features presented in this section. For instance, [Traum et al., 2006] uses current and previous addressee, the speaker information but does not use focus information, previous DA etc. Similarly, [Baba et al., 2011] only uses acoustic features and head orientation for addressee detection. The proposed research work predicts addressee of the current utterance by training machine learning models on maximum number of features, that exist in one dataset and that have been recommended as addressee marker from existing research work.

Existing works show that focus is an important clue for addressee detection. Since in this research work, addressee is predicted at the level of DA and focus can shift multiple times during a DA, there can be various ways to encode speaker and listener focus. However, to the best of our knowledge, none of the existing works study the impact of different focus encoding schemes on addressee detection.

4.4 Focus Encoding Schemes

Focus of listeners and speakers is a marker for addressee detection. To further investigate the importance of focus, both AMI and MULTISIMO have been analysed. In AMI, a person can look at 3 other participants, a slide screen, a table and a whiteboard. When neither the participants nor the objects are in focus, the focus is marked as *other*. In MULTISIMO, a participant can only look at 2 other participants, otherwise the focus is marked as *other*. Table 4.4 shows a relationship between addressee of an utterance and speaker focus during the utterance. For instance, the first row of table 4.3 illustrates that, if the majority of the focus is A, 52.5% of the time A is the addressee. Similarly, this value is 58.5% in MULTISIMO (first row of table 4.4). The difference in the share of focus can be attributed to the number of participants or objects that a person can look at (7 entities in AMI and 2 in MULTISIMO).

A brief description of the focus encoding schemes is given in the two following subsections.

4.4.1 One-hot Focus Encoding

During the course of a DA, the focus of attention can be any individual or object. The following steps explain the process of focus annotation of each utterance:

Focus / Addressee	A	B	C	PM	Group
A	0.525	0.034	0.025	0.031	0.385
B	0.043	0.481	0.027	0.023	0.426
C	0.015	0.036	0.479	0.025	0.446
PM	0.020	0.011	0.034	0.509	0.426
Multiple	0.054	0.053	0.076	0.071	0.745
Other	0.084	0.101	0.081	0.158	0.576
Slide Screen	0.052	0.080	0.065	0.279	0.524
Table	0.121	0.074	0.110	0.143	0.552
Whiteboard	0.141	0.054	0.079	0.116	0.610

Table 4.3 – Relationship between speaker focus and addressee of an utterance in AMI

Focus/ Addressee	A	B	PM	Group
A	0.585	0.032	0.064	0.319
B	0.000	0.707	0.096	0.192
PM	0.037	0.037	0.916	0.009
Other	0.24	0.19	0.30	0.27
Multiple	0.121	0.121	0.152	0.606

Table 4.4 – Relationship between speaker focus and addressee of an utterance in MPR

1. if there are more than two participants and/or objects in focus, the focus is marked as *multiple*;

2. if the focus is on a participant and an object then

   a) if the person is first in focus and then the object, the focus is marked as the person,

   b) otherwise the focus is marked as multiple.

From figure 4.1, for one-hot encoding only one of the participants or objects can have a value of 1 while the remaining entities have a value of 0.

4.4.2 Shared Focus Encoding

In shared focus encoding, ratios of total focus of a participant during an utterance are stored depending upon the time that the person focus is shared among other participants and objects (if any). A ratio of focus for a participant or object e.g. X is calculated as follows:

$$\text{Ratio of Focus for (X)} = \frac{\text{Overall duration of focus towards participant X}}{\text{Total duration of utterance}}$$

(4.1)

Figure 4.1 contains example of shared focus. In the given example for AMI, the value for PM is 0.4 which means that during the utterance the focus of the participant was PM for 40% of the total duration of the utterance. Similarly, the focus was respectively on participant C or on the table for 10% or 20% of the time, whereas the participant looked at something else during the remaining 10 percent.

An important point is that participants' focus seems not follow any particular sequence during an utterance. When a participant looks at a person or object multiple times, the values for the corresponding focus durations are simply added to calculate the final focus value.

4.4.3 Discussion

In figure 4.1, the shared focus vectors for AMI and MULTISIMO have different values while the corresponding vectors for one-hot encoded focus is similar. This illustrates that one-hot encoding scheme tends to make the representations uniform.

Another disadvantage of one hot encoding is that some degree of biasness is integrated in the participants focus. For instance, if a participant A looks at PM for 90% and at participant B for 10% of time during an utterance, the focus is marked as multiple, although participant A only glanced at participant B for a very short period. Alternatively, the focus is again marked as multiple if person B has 90% of the focus while PM has 10% of the focus. Hence, two different focus situations are marked as having similar multiple focus. In case of shared focus, the exact focus ratios during an utterance are stored.

4.5 Problem Formalization Datasets

4.5.1 Addressee Detection Model

Figure 4.2 depicts the proposed addressee detection model based on the features discussed in section 4.3 and 4.4. The figure shows an example of shared focus encoding scheme dedicated to AMI configuration. All the features except focus features are directly transmitted to the machine learning models. The focus features are first encoded via one-hot or shared focus encoding scheme and then passed to the ML models.

For predictions in real conditions (r4), the model is modified in two ways: (i) no previous addressee is used during the training and testing, (ii) the value of real-time

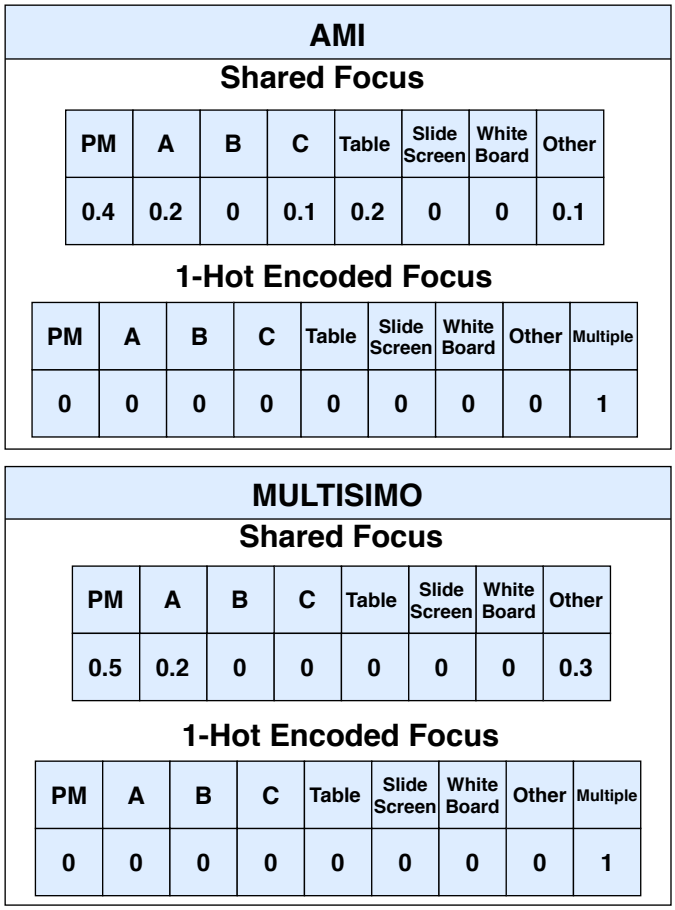

Figure 4.1 – Examples of one-hot and shared focus encoding vectors for AMI (top) and MULTISIMO (bottom) datasets.

previous addressee is exploited for training and predicted previous addressee is used for testing.

4.5.2 Datasets

The experiments are performed on the AMI and the MULTISIMO corpora since they are the only datasets that contain all the features required to train machine learning models in this research work. The MPR corpus also contains all the required features but unfortunately it was developed after the experimentation on addressee detection were completed.

Several differences exist between AMI and MULTISIMO. In AMI, each meeting has 4 participants: PM, UI, ID, and ME, whereas the meetings in MULTISIMO have 3 participants: Facilitator, Left Player and Right Player. Therefore, both corpora have been

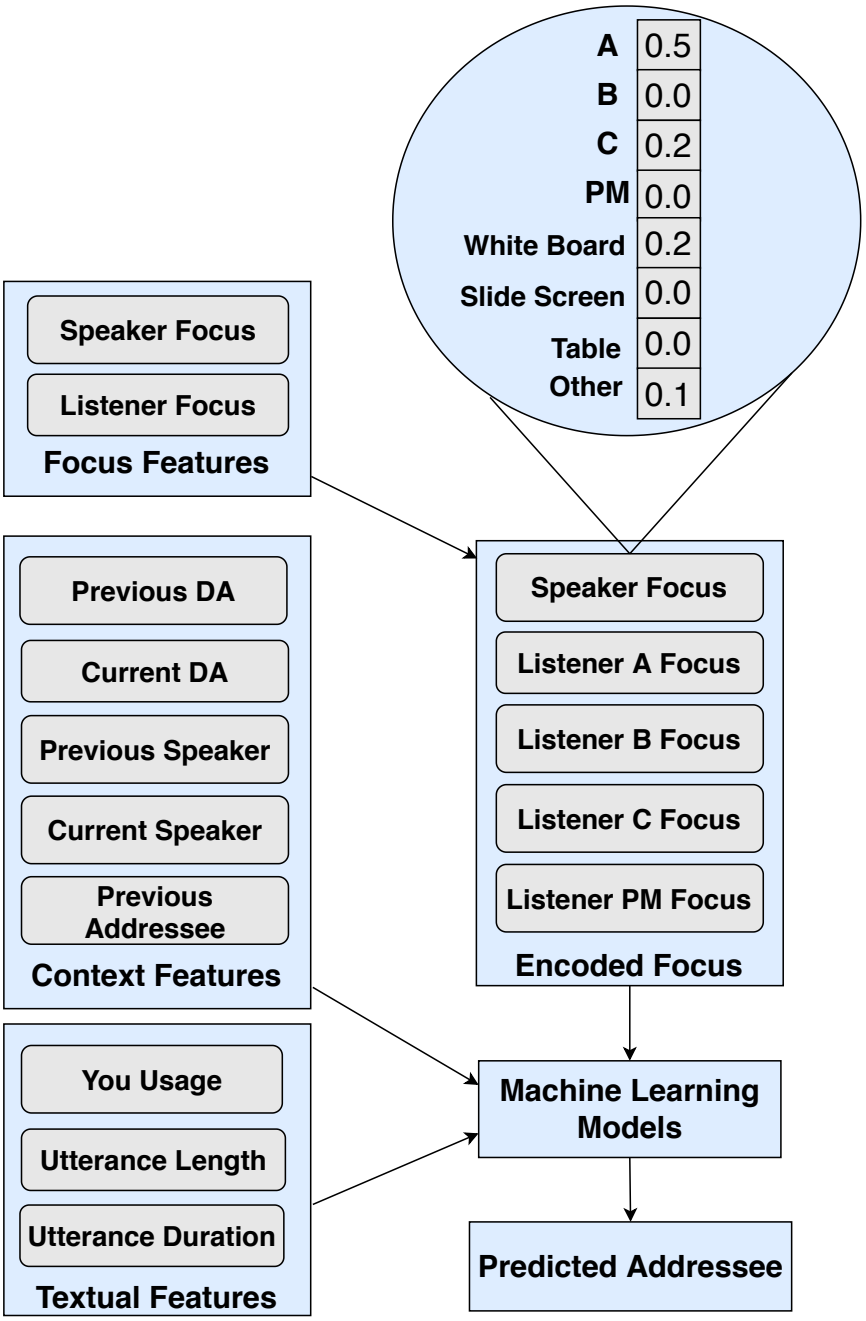

Figure 4.2 – Addressee detection model (Example of shared focus encoding from AMI dataset

AMI	MULTISIMO	New Name
PM	FC	PM
ME	Left Player	A
ID	Right Player	B
UI	-	C

Table 4.5 – Renaming convention of the participants

processed to rename the participants since (r2) states that the model should not depend upon the positioning of the participants. The table 4.5 contains the renaming convention.

In AMI, the PM acts as a moderator similarly to the facilitator in MULTISIMO. As both acts to regulate the meeting, they were given the same label for the sake of uniformity. In MULTISIMO, the two meetings containing gaze information are not annotated with DA and addressee. Therefore, the meetings were manually annotated for addressee and DA by two annotators[1]. Finally, in AMI the addressee has not been annotated for 'minor' DAs such as stalling, fragment, back-channel and others. Thus, the utterances containing these kinds of DA are also removed from MULTISIMO. Further details of the AMI and MULTISIMO datasets is presented in Section 2.6.2.

4.5.3 Problem Formalization

Table 4.6 highlights the working hypothesis. Hypothesis (h1) claims that the models proposed in this research work should improve existing baseline accuracies for both AMI and MULTISIMO. Hypothesis (h2) states that the model having a N participants should achieve similar or better classification performance when tested on a dataset with equal or less than N of participants. H2 is based on the intuition that the default probability of addressee prediction is higher in meetings with less participants as compared to the meetings with more participants. Hypothesis (h3-a) claims that in real-time the classification performance for addressee detection decreases since the ground-truth value of the previous addressee is not available. Hypothesis (h3-b) states that the classification performance is better when models are trained and tested without previous addressee as compared to the models trained with ground-truth previous addressee and tested with predicted previous addressee.

Given the set of features mentioned in Figure 4.2, the task performed is to predict, whether the current utterance of the speaker is addressed to listener A, B, C, PM or the whole group (in AMI) or listener A, B, PM or the whole group (in MULTISIMO).

[1]The annotation is available at: `http://pagesperso.litislab.fr/~jsaunier/ms_addressee_detection.csv`

Hypothesis.Id	Description
h1	The models proposed in this research work should perform better than baselines on AMI and MULTISIMO
h2	Models with participants N should achieve at least similar or better performance when tested on equal or less than N participants
h3	**(a)**: Model performance decreases when the ground-truth value of previous addressee is not available. **(b)**: Models trained and tested without previous addressee perform better than models tested with predicted previous addressee

Table 4.6 – Working hypotheses for the model

4.6 Experiments and Results

The section contains the detail of different experiments performed for addressee detection, along with the experimental process and the results obtained.

4.6.1 Experiments

The section describes the procedure followed for experiments, to satisfy requirements and validate hypotheses.

The set of experiments performed have been summarized in table 4.7. Eight experiments are performed to develop models for self-learning and inter-corpora-learning. In this research work, the term self-learning refers to experiments where models are trained and tested on the same dataset. On the contrary, inter-corpora-learning refers to experiments where models are either trained on AMI and tested on MULTISIMO, or vice-versa. The experiments for inter-corpora learning are performed to study the stability of the proposed models when evaluated in unseen meeting scenarios.

4.6.1.1 Experiments With ground-truth Previous Addressee Values

In experiments AA, MM, AM and MA, the machine learning models are trained and tested are with ground-truth previous addressee. Experiments AA and MM are performed to validate h1 which states that the proposed models should improve existing baselines for both AMI and MULTISIMO. These models are self-learning models where both training and test sets belong to the same dataset.

In AM and MA, the models are inter-corpora-learning models, trained on AMI and tested on MULTISIMO, and vice versa. They are performed to validate h2.

Experiments	Description
AA	Algorithms trained and tested on AMI
MM	Algorithms trained and tested on MULTISIMO
AM	Algorithms trained on AMI and Tested on MULTISIMO
MA	Algorithms trained on MULTISIMO and tested on AMI
AM-PN	Algorithms trained and tested on AMI without ground-truth previous addressee
MM-PN	Algorithms trained and tested on MULTISIMO without ground-truth previous addressee
AM-PN	Algorithms trained on AMI and tested on MULTISIMO without ground-truth previous addressee
MM-PN	Algorithms trained on MULTISIMO and tested on AMI without ground-truth previous addressee

Table 4.7 – Summary of the Experiments Performed

4.6.1.2 Experiments for Real-time Addressee Detection

To satisfy requirement r4 and to verify hypothesis h3, experiments are performed without ground-truth values for previous addressee since the ground-truth values are not available at real-time.

There exist two possible solutions to this problem: either the algorithms can be trained with ground-truth values for previous addressee but tested at real-time using as input the previously predicted addressee, or the algorithms are trained and tested without any previous addressee. In experiments AA-PN, MM-PN, AM-PN and MA-PN, the models are tested first with the predicted previous addressee and then without any previous addressee. The results are compared to find the best approach for addressee detection in real interactions where ground-truth value of the previous addressee is not available.

4.6.1.3 Experimental Process

A conventional machine learning pipeline is followed to perform experiments. Each type of experiment listed in Table 4.7 is performed with the two focus encoding schemes: shared focus and one-hot focus. The proposed model is independent of the number of participants. However, in this research work the model has been tested with 4 participants (AMI) and 3 participants (MULTISIMO). In case of a meeting with less than 4 participants, the focus encoding scheme does not have to be changed since 0 can be added in place of the vectors for the participants not present. For instance in MULTISIMO, C is not present, hence 0 can be added in place of C in both one-hot and shared focus encoding. In case of more than 4 participants, one vector has to be added for each new

participant and a new dimension needs to be added in all the existing focus vectors. To generalize, for N participants, the number of one-hot encoded focus vectors is N and the number of dimensions in each focus vector is N-1.

Feature encoding is followed by the the training of machine learning models on the preprocessed datasets that includes contextual and textual features as well as encoded focus. Nine of the most common classic machine learning classifiers have been trained and tested: Multilayer Perceptron (MLP) [Kruse et al., 2013], Random Forest (RF) [Liaw et al., 2002], Logistic Regression (LR) [Hosmer Jr et al., 2013], Support Vector Machines (SVM) [Hearst et al., 1998], Naive Bayes (NB) [Rish et al., 2001] and K-Nearest Neighbours (KNN) [Zhang and Zhou, 2005], XGboost (XGB) [Chen and Guestrin, 2016], Adaboost (ADB) [Hastie et al., 2009] and Extra Tree (ET) [Geurts et al., 2006]. Finally, tests are also performed with 3 deep learning algorithms: Long Short Term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997], One-Dimensional Convolutional Neural Network (1D-CNN) [Kiranyaz et al., 2019], and Bi-Directional Long Short Term Memory Network (Bi-LSTM) [Melamud et al., 2016]. Parameters of the classifiers used for AA, MM, AM and MA are summarized in Appendix B.

Concerning AA-PN, MM-PN, AM-PN and MA-PN, only the ensemble algorithms i.e. XGB, ET, ADB and RF have been tested since they show overall best accuracies for the experiments with ground-truth values for previous addressee.

For all the classifiers except LSTM, Bi-LSTM and 1D-CNN, parameter selection is performed using grid search [Smit and Eiben, 2009]. Unmentioned hyper-parameters are set to default values as specified in Python's Sklearn library [Pedregosa et al., 2011]. For the deep learning classifiers: LSTM, Bi-LSTM and 1D-CNN, parameters mentioned in Appendix A, are randomly initialized. Since models start to over fit [Hawkins, 2004] on training set using these parameters, no further parameters are tested. Five fold cross validation is performed to ensure that the models do not over fit and that the obtained results are stable. Finally, accuracy and F1 measure have employed used to evaluate performances. Accuracy is used to compare the results with baselines and F1 measure is considered since the class distribution is irregular in both datasets [Smit and Eiben, 2009].

4.6.2 Baselines

From existing works, Akker and Traum [Akker and Traum, 2009] use multimodal information from the AMI dataset to achieve an accuracy of 65% on the addressee detection detection. This work can be considered as one of the baselines as this is the only work that respects requirements r1, r2 and r3. For a fair comparison between the

baseline and the solution proposed, the Akker and Traum's algorithm is also tested on the AMI subset used in this research work, as a second baseline. Finally, we have observed that "Group" is the majority class in the AMI. Therefore, the majority class for addressee is proposed as the third naive baseline for AMI.

To have a baseline for MULTISIMO , the Akker and Traum's algorithm is also tested on MULTISIMO. The majority class of addressee is used as second baseline for MULTISIMO as well.

Although Akker and Akker [Akker and Akker, 2009] report a higher accuracy of 92% on AMI, their algorithm only supports 4 participants, with fixed sitting positions, hence violating requirements r1 and r2 considered in this research work.

4.6.3 Results

This section shows the results of all the experiments.

4.6.3.1 Using Ground-Truth Previous Addressee Values

In experiments AA, MM, AM and MA the ground-truth addressee value is used for both training and testing the algorithms. The results for these experiments have been summarized in tables 4.8-4.11.

The accuracies for AA, where models are trained and tested on AMI, are presented in table 4.8. The maximum accuracies of 78.77% and 74.26% are respectively achieved for shared-focus encoding and one-hot focus encoding via XGBoost (XGB) classifier which is higher than all the three baselines. The results show that shared focus encoding outperforms the one-hot focus encoding for experiments AA. The results are significant at $p < 0.05$ (p = 0.005).

Similarly for MPR, Table 4.9 depicts the accuracies for MM where models are trained and tested on MULTISIMO. The results show that for the shared focus encoding Extra Tree classifier (ET) achieves the highest accuracy of 88.81% which greater than the three baselines. For one-hot focus encoding, a maximum accuracy of 84.80% is achieved via the linear regression algorithm. The results for experiment MM also show that shared focus encoding outperform the one-hot focus encoding. However the results are not significant at $p < 0.05$ (p = 0.141) when all algorithms are compared.

For inter-corpora-learning, table 4.10 contains accuracies for AM where models are trained on AMI and tested on MULTISIMO. The results show that one-hot focus encoding approach achieves the highest accuracy of 77.19% with naive bayes. The best

Classifier	**Shared Focus**	**One-Hot Focus**
XGB	**78.77** ±0.016 (0.79)	**74.26** ±0.014 (0.74)
ET	77.28 ±0.012 (0.77)	72.17 ±0.012 (0.71)
ADB	75.02 ±0.014 (0.75)	71.24 ±0.017 (0.68)
RF	75.27 ±0.015 (0.75)	74.66 ±0.032 (0.73)
MLP	76.36 ±0.008 (0.76)	**74.26** ±0.029 (0.74)
SVM	76.78 ±0.009 (0.77)	73.09 ±0.020 (0.73)
KNN	73.76 ±0.012 (0.73)	68.65 ±0.002 (0.67)
LR	77.20 ±0.009 (0.77)	74.09 ±0.003 (0.72)
NB	75.85 ±0.013 (0.75)	64.12 ±0.030 (0.63)
LSTM	60.90 ±0.018 (0.60)	60.58 ±0.020 (0.60)
Bi-LSTM	58.14 ±0.020 (0.58)	59.35 ±0.001 (0.58)
1D CNN	57.05 ±0.012 (0.57)	58.26 ±0.019 (0.58)
Baseline 1: [Akker and Traum, 2009]	65%	
Baseline 2: [Akker and Traum, 2009]: same AMI subset	60%	
Baseline 3: Majority Class	54%	

Table 4.8 – Accuracies for AA (F1 in brackets)

Classifier	**Shared Focus**	**One-Hot Focus**
XGB	85.52 ±0.018 (0.86)	83.57 ±0.050 (0.79)
ET	**88.81** ±0.042 (0.88)	83.92 ±0.041 (0.84)
ADB	87.40 ±0.021 (0.87)	79.89 ±0.038 (0.80)
RF	86.12 ±0.028 (0.88)	82.50 ±0.021 (0.83)
MLP	81.88 ±0.024 (0.82)	83.82 ±0.028 (0.85)
SVM	84.25 ±0.037 (0.84)	83.34 ±0.041 (0.83)
KNN	81.88 ±0.046 (0.82)	84.32 ±0.032 (0.80)
LR	82.67 ±0.033 (0.82)	**84.80** ±0.037 (0.85)
NB	75.91 ±0.046 (0.75)	75.89 ±0.029 (0.75)
LSTM	35.50 ±0.031 (0.28)	46.85 ±0.027 (0.40)
Bi-LSTM	33.60 ±0.047 (0.25)	49.55 ±0.025 (0.46)
1D CNN	31.20 ±0.010 (0.27)	45.95 ±0.035 (0.43)
Baseline 1:[Akker and Traum, 2009]: Applied on MULTISIMO	77.01%	
Baseline 2: Majority Class	32.49%	

Table 4.9 – Accuracies for MM (F1 in brackets)

Classifier	Shared Focus	One-Hot Focus
XGB	**75.55** ±0.019 (0.75)	76.48 ±0.022 (0.77)
ET	69.55 ±0.032 (0.71)	73.60 ±0.029 (0.75)
ADB	64.30 ±0.019 (0.66)	55.45 ±0.022 (0.69)
RF	68.45 ±0.025 (0.70)	73.35 ±0.010 (0.73)
MLP	66.56 ±0.038 (0.66)	75.61 ±0.026 (0.76)
SVM	30.44 ±0.010 (0.14)	70.91 ±0.005 (0.70)
KNN	21.29 ±0.005 (0.08)	70.91 ±0.010 (0.69)
LR	31.54 ±0.026 (0.19)	76.66 ±0.002 (0.77)
NB	22.39 ±0.031 (0.13)	**77.19** ±0.009 (0.77)
LSTM	29.04 ±0.022 (0.27)	38.20 ±0.027 (0.37)
Bi-LSTM	32.11 ±0.019 (0.28)	39.31 ±0.230 (0.38)
1D CNN	30.52 ±0.023 (0.30)	31.70 ±0.025 (0.30)

Table 4.10 – Accuracies for AM (F1 in brackets)

Classifier	Shared Focus	One-Hot Focus
XGB	53.42 ±0.021 (0.52)	53.45 ±0.029 (0.52)
ET	43.40 ±0.035 (0.43)	47.26 ±0.031 (0.46)
ADB	46.76 ±0.015 (0.46)	42.11 ±0.029 (0.43)
RF	41.29 ±0.036 (0.40)	47.02 ±0.024 (0.47)
MLP	37.50 ±0.041 (0.37)	**56.27** ±0.017 (0.56)
SVM	38.40 ±0.005 (0.37)	46.66 ±0.009 (0.44)
KNN	43.50 ±0.010 (0.44)	43.12 ±0.004 (0.41)
LR	**53.77** ±0.031 (0.52)	54.64 ±0.021 (0.55)
NB	23.40 ±0.026 (0.26)	25.00 ±0.031 (0.08)
LSTM	15.63 ±0.019 (0.12)	26.51 ±0.008 (0.20)
Bi-LSTM	13.90 ±0.024 (0.14)	27.41 ±0.004 (0.24)
1D CNN	12.35 ±0.019 (0.11)	37.99 ±0.013 (0.35)

Table 4.11 – Accuracies for MA (F1 in brackets)

case accuracy achieved via shared focus is 75.55% via XGboost (XGB) algorithm. For experiments AM. The results are significant at $p < 0.05$ (p = 0.008) for all algorithms.

The accuracies for MA where models are trained on MULTISIMO and tested on AMI are presented in table 4.11. The results show that the highest accuracy achieved is 56.27% using one-hot focus encoding via Multilayer Perceptron. The results are significant at $p < 0.05$ (p = 0.009) for all algorithms.

4.6.3.2 Results for Real-time Addressee Detection

This section concerns the experiments where models are either trained with ground-truth previous addressee values and tested with predicted previous addressee, or trained

Approach	**Shared Focus**	**One-Hot Focus**
XGB (PA)	69.71 ±0.027 (0.69)	63.08 ±0.019 (0.62)
ET (PA)	65.18 ±0.019 (0.65)	61.32 ±0.034 (0.58)
ADB (PA)	65.77 ±0.016 (0.66)	61.82 ±0.014 (0.53)
RF (PA)	67.07 ±0.021(0.61)	61.07 ±0.018 (0.58)
XGB (NA)	**73.84** ±0.014 (0.73)	67.35 ±0.023 (0.68)
ET (NA)	73.09 ±0.015 (0.73)	**68.33** ±0.0012 (0.68)
ADB (NA)	69.90 ±0.021 (0.70)	65.63 ±0.024 (0.61)
RF (NA)	70.31 ±0.014 (0.74)	67.72 ±0.023 (0.65)

Table 4.12 – Accuracies for AA-PN. F1 in Brackets. (PA= Predicted Previous Addressee, NA = No Previous Addressee)

and tested without the previous addressee in the feature set. The results for these experiments, i.e. AA-PN, MM-PN, AM-PN and MA-PN, are summarized in tables 4.12-4.15. The first four rows of all these tables (PA) contain results for the experiments where models are trained with ground-truth previous addressee vales and tested with predicted previous addressee values. The last four rows (NA) contain results where models are trained and tested without ground-truth previous addressee values.

The accuracies for AA-PN, as depicted in table 4.12, show that the highest accuracy of 73.84% is achieved using XGBoost (XGB) with shared focus encoding when the model is trained without the previous addressee, which is 4.93% less than the accuracy achieved when the models are trained and tested with ground-truth previous addressee values (Experiment AA). Similarly for one-hot encoding, the highest accuracy of 67.35% is achieved without the previous addressee which is less than the maximum accuracy of 74.26% achieved actual previous addressee is included in the feature set (Experiment AA). The results are significant at $p < 0.05$ (p = 0.003 for shared focus encoding and p = 0.003 for one-hot encoding) for all algorithms.

Similarly, table 4.13 depicts accuracies for MM-PN. The results show that the highest accuracy of 86.61% is achieved using XGBoost (XGB) with shared focus encoding when the model is trained without the previous addressee, which is less than 88.81%, the accuracy achieved when the models are trained and tested with ground-truth previous addressee values (Experiment MM). Similarly for one-hot encoding, the highest accuracy of 80.35% is achieved without the previous addressee which is less than the maximum accuracy of 84.80% achieved actual previous addressee is included in the feature set (Experiment MM). The results are not significant at $p < 0.05$ (p = 0.21 for shared focus encoding) but significant for for one-hot encoding at (p = 0.00071).

The accuracies for AM-PN and MA-PN, performed to evaluate the inter-corpora-learning capabilities of the models trained using ground-truth previous addressees

Approach	Shared Focus	One-Hot Focus
XGB (PA)	82.52 ±0.014 (0.82)	79.27 ±0.018 (0.78)
ET (PA)	80.15 ±0.029 (0.80)	77.47 ±0.013 (0.78)
ADB (PA)	81.74 ±0.034 (0.82)	76.57 ±0.025 (0.77)
RF (PA)	84.92 ±0.017 (0.85)	77.47 ±0.024 (0.78)
XGB (NA)	**86.61** ±0.026 (0.87)	**80.35** ±0.021 (0.80)
ET (NA)	86.18 ±0.031 (0.88)	81.25 ±0.031 (0.81)
ADB (NA)	85.03 ±0.026 (0.85)	76.78 ±0.031 (0.77)
RF (NA)	86.61 ±0.013 (0.87)	80.35 ±0.023 (0.81)

Table 4.13 – Accuracies for MM-PN. F1 in Brackets. (PA= Predicted Previous Addressee, NA = No Previous Addressee)

Approach	Shared Focus	One-Hot Focus
XGB (PA)	67.35 ±0.134 (0.68)	68.94 ±0.016 (0.70)
ET (PA)	54.10 ±0.031 (0.57)	62.47 ±0.027 (0.65)
ADB (PA)	54.25 ±0.017 (0.56)	55.83 ±0.024 (0.56)
RF (PA)	60.88 ±0.021 (0.63)	62.11 ±0.031 (0.65)
XGB (NA)	**69.32** ±0.023 (0.70)	**69.78** ±0.024 (0.71)
ET (NA)	61.45 ±0.029 (0.64)	65.45 ±0.036 (0.68)
ADB (NA)	57.97 ±0.041 (0.60)	58.43 ±0.029 (0.58)
RF (NA)	65.71 ±0.036 (0.67)	66.02 ±0.031 (0.67)

Table 4.14 – Accuracies for AM-PN (F1 in brackets)

(PA), and without previous addressee (NA), are reported in Table 4.14 and Table 4.15, respectively.

Concerning AM-PN where models are trained on AMI and tested on MULTISIMO, maximum accuracy of 69.78% is achieved with one-hot focus encoding using XGBoost (XGB) and no previous addressee. The results further show that algorithms trained without previous addressee outperform the algorithms trained using predicted previous addressee. The results are significant at $p < 0.05$ (p = 0.01 for shared focus encoding and p = 0.03 for one-hot encoding) for all algorithms.

Finally, for MA-PN, maximum accuracy of 49.13% is achieved with no previous addressee using one-hot focus encoding and XGboost (XGB) algorithm. For all the algorithms, experiments performed without using previous addressee in the feaure set achieve better results than the experiments performed using predicted previous address. The results however are only significant for one-hot encoding at $p < 0.05$ (p = 0.000021) and not for shared focus encoding (p = 0.06).

The discussions and the insights obtained from the results are presented in the next section.

Approach	Shared Focus	One-Hot Focus
XGB (PA)	48.55 ±0.023 (0.47)	48.47 ±0.023 (0.48)
ET (PA)	42.94±0.021 (0.42)	44.94 ±0.019 (0.44)
ADB (PA)	37.68 ±0.022 (0.38)	41.86 ±0.031 (0.41)
RF (PA)	39.44 ±0.013 (0.38)	44.64 ±0.024 (0.43)
XGB (NA)	**48.58** ±0.012 (0.47)	**49.13** ±0.024 (0.49)
ET (NA)	44.96 - ±0.023 (0.44)	45.58 ±43.01 (0.42)
ADB (NA)	41.51±0.021 (0.41)	42.49 ±0.026 (0.42)
RF (NA)	40.34 ±0.031 (0.39)	45.35 ±0.042 (0.45)

Table 4.15 – Accuracies for MA-PN (F1 in brackets)

4.7 Discussion & Perspectives

The results from experiments AA and MM show that both shared and one-hot focus encoding schemes in experiments AA and MM improve the existing baseline results for AMI and MULTISIMO, thus validating hypothesis h1 which states that the proposed model performs better than the baseline models for AMI and MULTISIMO.

Furthermore, the results from experiments AA and MM depict that for models trained and tested on a single dataset with ground-truth previous addressee values, shared focus encoding scheme significantly outperforms one-hot encoded focus scheme for both AMI and MULTISIMO, in the best case. For all algorithms, the results are significant for AMI dataset at $p < 0.05$ (p=0.005) but not significant for MPR (p = 0.13). The poor performance of one-hot encoding in case of self-learning can be attributed to the fact that in one-hot focus encoding there is a very high probability of information loss since the focus is being approximated,while shared focus encoding captures real focus ratios for all the participants and objects in the meeting.

Experiments AM and MA are performed to evaluate inter-corpora-learning capabilities of the models. The results show that one-hot encoded focus scheme outperforms the shared focus scheme. The reason for better performance of one-hot encoded focus for inter-corpora-learning can be attributed to the focus vector similarity between the focus of AMI and MULTISIMO as exemplified by figure 4.1 where although the shared focus vectors for AMI and MULTISIMO are different, the corresponding one-hot encoded focus vectors are similar. Thus, one-hot encoding reduces model over-fitting in case of inter-corpora-learning. The results are significant at $p < 0.05$ (p=0.008 for AMI, and p = 0.009 for MPR)

The results of experiments AA and AM show that in case of shared focus encoding, the maximum accuracy of 78.77% is achieved with experiment AA. This value is greater than the maximum accuracy of 75.55% with shared focus encoding for experiment AM

where models are trained on AMI containing 4 participants, and tested on MULTISIMO containing 3 participants. On the contrary, with one hot-focus encoding, the accuracy achieved with AM is higher compared to AA. Hence, hypothesis h2 which states that N should achieve at least similar or better classification performance when tested on a dataset with a number of participants equal or less than N, is only validated for shared focus encoding. The results is significant at $p < 0.05$ (p=0.000025)

The comparison of results for MM and MA shows that the opposite of hypothesis h2 (models trained with N participants should perform better or at least equal to when tested with more than N participants) is not true, hence it is important to mention that to get better results the algorithms should be trained on datasets with more participants than the test dataset.

Experiments AA-PN, MM-PN, AM-PN and MA-PN show results for the models trained with predicted previous addressee or no previous addressee. The results show that the accuracy decreases when the ground-truth previous addressee values are not used, which validates hypothesis h3-a. The results are significant for on AMI for both shared and one-hot focus encoding at $p < 0.05$ (p = 0.00045 for shared focus encoding on AMI, p = 0.0020 for one-hot focus encoding on AMI. For MPR, the results are only significant for one-hot focus encoding at $p < 0.05$ (p=0.00071) and not for shared focus encoding (p=0.21) for all algorithms.

The investigation of the models trained using ground-truth previous addressee and tested with predicted previous addressee, and trained and tested without previous addressee, reveals that models trained and tested without previous addressee perform better than the models with predicted previous addressee, which validates hypothesis h3-b. The possible explanation is that the error generated due to wrong predictions for the previous addressee propagates to all the next addressee predictions.

Among the algorithms, the ensemble learning algorithms, particularly XGBoost (XGB), achieve the highest accuracy for 6 out of 8 experiments (AA, AM, AA-PN, MM-PN, AM-PN, MA-PN). One can note that deep learning models perform poorly on both AMI and MULTISIMO. There can be two possible reasons: (i) there is not enough data to train deep learning classifier as the maximum accuracy achieved for MM with 634 records, using B-LSTM, is much lower (49.55%) compared to the accuracy for AA (60.90%) with 5,628 records. For all the rest of the algorithms, the accuracies achieved for MM are better compared to the accuracies for AA although there are less records; (ii) the addressee detection problem cannot be treated as a sequence problem. The second reason seems sustained by the results achieved for the experiments comparing predicted previous addressee and no previous addressee, since all the models with no previous addressee

performed better than the predicted previous addressee.

The overall results show that the proposed models are capable of addressee detection (requirement r3) on multiple datasets, with a varying number of participants (requirement r1), and with no dependence upon the positioning of the participants (requirement r2). Furthermore, a variation of the model is proposed that works in real-time without using ground-truth previous addressee (requirement r4).

The proposed models achieve accuracies of 78.77% on AMI dataset, and 84.80% on MULTISIMO dataset, which are better than the baseline accuracies for the respective datasets. Thus, the hypothesis h1, that *the models proposed in this research work should perform better than baselines on AMI and MULTISIMO* is validated. Hypothesis h2, that *models with participants N should achieve at least similar or better performance when tested on equal or less than N participants* is validated only for models trained with one-hot focus encoding. Finally at real-time, the performance of the proposed model is reduced, which validates hypothesis h3-a, that *Model performance decreases when the ground-truth value of previous addressee is not available*. In such cases, the hypothesis h3-b, that *the models trained and tested without previous addressee perform better than models tested with predicted previous addressee*, is also verified.

CHAPTER 5

TURN CHANGE AND NEXT SPEAKER PREDICTION

The previous chapter addressed the addressee detection problem in multiparty interaction. This chapter investigates the turn change and next speaker prediction problem using machine learning algorithms and smart feature selection. The proposed models show that the performance of the turn change prediction and next speaker prediction can be improved using psycholinguistic features such as pause duration [O'Connell and Kowal, 2012].

The chapter is divided into six sections. The first section contains a brief discussion of turn change management in conversations. Section 5.2 presents the related work. The process of feature selection is explained in section 5.3. Section 5.4 formalizes the problem and explains the experimental methodology while experiments and results are described in section 5.5. The chapter ends with a discussion in section 5.6.

5.1 Turn Change Mechanism in Conversations

An utterance addressed to an individual or a group of addressees, depending upon the dialogue act and the context, may or may not induce response from the listeners. For instance, a speaker can ask a question from one or multiple participants that prompts one of the participants to talk and reply. This change of speaker is referred to as turn change and the process of distribution of turns among the meeting participants is called turn management [Allwood, 2000].

In human-human interactions, turn management is mostly implicit. The interaction participants are expected to know when to speak, rather than explicitly being told to

talk. Furthermore, at each time the speaker can continue or can be interrupted by any of the listeners, hence there is no 'correct' next speaker and the 'real' next speaker is only known once a participant speaks in response to the current utterance. Humans learn from childhood how to smoothly manage dialogue turns. Speaker and listeners exploit language as well as various co-verbal and non-verbal signals such as pitch, gaze, head and hand gestures, to implicitly negotiate turn change [De Kok and Heylen, 2009; Guntakandla and Nielsen, 2015; Meshorer and Heeman, 2016]. Therefore, turn management is a multimodal process.

Human-agent interactions systems can be improved if the turn change and the speaker of the next turn can be correctly predicted. In multiparty context, turn change and next speaker prediction does not only allow the agent to understand when to contribute to an interaction, but also help the agent detect who should speak next resulting in appropriate behaviour generation for an agent. For instance, even if the agent does not intend to talk, it can turn toward the predicted next speaker before she says anything, demonstrating a human-like behaviour.

Though any of the speaker can talk at any time during a multiparty interaction, the process of turn taking can be broadly divided into three categories. Speakers signal that (i) they want to have turn when it is available, also known as turn taking, (ii) they are ready to accept turn when it is given to them by the previous speaker, which is known as turn accepting, and (iii) they want to take the turn even if the turn is not available which is called turn grabbing [Petukhova and Bunt, 2009].

The next section reviews existing works for turn change prediction and next speaker prediction.

5.2 Related Works for Turn Change & Next Speaker Prediction

Existing works for turn management fall in two categories. Works that study turn change prediction as an independent problem and works that combine turn change and next speaker prediction.

5.2.1 Models for Turn Change Prediction

One of the earliest turn management model is proposed by [Harvey Sacks and Jefferson, 1974]. The authors report that conversations proceed smoothly with only one people speaking at a time, and that sub-dialogues with multiple participants speaking simulta-

neously are short. They claim that there exists natural stages in a conversation, called Transition Relevance Places (TRPs), that mark the end of a turn and initiate a new turn. Relying on TRPs, Sacks *et al.* propose the following rules for turn management [Harvey Sacks and Jefferson, 1974]:

1. If the current speaker (S) selects the next speaker (N) in the turn, S is expected to stop speaking, and N to speak next.

2. If S does not select the next speaker, then any other participant may self-select and whoever speaks first gets the turn.

3. If no speaker self-selects, S may continue.

Though these rules are sufficiently generic to be applicable in various situations, they are not specific enough to explain (i) which interaction characteristics or features signal turn change, and (ii) which aspects of the current utterance are useful to predict the speaker of the next utterance in case of turn change. In addition, [Sacks et al., 1974] only propose a set of rules and TRPs for turn change. They did not evaluate the proposed rule based models on any test set and hence no objective results are available from the proposed model. Anyway, the work from [Sacks et al., 1974] opened door for research developing model for turn change prediction.

[Guntakandla and Nielsen, 2015] employ a J48 decision tree for turn prediction using n-grams of current and previous Dialogue Acts (DA) and current and previous speakers as features [Guntakandla and Nielsen, 2015]. The model is trained on Switchboard dataset [Godfrey et al., 1992]. They report an overall accuracy of 62.70% for turn change prediction.

[Meshorer and Heeman, 2016] propose a model that exploits random forests to predict turn changes. The model is also trained on Switchboard and uses current and previous DA, the *relative turn length* [1] and the *relative turn floor control* [2] to predict turn change. They report an accuracy of 76.05% and a F1 score of 0.74 for turn change prediction.

[Aldeneh et al., 2018] consider turn change as a sequence problem where turn change occurs in sequential manner. They propose an end-of-turn detection model that employs LSTM to learn end-of-turn prediction using speech features such as loudness, intensity,

[1] Two versions: (i) duration of a turn divided by the average speaker's turn duration, and (ii) number of words of a turn divided by the average number of words of the speaker's turns.

[2] Two versions: (i) total duration of the speaker's turns divided by the duration of the whole conversation, and (ii) total number of words the speaker's turns divided by the total number of words in the whole conversation.

zero-crossing rate. The proposed model is trained on Switchboard corpus. They report a F1 score of 0.65 for turn change prediction.

[De Kok and Heylen, 2009] propose a machine learning based model for end of turn prediction using 14 meetings from the AMI dataset[Carletta, 2007]. The proposed model uses DA, focus of attention, head gesture and prosody as features to train random conditional fields and hidden markov models. [De Kok and Heylen, 2009] report a F1 score of 0.61 for end of turn prediction.

5.2.2 Combined Models for Turn Change and Next Speaker Prediction

In addition to turn change, various researchers have proposed combined models for turn change and next speaker prediction. To this end, [Petukhova and Bunt, 2009] studied the importance of various multimodal signals such as gaze directions, verbal signals, lip movements and posture shift for the identification of next speaker. The proposed model is based on two task-based meetings from the AMI corpus. The approach simply finds the correlation between the different multimodal features and the turn types such as turn taking, turn grabbing and turn accepting. The videos of the meetings are shown to 15 end users who evaluate whether turn change occurred at a certain instance or not. The results depict that the gaze direction is the most important non-verbal for turn change prediction.

[Kawahara et al., 2012] proposes machine learning based turn change and next speaker prediction model that relies on a combination of gaze, prosody and head movements. The dataset used to train machine leaning models consist of 3 participants where one of the participants describe a poster while two other participants ask questions regarding this poster. They report an accuracy of 70.60% for turn change and 69.06% for next speaker prediction using Support Vector Machines (SVM).

[Ishii et al., 2013] propose a turn change and next speaker prediction model based on participants gaze transition patterns near the end of an utterance in multiparty interaction. A total of 12 gaze transition patterns have been used for predictions. They propose a Probabilistic model that depends on the relationship between different gaze transition patterns and the probability of turn change and next speaker selection. They achieve a F1 score 0.76 for turn change and an accuracy of 59.20% for next speaker prediction.

[Ishii et al., 2015a] use human gaze and respiratory behaviour for prediction of turn change and next speaker in multiparty interaction. They propose individual models

based on gaze and respiratory behaviour and a combined model which fuses gaze and respiration for the identification of the next speaker. The tests are performed on a custom dataset of 4 participants. Sequential minimal optimization (SMO) algorithm which is a variation of SVM is used to train the models. The results show that for turn change prediction, the model based on late fusion of eye gaze and respiratory behaviour yield a F1 score of 0.75. Similarly, for next speaker prediction the model based on the fusion of respiration and gaze behaviour result in a better performance (F1=0.52) compared to the models based individually on gaze (F1=0.45) or respiration (F1=0.47).

[Ishii et al., 2015b] propose a two-step machine learning based model that initially predicts whether or not a turn change occurs and in the next step predicts who is the next speaker. Both steps are based on head movement analysis such as the amplitude and frequency of the movement of head position of the speaker as well as for the listeners near the end of the utterance. The model is trained via SVM on a custom dataset of 4 participants. They achieve an accuracy of 75.00% for turn change prediction and an accuracy of 55.20% for next speaker prediction.

[Ishii et al., 2019] investigate the role of mouth-opening transition pattern (MOTP) which refers to change of mouth-opening degree at the end of the utterance and uses this information to predict the time interval between the two successive utterances, and the next speaker in multiparty interaction. They employ SVM for turn change and next speaker prediction using a custom dataset with 4 meeting participants. They report a F1 score of 0.80 for turn change prediction and a F1 score of 0.47 for next speaker prediction.

5.2.3 Summary and Discussion

A summary of existing works for turn change and next speaker prediction is presented in Table 5.1.

Most of the researchers have tried to predict turn change as a binary classification problem (turn change occurs or not). Both rule based as well as machine learning based approaches are used for turn change prediction. The existing accuracy for turn prediction vary between 50% and 80%. The most commonly used features for turn change prediction are DA, prosody, gaze information, head movements and speaker information.

Existing works for next speaker prediction are limited. Most of them use machine learning based approaches for next speaker prediction. The accuracies for next speaker prediction are significantly lower compared to accuracies achieved for turn change prediction. The reason can be that next speaker prediction is a multiclass classification problem which increases the chance of miss-classification compared to turn change prediction which is a binary classification task. Another reason could be the uncertainty

Reference	**Approach**	**Dataset**	**Salient Features**	**Result (Turn Change Prediction)**	**Result (Next Speaker Prediction)**
[Guntakandla and Nielsen, 2015]	Decision Tree (J48)	Switchboard	Current and previous DA, current and previous speakers	62.00% Accuracy	NA
[Meshorer and Heeman, 2016]	Random Forest	Switchboard	Current and previous DA, relative turn length, relative floor control	76.05% Accuracy	NA
[Aldeneh et al., 2018]	Single Dimensional LSTM	Switchboard	Speech features e.g loudness, intensity, zero-crossing rate	0.65 F1 Score	NA
[Petukhova and Bunt, 2009]	Correlation	AMI	Gaze, head movements, hand gestures	Correlations for various features.	Correlations for various features.
[De Kok and Heylen, 2009]	CRF & HMM	AMI	DA, focus of attention, head gestures, prosody	0.65 F1 Score	NA
[Kawahara et al., 2012]	Support Vector Machines	Custom Dataset	Gaze, prosody and head movements	70.00% Accuracy	69.06% Accuracy
[Ishii et al., 2013]	Probabilistic	Custom Dataset	Gaze transition patterns	0.76 F1 Score	0.76 F1 Score
[Ishii et al., 2015a]	Support Vector Machines	Custom Dataset	Gaze and respiratory behaviour	0.75 F1 Score	0.52 F1 Score
[Ishii et al., 2015b]	Support Vector Machines	Custom Dataset	Head movements such as the amplitude and frequency of the movement	75.00% Accuracy	55.20 % Accuracy
[Ishii et al., 2019]	Support Vector Machines	Custom Dataset	Mouth opening transition patterns	0.80 F1 Score	0.47 F1 Score

Table 5.1 – Summary of related works for turn change and next speaker prediction

of task since there can be multiple potential next speakers after a certain utterance since any of the meeting participants can speak at any time.

For both turn change and next speaker prediction, various features have been used by existing researchers. Surprisingly, to the best of our knowledge, the pause duration which

is the gap between two successive utterances is never exploited for turn change and next speaker prediction despite psycholinguistic experiments [Hilbrink et al., 2015; O'Connell and Kowal, 2012; Ten Bosch et al., 2004]. In the following, we propose machine learning based turn change and next speaker prediction models by selecting useful features as mentioned in the literature review and on the basis of the analysis of various datasets.

5.3 Feature Selection

The proposed turn change and next speaker prediction models learn from datasets via supervised machine learning techniques. Supervised machine learning techniques learn to find relationship between features and the output labels. Thus selecting the suitable features for turn change and next speaker prediction not only improves the model training time but also improves the model accuracy.

Section 2.7 contains details about various features used for different multimodal human-agent interaction tasks. In this chapter, feature list is compiled on the basis of relevance of features for turn change and next speaker prediction. To reaffirm the importance of the selected features for turn change and next speaker prediction, a brief analysis of some of the features is performed on the AMI and MPR datasets. The features that have never been used in the existing works but seem relevant after data analysis are also added in the feature set. The features selected for turn change and next speaker prediction are: speaker role, addressee role, start and end time of utterance, dialogue act (DA), gaze information, and pause duration.

5.3.1 Speaker Role

Speaker role refers to identity or name used to uniquely identify the current speaker during the interaction. Since every participant in a multiparty interaction can speak, each participant has a unique speaker role.

Work from [Guntakandla and Nielsen, 2015] highlights the importance of including current speaker for turn change prediction. Furthermore, analysis of the AMI dataset show that the speaker PM, being the moderator of the meeting is more likely to keep the turn compared to the other meeting participants. Similarly, in MPR, the robot NAO, being the moderator of games played in MPR, is more likely to keep turn than the remaining participants. In both of these examples, speaker role plays an important role for turn change prediction.

5.3.2 Addressee Role

Addressee role is another feature selected for training turn change and next speaker prediction models. Though there is no evidence from existing works that show the importance of addressee role for turn change and next speaker prediction. However, we propose to add addressee role in the feature set based on the data analysis of AMI and MPR datasets. The rationale is that if a speaker asks a question to a particular addressee, that addressee is more likely to take the turn and respond to the speaker. This hypothesis is substantiated by the analysis of the AMI and MPR datasets. In AMI, the addressee of the current utterance is the speaker of the next utterance in 47% of the utterances. If the DA of the current utterance involves a question, this percentage rises up to 65%. Similarly, for MPR, the addressee of the current utterance is next speaker for 58% of the time. The analysis of AMI also reveals that utterances addressed to individuals rather than groups are more likely to cause turn change.

5.3.3 Dialogue Act

DA is an important feature for turn change and next speaker prediction. For instance, a DA involving a question often prompts a turn change and leads (one of) the addressee(s) to answer. Similarly, DAs such as stalling express that the speaker has not yet finished talking and wants to keep the turn. The importance of DA for the two tasks is highlighted in several existing works [Petukhova and Bunt, 2009], [De Kok and Heylen, 2009], [Guntakandla and Nielsen, 2015], [Meshorer and Heeman, 2016].

The analysis of AMI and MPR datasets show that some dialogue acts such as *feedback elicitation, thanking, answering* are more likely to cause turn change (100%, 93 %, and 86.78%, respectively) as compared to *time-management* DA which on average cause turn change in 45% of utterances.

5.3.4 Start and End Time of Utterance

The start and end time of an utterance can also play an important role for making various decision in human-agent interaction [Meshorer and Heeman, 2016]. The analysis of the AMI and MPR datasets show that the utterances at the beginning of a conversation are less likely to have turn change compared to utterances at later stages of a meeting which shows that start and end time of utterance can also be used as a separate feature, in addition to calculating utterance duration.

5.3.5 Pause Duration

Psycho-linguistic evidence highlights a strong correlation between pause duration and turn change [O'Connell and Kowal, 2012], [Hilbrink et al., 2015]. A quantitative analysis [O'Connell and Kowal, 2012] of speaker changes reports that 91% of the speaker changes occur with a pause between two utterances whereas only 8% of the turn changes occur with no pause, and 1% of the utterances are overlapped by another one. Furthermore, Takeuchi *et al.* show that pause duration can be employed for predicting response time for stalling or turn taking in dyadic interactions [Takeuchi et al., 2003]. Since robots speak after a certain response time, which corresponds to turn change in dyadic interaction, pause duration can be used for turn change and next speaker prediction in multiparty interactions as well.

The analysis of MPR and AMI dataset shows that on average, in case of turn change, pause duration is approximately three times lesser compared to the cases when the current speaker retains the turn.

5.3.6 Focus of Attention

The importance of gaze or focus of attention as a marker for turn change and next speaker prediction has been investigated by several researchers [Ishii et al., 2015a], [Ishii et al., 2013], [Kawahara et al., 2012], [De Kok and Heylen, 2009], [Petukhova and Bunt, 2009]. The results from these research works show that focus of attention is fundamental to turn change and next speaker prediction. Particularly, the participant in focus near the end of the utterance of the current speaker is often the next speaker.

5.3.7 Discussion

Several works predict turn change and next speaker with machine learning techniques as summarized in Table 5.1. Despite psycholinguistic evidences, to the best of our knowledge none of the existing works have exploited pause duration as a feature in a machine learning model for turn change and next speaker prediction in multiparty interaction. A reason can be that pause duration is a dynamic attribute: pause duration between two utterances cannot be calculated before the start of the next utterance. A solution would be to regularly evaluate pause duration until it exceeds a certain threshold that triggers the prediction of a turn change. This process is similar to human evaluation: if the speaker continues to speak immediately after an utterance, listeners often do not take the turn. On the other hand, if the speaker pauses for a long time, listeners are more likely to take the turn [O'Connell and Kowal, 2012].

Addressee role is another feature that, to the best of our knowledge, has not been exploited yet for turn change and next speaker prediction. This feature has been intuitively selected based on the analysis of AMI and MPR datasets that reveal that the utterances addressed to individuals are more likely to evoke a response causing turn change, compared to the utterances addressed to a group. Furthermore, the addressee of the current utterance is often the next speaker.

5.4 Problem Formalization and Methodology

The task of turn change and next speaker is divided into two sub-tasks: turn change prediction and next speaker prediction. Two approaches are proposed to solve these task. In the first approach, two independent models for turn change and next speaker prediction are proposed. In the second approach, a connected model for turn change and next speaker prediction is proposed where turn change is predicted in the first step and then the next speaker is predicted by including turn change in the feature set.

5.4.1 Turn Change Prediction and Next Speaker Prediction Considered as Independent Problems

Given a set of features, the first task consists in predicting, whether the speaker of the current utterance and the next utterance are different or not. If the speaker of the current and the next utterances are different, it is assumed that turn change has occurred, else turn change did not occur. Turn change prediction is a binary classification problem since there are only two possible outputs.

The second task performed is to predict who is the speaker of the next utterance. Next speaker prediction is a multiclass classification problem since there are more than two possible outputs in multiparty interaction. For instance in AMI, the next speaker can be any of the four participants. Similarly, in the MPR , the next speaker can be one of the interaction participants A, B, C, or the robot NAO.

An utterance can consist of different DAs and vice versa a DA can span various utterances. However for the sake of simplification, in this research work, it is assumed that an utterance consists of a single DA and thus both turn change and next speaker are predicted after every DA.

Figure 5.1 shows the experimental methodology considering turn change and next speaker prediction as independent models. Input features, divided into contextual features, focus features, and textual and speech features, are fused together to form a

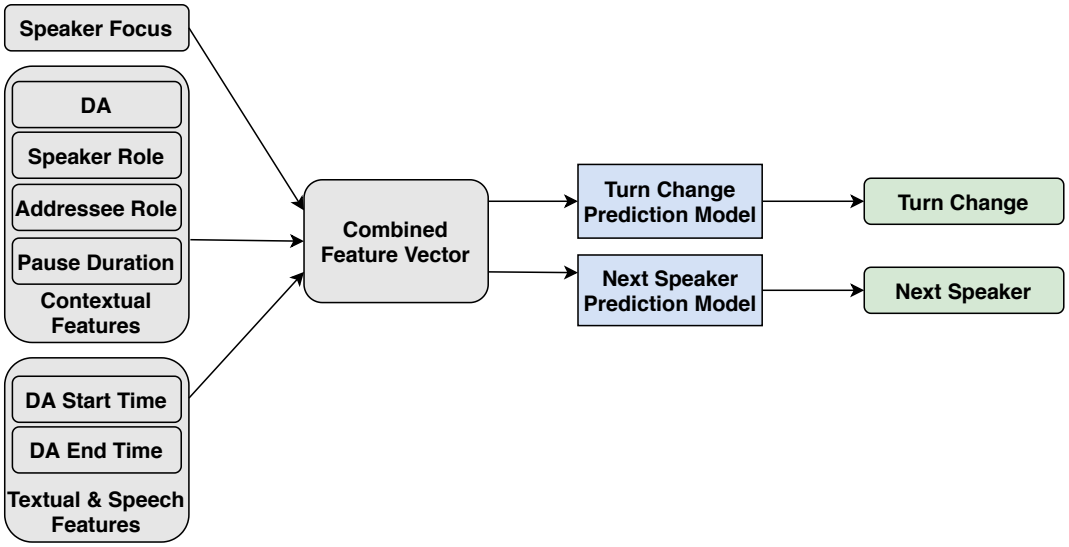

Figure 5.1 – Independent turn change and next speaker prediction models

combined feature vector that is used to train both models. The features used are divided into three groups: contextual features, focus features, and textual features. Contextual features are DA, speaker role and addressee roles, and pause duration, focus features contains the focus information of the speaker, whereas textual or speech features include the start and end times of a DA. The turn change prediction model outputs a binary value (a turn change occurs / no turn change). The next speaker prediction model predicts the speaker for the next utterance among the meeting participants.

5.4.2 Combining Turn Change and Next Speaker Prediction

Turn change and next speaker prediction are two related tasks as next speaker prediction depends on turn change: a turn change signals that the speaker of the next utterance cannot be the current speaker and someone different from the current speaker has to speak next. Thus, predicted turn change can also be used as an additional feature.

Figure 5.2 depicts the experimental methodology to combine turn change and next speaker predictions. The next speaker is predicted in two steps: 1) the input features are used to predict turn change similarly to Figure 5.1; 2) the predicted turn change value is then exploited as additional input feature, in order to predict the next speaker. The real turn change values are used to train the model whereas at run time the predicted turn change value is exploited as additional feature.

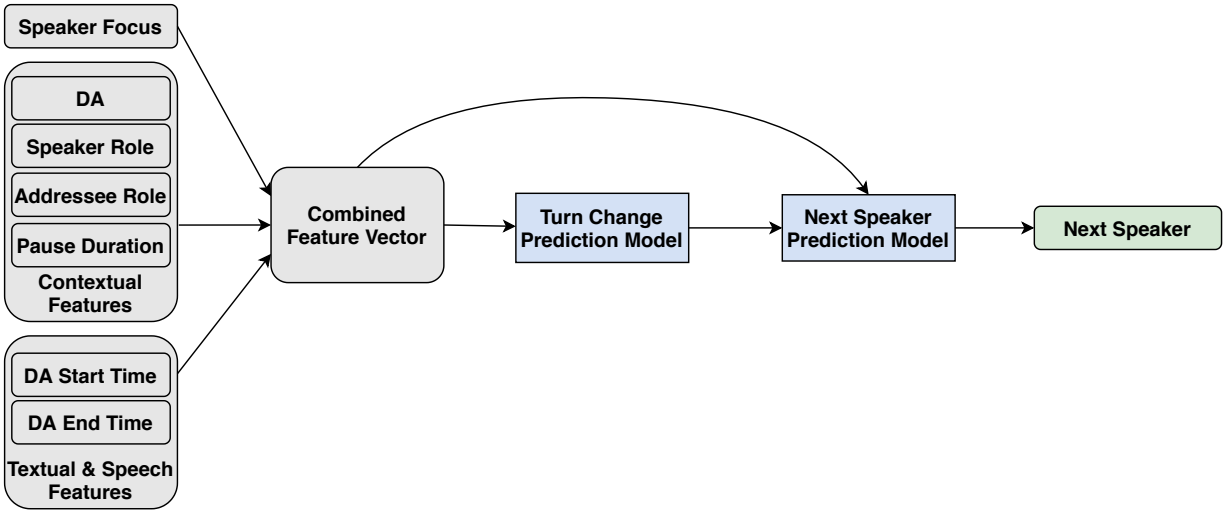

Figure 5.2 – Combined turn change and next speaker prediction model

5.5 Experiments and Results

This section describes the datasets, the procedure followed to perform the experiments and the results obtained.

5.5.1 Datasets

The datasets used to train and test the turn change and next speaker prediction models are AMI [McCowan et al., 2005] and MPR [Funakoshi, 2018]: they are the only corpora that contain annotated data for all the features selected in Section. 5.3. The details of the AMI and MPR dataset has been presented in Section 2.6.2.

5.5.2 Baselines

To evaluate our turn change prediction models, two baselines are selected: (i) [Meshorer and Heeman, 2016], and (ii) majority class for turn change. The first baseline is selected because (a) it returns the highest performance on Switchboard and (b) its features are available in most of the existing datasets and therefore results can be reproduced.

To evaluate our next speaker prediction model, the majority class for next speaker prediction is chosen as baseline. Indeed, the features proposed are not available in any of the datasets exploited by existing research works, nor it is possible to easily reproduce their results on the commonly used datasets (i.e. Switchboard and AMI) as they are funded on unavailable features. Hence, unfortunately, it is not possible to compare our proposed model with any of the existing models as baseline for next speaker prediction.

5.5.3 Experiments

Three sets of experiments are performed : (i) experiments performed to evaluate performance of turn change and next speaker prediction models individually, as mentioned in Section 5.4.1, (ii) experiments performed for next speaker prediction using predicted turn change explained in Section 5.4.2, and (iii) experiments performed for ablation study in order to evaluate the importance of some of the features.

5.5.3.1 Experiments for Individual Turn Change and Next Speaker Prediction Models

Two separate sets of experiments are performed: one for turn change prediction model, and the other for next speaker prediction model.

A conventional machine learning pipeline is followed to carry out the experiments. The categorical features are one-hot encoded to convert them into a numerical form. The feature set is normalized using standard scaling. Training and test sets are created via five fold cross validation.

Eight of the most classic machine learning classifiers have been trained and tested: XGboost (XGB) [Chen and Guestrin, 2016], Multilayer Perceptron (MLP) [Kruse et al., 2013], Random Forest (RF) [Liaw et al., 2002], Logistic Regression (LR) [Hosmer Jr et al., 2013], Support Vector Machines (SVM) [Hearst et al., 1998], Naive Bayes (NB) [Rish et al., 2001] and K-Nearest Neighbours (KNN) [Zhang and Zhou, 2005]. In addition, to evaluate if turn change and next speaker prediction problems can be considered as sequential, tests are also performed using a combination of Long Short Term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997] and Densely Connected Neural network (DNN) in a multilayer perceptron.

For all the classifiers, default parameters as specified in Python's Sklearn library [Pedregosa et al., 2011] are used. Finally, accuracy and F1 measure have been employed to evaluate performances. Accuracy is used to compare the results with baselines and F1 measure is considered since the class distribution is irregular in both datasets [Pedregosa et al., 2011].

5.5.3.2 Experiments for Model Combining Turn Change and Next Speaker Prediction

Concerning the experiments performed to evaluate the model combining turn change and next speaker prediction, the predicted turn change is included in the feature set in order to predict the next speaker as shown in Figure 5.2. The values for turn change

prediction are obtained with the model that yields best result for turn change prediction during the experiments presented in Section 5.5.3.1. The machine learning pipeline, the algorithms and the evaluation metrics are similar to those of Section 5.5.3.1.

5.5.3.3 Ablation study

In the proposed models, two new features (*i.e.* pause duration and addressee role) are added. To the best of our knowledge, these features have not been previously exploited for turn change and next speaker prediction. Experiments are performed with and without pause duration and addressee roles to evaluate if these features actually significantly improve turn change and next speaker predictions.

Since the importance of different features is studied individually on turn change and next speaker prediction models, two sets of experiments are performed: (i) Turn change prediction, (ii) and Next speaker prediction.

The naming convention for the experiments follows the pattern *dataset-task-experiment*, where: dataset refers to ami or mpr; task refers to tc (for turn change prediction) and ns (for next speaker prediction); experiment refers to one of the 4 different experiment types *i.e.* -ar-pd, +ar-pd, -ar+pd, and +ar+pd.

In experiments -ar-pd, the models are trained using the feature set presented in Section 5.3 without addressee role and pause duration. In experiments +ar-pd, addressee role is added to the feature set of experiments -ar-pd. In experiments -ar+pd pause duration is added to the feature set of -ar-pd. Finally in experiments +ar+pd, the whole feature set is used, including addressee role and pause duration. To find statistical significance, t-tests are performed between pairs of experiments (-ar-pd, +ar-pd), (-ar-pd, -ar+pd), and (-ar-pd, +ar+pd).

The machine learning pipeline, the algorithms and the evaluation metrics for the ablation experiments are the same as the ones used for experiments in Section 5.5.3.1.

5.5.4 Results

The results are divided into three sections: (i) results for experiments for individual turn Change and next speaker prediction models, (ii) results for next speaker prediction model using predicted turn change, and (iii) ablation study. The results show the average values obtained for five folds cross validation and depict the accuracy and standard deviation and F1 measure (mentioned in brackets).

5.5.4.1 Results for Individual Turn Change and Next Speaker Prediction Models

The results from Table 5.2 show that for both the MPR and AMI datasets, the proposed turn change model perform better than both the baselines. A maximum average accuracy of 85.83% is obtained on the AMI using XGboost algorithm which is better than 61.57% obtained by the baseline 1, and 54.55% obtained by the baseline 2. On the MPR, the proposed model achieved a maximum accuracy of 83.08% which is better than baseline 1 (76.89%) and baseline 2 (74.37%)

T-tests are performed to find the statistical significance between the results from proposed models and baselines 1 and 2 on both the AMI and MPR. For baseline 1, the results are significant at $p < 0.05$ (p=0.000018 on AMI and p=0.000056 on MPR). For baseline 2, the results are also significant at $p < 0.05$ (p=0.000011 on AMI and p=0.0000078 on MPR).

Algorithtm	**MPR**	**AMI**
XGB	**83.08 ±0.62 (0.82)**	**85.83 ±1.10 (0.85)**
RF	82.13 ±1.09 (0.81)	84.95 ±1.59 (0.84)
MLP	79.53 ±1.16 (0.79)	72.47 ±1.64 (0.72)
SVM	81.20 ±0.89 (0.79)	65.78 ±2.76 (0.65)
KNN	80.91 ±1.10 (0.80)	60.20 ±2.71 (0.60)
LR	78.41 ±0.99 (0.75)	64.13 ±3.39 (0.64)
NB	26.69 ±0.88 (0.13)	62.04 ±2.47 (0.61)
LSTM+DNN	78.21 ±1.34 (0.78)	65.66 ±2.41 (0.65)
Baseline 1 ([Meshorer and Heeman, 2016])	**76.89 ±1.05 (0.76)**	**61.57±1.89 (0.61)**
Baseline 2 (Majority Class)	**74.37±0.76 (0.74)**	**54.55 ±2.73 (0.54)**

Table 5.2 – Results for turn change prediction for MPR and AMI datasets. Results are in %, F1 values in brackets.

Table 5.3 depicts results for the experiments performed for next speaker prediction using the AMI and MPR datasets. The results show that for the AMI the best case accuracy of 64.77% is achieved via the XGB algorithm which is better than the baseline accuracy if 35.77%. Similarly, for MPR a maximum accuracy of 64.73% is achieved via the XGB algorithm, which outperforms the baseline accuracy of 38.75%.

T-tests are performed to find the statistical significance between the proposed model and the baseline for both the AMI and MPR datasets. The results obtained are significant at $p < 0.05$ (p=0.00038 on AMI and p=0.0000082 on MPR).

Algorithm	**MPR**	**AMI**
XGB	**64.73 ±1.67 (0.64)**	**64.77 ±3.55 (0.64)**
RF	62.62 ±1.00(0.62)	63.02 ±3.14(0.63)
MLP	59.71 ±0.54(0.59)	45.92 ±4.48(0.46)
SVM	61.69 ±1.50(0.60)	51.71 ±3.57(0.51)
KNN	60.79 ±1.10(0.60)	46.99 ±2.41(0.47)
LR	58.65 ±1.68(0.58)	51.04 ±3.30(0.50)
NB	25.17 ±2.70(0.16)	25.16 ±2.93(0.14)
LSTM + DNN	60.02 ±0.70(0.59)	44.52 ±3.54(0.45)
Baseline (Majority Class)	**38.75 ±1.80 (0.38)**	**35.77 ± 6.29 (0.35)**

Table 5.3 – Results for next speaker prediction for AMI and MPR datasets. Results are in %, F1 values in brackets.

5.5.4.2 Results for Model Combining Turn Change and Next Speaker Prediction

Table 5.4 depicts the results for the combined turn change and next speaker prediction model, where predicted turn change is included in the feature set of the next speaker prediction model. The results show that for AMI dataset, a maximum accuracy of 65.53%, and a F1 value of 0.65 is obtained via the XGB algorithm. This value is greater than the value achieved (64.77% and F1= 0.64) when next speaker prediction model is considered as an individual model as shown in Table 5.3.

Concerning MPR, a maximum accuracy of 64.46% and a F1 value of 0.63 are obtained using the XGB algorithm, which is almost similar to the values (64.73% and F1=0.64) obtained via individual next speaker prediction model.

The mutual comparison of individual and combined next speaker prediction models show that the performance difference between the two models is not significant at $p < 0.05$ (p=0.16 on AMI and p=0.30 on MPR).

For both individual and combined turn change and next speaker prediction models, the performance of LSTM+DNN model is quite similar to the MLP algorithm which is a DNN. The result shows that LSTM does not improve the accuracy of turn change and next speaker prediction tasks which may point to the non-sequential nature of both turn change and next speaker prediction tasks.

5.5.4.3 Results of the Ablation Study

Results for experiments performed for ablation study are mentioned in Tables 5.5 to 5.8.

Table 5.5 and Table 5.6, shows the results obtained during the ablation study performed for turn change prediction, respectively on AMI and MPR datasets. The results

Algorithm	**MPR**	**AMI**
XGB	**64.46 ±1.64 (0.63)**	**65.53 ± 2.67 (0.65)**
RF	62.62 ±1.01(0.62)	64.07 ±2.16(0.64)
MLP	59.93 ±2.57(0.58)	50.36 ±3.17(0.50)
SVM	61.06 ±1.30(0.60)	49.60 ±3.53(0.49)
KNN	61.11 ±0.96(0.60)	45.57 ±2.41(0.45)
LR	59.62 ±2.84(0.58)	50.57 ±3.46(0.50)
NB	25.18 ±2.68(0.16)	25.11 ±2.93(0.14)
LSTM + DNN	60.10 ±1.89(0.59)	51.21 ±2.81(0.50)
Baseline 1 (Majority Class)	**38.75 ±1.80 (0.38)**	**35.77 ± 6.29 (0.35)**

Table 5.4 – Results for Model Combining Turn Change and Next Speaker Prediction (accuracies in %, F1 values in brackets).

Algorithm	**ami-tc-ar-pd**	**ami-tc+ar-pd**	**ami-tc-ar+pd**	**ami-tc+ar+pd**
XGB	65.85 ±1.10(0.66)	66.84 ±1.15(0.66)	**85.21 ±1.19 (0.85)**	**85.83 ±1.10 (0.85)**
RF	62.11 ±1.25(0.62)	64.15 ±1.24(0.65)	84.51 ±1.67(0.83)	**84.95 ±1.59 (0.84)**
MLP	62.17 ±1.14(0.62)	63.25 ±2.21(0.63)	**74.31 ±1.39 (0.74)**	72.47 ±1.64(0.72)
SVM	66.53 ±2.13(0.67)	66.81 ±2.40(0.67)	64.35 ±2.40(0.64)	**66.98 ±2.76 (0.66)**
KNN	60.06 ±2.15(0.65)	60.13 ±2.13(0.66)	**61.21 ±2.47 (0.62)**	60.20 ±2.71(0.60)
LR	63.40 ±2.84(0.63)	63.64±2.36(0.64)	63.18 ±3.10(0.63)	**64.13 ±3.39 (0.64)**
NB	62.11 ±2.18(0.62)	62.11 ±3.10(0.62)	**62.17 ±2.98 (0.62)**	62.04 ±2.47(0.61)
LSTM+DNN	62.10 ±1.98(0.62)	62.15 ±2.65(0.62)	**66.24 ±1.89 (0.66)**	65.66 ±2.41(0.65)

Table 5.5 – Results for turn change prediction for AMI. Results are in %, F1 values in brackets.

show that for both AMI and MPR, in the best case (using the XGB algorithm) the models trained using both addressee role and pause duration (Table 5.5: ami-tc+ar+pd, Table 5.6: mpr-tc+ar+pd), outperform the models trained without these features (Table 5.5: ami-tc-ar-pd, Table 5.6: mpr-tc-ar-pd) and the models trained including only one of these features in the feature set (Table 5.5: ami-tc+ar-pd, and ami-tc-ar+pd, Table 5.6: mpr-tc+ar-pd, and mpr-tc-ar+pd). For both AMI and MPR, the best results are obtained via the XGB algorithm. These results are significant at $p < 0.05$ (p=0.033 on AMI and p=0.001 on MPR).

The importance of addressee role and pause duration has been individually studied concerning turn change prediction. For both AMI and MPR, the best case performance obtained via XGB by including addressee role (Table 5.5: ami-tc+ar-pd, Table 5.6: mpr-tc+ar-pd) are slightly higher or similar to the models that do not use addressee role as a feature (Table 5.5: ami-tc-ar-pd, Table 5.6: mpr-tc-ar-pd). The results are significant at $p < 0.05$ (p=0.02 on AMI and p=0.007 on MPR) when all the algorithms are compared.

The results of experiments ami-tc-ar+pd and mpr-tc-ar+pd further show that the

Algorithm	**mpr-tc-ar-pd**	**mpr-tc+ar-pd**	**mpr-tc-ar+pd**	**mpr-tc+ar+pd**
XGB	76.28 ±0.74(0.76)	76.34 ±1.13(0.77)	82.43 ±0.84(0.82)	**83.08 ±0.62 (0.82)**
RF	69.59 ±0.89(0.70)	70.15 ±1.20(0.70)	81.45 ±1.10(0.81)	**82.13 ±1.09 (0.81)**
MLP	74.47 ±1.45(0.74)	74.59 ±1.92(0.74)	79.27 ±1.41(0.79)	**79.53 ±1.16 (0.79)**
SVM	76.17 ±2.10(0.76)	76.39 ±2.10(0.76)	**81.44 ±1.32 (0.81)**	81.20 ±0.89(0.79)
KNN	74.02 ±1.91(0.74)	74.87 ±2.31(0.75)	**81.22 ±1.74 (0.81)**	80.91 ±1.10(0.80)
LR	76.31 ±2.30(0.76)	76.44 ±1.60(0.76)	77.24 ±0.82(0.77)	**78.41 ±0.99 (0.75)**
NB	25.86 ±1.82(0.25)	26.00 ±1.76(0.26)	25.86 ±1.43(0.25)	**26.69 ±0.88 (0.13)**
LSTM+DNN	73.04 ±1.20(0.73)	73.51 ±0.95(0.73)	**80.19 ±1.57 (0.80)**	78.21 ±1.34(0.78)

Table 5.6 – Results for turn change prediction for MPR. Results are in %, F1 values in brackets.

best case results are obtained via the XGB algorithm which implies that adding pause duration as individual feature outperforms the best case results obtained when pause duration is not included in the feature set (Table 5.5: ami-tc-ar-pd, Table 5.6: mpr-tc-ar-pd). In addition, adding pause duration individually to the features set significantly improve the results at $p < 0.05$ (p=0.037 on AMI and p=0.00023 on MPR) when all the algorithms are compared.

The Table 5.7 and Table 5.8 depicts the results of the ablation study performed for next speaker prediction using the AMI and MPR datasets, respectively. The results show that for both AMI and MPR, the models trained using jointly addressee role and pause duration (Table 5.7: ami-ns+ar+pd, Table 5.8: mpr-ns+ar+pd), outperform the models trained without these features (Table 5.7: ami-ns-ar-pd, Table 5.8: mpr-ns-ar-pd) and the models trained including only one of these features (Table 5.7: ami-ns+ar-pd, and ami-ns-ar+pd, Table 5.8: mpr-ns+ar-pd, and mpr-ns-ar+pd). For both AMI and MPR, the best results are obtained via the XGB algorithm. These results are significant at $p < 0.05$ (p=0.03 on AMI and p=0.000056 on MPR).

Finally, the importance of addressee role and pause duration has been individually tested concerning next speaker prediction. For both AMI and MPR, the best case performance obtained via XGB by including addressee role (Table 5.7: ami-ns+ar-pd, Table 5.8: mpr-ns+ar-pd) outperforms the models without addressee role (Table 5.7: ami-ns-ar-pd, Table 5.8: mpr-ns-ar-pd). The results are significant at $p < 0.05$ (p=0.04 on AMI and p=0.000018 on MPR).

The results from experiments ami-tc-ar+pd and mpr-tc-ar+pd further show that best case results obtained via XGB by adding pause duration as an individual outperform the results obtained without pause duration (Table 5.7: ami-ns-ar-pd, Table 5.8: mpr-ns-ar-pd). The the performance difference is significant at $p < 0.05$ (p=0.0006 on AMI and p=0.00085 on MPR).

Algorithm	ami-ns-ar-pd	ami-ns+ar-pd	ami-ns-ar+pd	ami-ns+ar+pd
XGB	45.14 ±2.58(0.45)	46.62 ±2.47(0.47)	61.77 ±3.10(0.62)	**64.77 ±3.55 (0.64)**
RF	39.99 ±1.65(0.40)	40.61 ±1.90(0.41)	60.49 ±2.20(0.61)	**63.02 ±3.14 (0.63)**
MLP	40.95 ±1.96(0.41)	41.12 ±1.46(0.41)	**47.77 ±1.90 (0.48)**	45.92 ±3.20(0.46)
SVM	45.07 ±2.20(0.45)	48.39 ±1.93(0.46)	49.50 ±3.23(0.48)	**51.71 ±3.57 (0.51)**
KNN	40.73 ±1.86(0.40)	47.82 ±2.26(0.42)	46.19 ±2.25(0.47)	**46.99 ±2.41 (0.47)**
LR	42.94 ±1.60(0.42)	42.56 ±1.25(0.43)	43.94 ±1.50(0.45)	**51.04 ±3.30 (0.50)**
NB	35.07 ±2.20(0.35)	35.10 ±1.32(0.35)	**36.80 ±2.50(0.36)**	25.16 ±2.93(0.14)
LSTM+DNN	40.86 ±1.71(0.41)	42.31 ±1.69(0.42)	**48.50 ±3.40 (0.47)**	44.52 ±3.54(0.45)

Table 5.7 – Results for next speaker prediction for AMI. Results are in %, F1 values in brackets.

Algorithm	mpr-ns-ar-pd	mpr-ns+ar-pd	mpr-ns-ar+pd	mpr-ns+ar+pd
XGB	48.70 ±0.56(0.48)	56.89 ±1.25(0.57)	56.89 ±1.72(0.57)	**64.73 ±1.67 (0.64)**
RF	41.69 ±0.79(0.41)	48.18 ±1.32(0.48)	56.08 ±1.90(0.58)	**62.62 ±1.00 (0.62)**
MLP	47.82 ±0.62(0.48)	53.83 ±1.15(0.54)	55.69 ±0.67(0.56)	**59.71 ±0.54 (0.59)**
SVM	46.74 ±0.91(0.46)	56.57 ±0.87(0.57)	53.45 ±1.40(0.54)	**61.69 ±1.50 (0.60)**
KNN	45.79 ±1.20(0.46)	53.08 ±1.21(0.53)	54.64 ±1.65(0.54)	**60.79 ±1.10 (0.60)**
LR	45.05 ±1.71(0.45)	55.01 ±1.15(0.55)	49.13 ±1.21(0.49)	**58.65 ±1.68 (0.58))**
NB	21.84 ±1.95(0.22)	34.97 ±1.42(0.35)	21.84 ±2.56(0.21)	**25.17 ±2.70 (0.16)**
LSTM+DNN	47.24 ±0.89(0.47)	52.74 ±0.90(0.53)	54.26 ±1.45(0.54)	**60.02 ± 0.70 (0.59)**

Table 5.8 – Results for next speaker prediction for MPR. Results in %, F1 values in brackets.

5.6 Discussion & Perspectives

Results depict that both individual and combined turn change and next speaker prediction models perform significantly better than baselines on AMI and MPR.

In addition, the results from the ablation study shows that both pause duration and addressee role are important features for turn change and next speaker prediction. The performance difference for both the AMI and MPR is significant when turn change and next speaker prediction are used as additional features as compared to the results obtained via models trained without these parameters.

The results show that using predicted turn change as additional feature improves the accuracy of next speaker prediction in the AMI . However the same result is not obtained for MPR , where the accuracy achieved without using predicted turn change as additional feature is similar to when predicted turn change is used as additional feature. One of the reason could be that for AMI, the maximum accuracy of predicted turn change is slightly higher as compared to the predicted turn change accuracy of MPR. Hence in case of AMI the ratio of propagated error from turn change to next speaker prediction is

less compared to MPR where higher error value is propagated to next speaker prediction task. This behaviour also confirms that correctly predicted turn change can actually improve the performance of next speaker prediction task.

The proposed turn change and next speaker prediction models use a static value for pause duration since while training and testing the model, the value for pause duration can be calculated by subtracting the end time of current utterance from the start time of the second utterance. For real time implementation, the static value of pause duration is not available since the start time of the next utterance is not known at the end of current utterance. Similarly, the focus value is not static and keeps on changing even after the utterance.

One way to solve these implementation problems is to start a thread as soon as an utterance ends. The thread monitors the time elapsed since the last ended and keeps track of the gaze information. The turn change and next speaker prediction models are then called iteratively after a certain time period. The dynamic values for time passed since the last utterance and the current focus of speaker are transmitted to the machine learning model to makes a series of decisions depending upon the number of iteration. As soon as the model predicts that a turn change has occurred the thread stops. The process continues at the end of each utterance. Even if this is one possible solution, a thread that executes iteratively may slow down the real-time response of turn change and next speaker prediction model. Hence a better optimization technique is required to dynamically call turn change and next predicting model with updated pause duration and speaker focus.

The proposed turn change and next prediction models can be very useful in real-time multiparty human-computer interaction. For instance, the turn change model informs an interaction agent whether or not the current speaker has finished speaking, allowing him to update its internal knowledge state about the discussion. Furthermore, the next speaker prediction model can help an agent decide if she is expected to speak or not at a particular time. In addition, the information about next speaker can help an agent adjust its behaviour generation. For instance, an agent can change its gaze direction towards the next potential speaker even before the speaker speaks.

Even if the proposed turn change and next speaker prediction models outperform existing baselines, there is still room for improvement. For instance, in addition to the features used in the proposed models, existing works show that prosody, head and hand gestures can also be exploited. Thus, adding these features can further improve the performance of turn change and next speaker prediction models. Currently, none of the existing datasets contains all these features along with the features proposed in this

research work. A solution could be to either develop new datasets or to annotate the existing AMI and MPR with these additional features.

Turn change and next speaker prediction models form the second component of this research work. The next chapter explains the third and final component we propose i.e. the Visual Focus of Attention (VFOA) behaviour generation model.

CHAPTER 6

VISUAL FOCUS OF ATTENTION BEHAVIOUR GENERATION MODEL

In typical human-agent interaction, an agent performs three tasks: it perceives the interaction environment including, makes decisions based on perceptions, and may or may not generate behaviours based on these decisions. In the previous chapters, a perception model for addressee detection, and decision making model for turn change and next speaker prediction are proposed. The research work proposed in this chapter concerns a behaviour generation component for intelligent agents.

In this regard, this chapter proposes a Visual focus of Attention (VFOA) behaviour generation model when an intelligent agent acts as a speaker and as listener in multiparty interaction. Machine learning as well as heuristic approaches are exploited to develop of different types of sub-models that contribute to predict the overall VFOA for an artificial companion at any of time.

This chapter is divided into 8 sections. Section 6.1 presents a general overview of VFOA behaviour in conversations. Section 6.2 discusses the related work while feature selection process is explained in section 6.3. Section 6.4 formalizes the problem. Proposed VFOA Generation models are discussed in Section 6.5. Experiments and results are presented respectively in Sections 6.6 and 6.7, while section 6.8 presents discussion and perspectives.

6.1 Visual Focus of Attention in Conversations

While verbal communication remains the primary mode of interaction between humans, nonverbal behaviours such as gestures [McNeill, 1992] and visual focus of attention (VFOA) [Argyle and Ingham, 1972] can improve and reinforce verbal communication to convey mental states. VFOA refers to the targets, i.e. the companion participants or objects, that are in focus of attention of a participant during an interaction. VFOA is a nonverbal behaviour that can be generated by identifying at any point of time during an interaction where a person is looking at.

The importance of VFOA for seamless human interaction has led researchers to replicate the VFOA behaviour in Human-Agent Interaction (HAI). The hypothesis is that for natural HAI, an agent should be able to exhibit VFOA behaviours similar to humans. One of the earliest researches for developing a VFOA model for HAI was initiated by the virtual agent community in the 1990s [Thórisson, 1994], [Cassel et al., 1998]. Meaningful gaze movements were also integrated in different robots such as Cog [Scassellati, 1996], Kismet [Breazeal and Scassellati, 1999], and Infanoid [Kozima and Ito, 1998].

Recent approaches to modeling VFOA behaviour come from cross disciplinary work spanning research areas from robotics, artificial intelligence, virtual agents and psychology. According to [Admoni and Scassellati, 2017], the techniques for generating VFOA behaviour for HAI can be broadly classified into three categories: (i) Biologically Inspired Systems, (ii) Data Driven Systems, and (iii) Heuristics Systems. Biologically inspired VFOA models stem from biological and cognitive mechanisms that control human VFOA behaviour. Systems in this category mimic the neurological functions of the human brain. Data driven systems are based on quantified observations of human interactions to train both rule-based and machine learning based VFOA models. Heuristic approaches model the appropriate VFOA behaviour without relying on any biological functions and human interaction corpora.

VFOA models for dyadic interaction are simpler compared to multiparty interaction. Consequently, there has been very limited work on the development of VFOA models for multiparty interaction [Vertegaal et al., 2000], [Vertegaal et al., 2001]. In this Chapter, a bottom-up, hybrid approach is proposed for developing VFOA models in multiparty interaction for speakers as well as listeners. The proposed VFOA behaviour generation models are divided into sub-models where each model solves a specific problem. The models are based on machine learning as well as heuristics. The combined functionality of the models generates the overall VFOA behaviour for an artificial companion at any given time.

6.2 Related Works

This section briefly reviews some of the existing visual focus of attention (VFOA) behaviour generation models for dyadic as well as multiparty interactions. As mentioned in the previous section, the approaches for developing VFOA models can be broadly classified into three categories: (i) Biologically Inspired Systems, (ii) Data Driven Systems, and (iii) Heuristics Systems.

6.2.1 Biologically Inspired Systems

Biologically inspired systems replicate different neurological and biological functions of the human brain in order to generate and interpret visual attention.

Many biologically inspired VFOA models focus on directing attention to areas of interest in a visual scene by replicating the neurological response to those visual stimuli. These models compute the saliencies of several features in parallel, then combine these saliencies into a single saliency map [Frintrop et al., 2010]. To this end [Hoffman et al., 2006] propose a gaze imitation and shared attention model that combines algorithm for estimating gaze vectors with bottom-up saliency maps of visual scenes to predict the objects being looked at. The Bayesian approach is used to develop a probabilistic model. The task is to imitate the VFOA behaviour of a teacher in a multiparty setting in order to look at a common object. A maximum accuracy of 93.4% is achieved for the gaze imitation task where a robot also looks at the same object which is being looked by an instructor.

Among biologically inspired systems, researchers have also proposed models to replicate high level human cognition that operates at a level of abstraction above the low level brain neurons. In this regard, [Trafton et al., 2008] propose a multiparty gaze generation model called ACT(R/E) that switches its gaze to speakers in order to perform conversational tracking. The task for a robot is to look at the person who is speaking. For the sake of evaluation, interaction videos are recorded where gaze information is generated using their proposed models and baseline models. Independent users view the videos and rate the naturalness of the proposed model in comparison with baselines.

[Lee et al., 2007] propose a model that tightly integrates gaze behaviors with the underlying cognitive model controlling reasoning. The model is implemented in virtual agents for speaker and listener gaze generation in multiparty scenarios. Experiments are performed where human users can interact with virtual agents and rate the naturalness of the model.

6.2.2 Data Driven Systems

Data driven VFOA models are trained using corpora containing human-human interactions involving VFOA as a nonverbal source of communication. Three main steps are involved in the development of data driven VFOA models: (i) a corpus of human interactions containing VFOA annotated with the person or object in focus and the duration of gaze is collected, (ii) a model of VFOA behaviour is developed either using a set of rules or features based on the dataset, and (iii) the proposed model is evaluated on HAI.

[Liu et al., 2012] develop a rule based approach for nodding, head tilting and gaze generation in multiparty interaction. Authors use a custom dataset for the extraction of rules. Experiments are performed where a human user interacts with two robots who play a role of receptionists at an information desk. Users are asked to interact with the robots and rate subjective naturalness of robots.

[Pelachaud and Bilvi, 2003] propose a statistical gaze generation model in dyadic interaction based on Bayesian networks. The model uses a custom dataset of 20-30 minutes videos where two users interact with each other. To predict the next gaze behaviour, the feature set contains communicative functions, previous gaze information and the running time of the current gaze. The model achieves an accuracy of 75.28% for correctly predicting speaker VFOA target and 84.37% for correctly predicting listener VFOA target.

[Admoni and Scassellati, 2014] develop a non-verbal behavior generation model for a tutoring application in dyadic settings. The model is trained using a custom dataset of two humans where one human acts as teacher while the other acts as a student. The model is trained using K-Nearest Neighbours. The features are the dialogue acts, deixis and mutual gaze and gestures of the meeting participants. For gaze behaviour generation, the proposed model matches human behaviour 52% of the time.

One of the main advantages of data driven systems is that with sufficient amount of data, a model can be trained with high accuracy for a particular task. However, the need for a large amount of data is a major downside to data driven systems. Also, the development of general purpose a gaze generation model which does not depend on any particular dataset, adds to the complexity of the task and therefore leads to the development of heuristic systems.

6.2.3 Heuristic Systems

Heuristic approaches for VFOA behaviour generation rely on loosely defined rules based on human psychology and knowledge of multimodal behaviour. The heuristics are defined to generate the most appropriate VFOA behaviour without relying on any biological observations or depending upon large corpora.

[Mao et al., 2009] exploit eye movement parameters selected from the AU-Coded facial expression database [Kanade et al., 2000] and real-time eye movement data (blink rate, pupil size, and saccade), to develop a rule-based approach for generating VFOA behaviours that represent various primary (joyful, sad, angry, afraid, disgusted and surprise) and intermediate emotions (emotions that can be represented as the mixture of two primary emotions). The research work only provides guidelines to generate gaze behaviour representative of emotions. No experiments are performed on virtual agents and results have not been reported.

Another heuristic is to generate gaze behaviour based on the gaze information of the other participants in the interaction. [Sisbot and Alami, 2012] propose gaze behaviour generation model for a robot that hands an item to humans in multiparty settings. Using the gaze information of the human, the robots plan the place of handover and then convey this information to the human by looking at the place where the object will be handed over. The robot performance is evaluated in terms of naturalness via user interaction.

[Peters et al., 2005] develop a gaze generation model that monitors participant gaze to access their level of engagement and then generates gaze movements to maintain and enhance participant's engagement in the conversation. The model consists of a set of rules for both speaker and listener VFOA behaviour generation in multiparty settings.

6.2.4 Summary & Discussion

A summary of the existing works for VFOA behaviour generation is presented in Table 6.1.

Human VFOA is not a direct function of any specific input. Several factors are involved in the duration and direction of visual attention. In addition, human VFOA behaviour generation mechanisms involve multiple tasks. For instance, humans change their VFOA multiple times and look at some participants or objects for longer duration than at others. Therefore, VFOA necessitates to predict not only a direction but also shifts as well as duration for each VFOA turn. However, most of the existing VFOA behaviour generation models for agents are developed specifically to accompany certain actions, and rely on individual stimuli e.g. in [Peters et al., 2005; Sisbot and Alami, 2012].

Reference	**Approach**	**Task**	**Meeting Type**	**Evaluation**
[Hoffman et al., 2006]	Biologically Inspired	Human Gaze Imitation	Multiparty	93.4% accuracy
[Trafton et al., 2008]	Biologically Inspired	Look at the current speaker	Multiparty	User Evaluation
[Lee et al., 2007]	Biologically Inspired	End-to-end gaze generation for both speaker and listener	Multiparty	User Evaluation
[Liu et al., 2012]	Data Driven (Rule Based)	End-to-end gaze generation for both speaker and listener	Multiparty	User Evaluation
[Pelachaud and Bilvi, 2003]	Data Driven (Belief Networks)	End-to-end gaze generation for both speaker and listener	Dyadic	75.28% for speaker gaze prediction, 84.37% for listener VFOA target
[Admoni and Scassellati, 2014]	Data Driven (KNN Algorithm)	End-to-end gaze behaviour generation for both speaker and listener	Dyadic	52.% overall accuracy
[Mao et al., 2009]	Heuristics	VFOA generation representing emotions	Multiparty	Results not reported
[Sisbot and Alami, 2012]	Heuristics	Mutual gaze and joint attention	Multiparty	User Evaluation
[Peters et al., 2005]	Heuristics	gaze generation to enhance participant's engagement	Multiparty	Not Reported

Table 6.1 – Summary of related works for speaker and listener VFOA generation in multiparty interaction

Moreover, models developed for end-to-end VFOA behaviour generation focus on a partial list of aspects of VFOA. For example, they predict only the duration and direction of VFOA (i.e. where and how long a speaker looks at a particular participant or object) during a whole utterance without any shift of VFOA [Lee et al., 2007; Liu et al., 2012].

In this research work, we propose a modular approach to develop a hybrid VFOA model that combines data driven and heuristic approaches to generate VFOA behaviour of agents in multiparty interaction. We divide VFOA generation into four sub-tasks: Number of VFOA turn prediction, duration of VFOA turn, VFOA target per turn and VFOA turn scheduling. A VFOA turn refers to a stretch of VFOA towards a single target which can be either a person or an object.

Our basic hypothesis is that VFOA estimation is a function of multiple input features

and the relationship between those input features and the VFOA can be learned via machine learning techniques. Furthermore, heuristic techniques are also used for instance to process outputs of sub-tasks.

6.3 Feature Selection

The proposed visual focus of attention (VFOA) behaviour generation model consists of a hybrid architecture that combines machine learning and heuristics. Section 2.7 contains details about different features used for various multimodal human-agent interaction tasks. As previously, feature list is compiled on the basis of relevance of features for speaker and listener VFOA generation behaviour as mentioned in the literature. To reaffirm the importance of the selected features for VFOA generation, a brief analysis of some of the features is performed on the AMI [Carletta, 2007] and MPR datasets [Funakoshi, 2018]. The features selected for speaker and listener VFOA generation behaviour are: current previous speakers, current and previous addressee, DA, utterance duration, and participant list.

6.3.1 Current and Previous Speakers

Speaker information refers to any information related to the speaker of the utterance such as name or role used to uniquely identify the speaker of the utterance. Works from [Buswell, 1935; Guy et al., 2019; Yarbus, 1967] show that the VFOA behaviour is dependent on individual participants of an interaction.

In addition, analysis of the AMI and MPR datasets shows that on average the number of VFOA target changes per utterance vary by speaker. Some speakers have a longer utterance duration but less VFOA target changes while others have shorter utterances but a higher number of VFOA target changes, which highlights the role of speaker information for VFOA generation in human-agent interaction.

6.3.2 Current and Previous Addresses

Since the current and previous addressee are also meeting participants. As explained in the previous section, the VFOA behaviour generation varies greatly with individual meeting participants, [Buswell, 1935], [Yarbus, 1967], [Guy et al., 2019]. In addition, it is shown in Section 4.3. that gaze is one of the most important features for addressee detection in multiparty interaction. Owing to the strong statistical relationship between

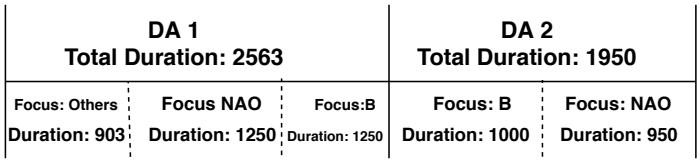

Figure 6.1 – Example of Duration and Direction of VFOA of Speaker turns during DA

the VFOA and addressee role, the current and previous addressees have been included in the feature set.

Another reason for including current and previous speakers and current and previous addressees is to deal with the cases where the VFOA overlaps between two continuous DAs as illustrated in Figure 6.1. The VFOA target of the DA1 and DA2 overlaps as the direction of the last VFOA turn of DA1 is equal to the direction of the first VFOA turn of DA2. The data analysis of MPR shows that on average when the current and previous addresses for two successive DAs are the same, the average VFOA duration of the first VFOA turn of the next DA is longer when compared to the cases where the previous and next addresses of the two DAs are different.

6.3.3 Utterance Duration

The duration of an utterance is directly related to the number of VFOA turns in a dialogue act. Longer dialogue acts tend to have higher number of VFOA turns compared to shorter dialogue act. This relation is confirmed by the analysis of AMI and MPR where a linear trend is observed between the number of turns and average duration of the DA containing the turns.

6.3.4 Dialogue Act

Past works have shown that DA and eye gaze are strongly related [Andrist et al., 2014; Poggi et al., 2000]. For instance in both AMI and MPR, the duration and the number of gaze turns are lower for the DAs signaling an Agreement or a Disagreement than for Information DAs. AMI uses a custom DA taxonomy whereas MPR is annoted with the DIT++ taxonomy [Bunt, 2009].

6.3.5 Start and End time of DA

Works from [Nakano and Ishii, 2010] highlight that user engagement and VFOA behaviour are correlated. The study show that the human engagement leads to the stability

of VFOA behaviour for longer periods. In addition [Baecker, 1993] explore that the early part of a meeting is more lively where user engagement is higher. After 15-20 minutes, attention lag sets in. The importance of start and end time of an utterance can therefore play an important role in generating VFOA behaviour particularly concerning listeners. For instance, in the early part of a meeting, user engagement and attention is higher and hence gaze patterns are more stable compared to the later part of the meeting when users attention lag occurs leading to an increase in unpredictability of VFOA behaviours.

6.3.6 Participant List

The feature *participant list* contains names of all the persons currently present in the interaction field. In a multiparty human-agent interaction, a participant may leave or enter a meeting. For instance, in MPR, the participants are free to move inside and outside of the field of interaction. Hence the number of meeting participants may change and it is important to record the participants present in the field of interaction.

6.3.7 Discussion

Several researchers have used statistical techniques for VFOA behaviour generation in multiparty interaction [Pelachaud and Bilvi, 2003], [Admoni and Scassellati, 2014]. Most of existing VFOA behaviour generation models for agents are developed specifically for accompanying certain actions, and rely on individual stimuli, e.g. [Peters et al., 2005; Sisbot and Alami, 2012]. Moreover, the models developed for end-to-end VFOA behaviour generation model only perform a specific task: they predict the duration and taeget of VFOA (i.e. where a speaker looks at a particular participant or object) during a whole utterance without any variation of VFOA target [Lee et al., 2007; Liu et al., 2012].

Human VFOA behaviour generation mechanism involves multiple tasks such as predicting number of VFOA turns, duration per turn, VFOA target, etc. Moreover, human VFOA is not a direct function of any specific input. There are several factors involved in deciding the duration and target of VFOA for instance the DA, the speaker role, the addressee role, etc.

Despite the evidence from existing literature, to the best of our knowledge, none of the existing works incorporate speaker and addressee roles, and start and end time of utterance as features to train a machine learning based model for VFOA behaviour generation. In addition, most of the existing approaches for VFOA behaviour generation work with a fixed number of participants and hence the feature participant list has never

been used as a feature previously. However the number of meeting participants can increase or decrease at anytime and this particularly should be taken into account.

In the proposed research work, three improvements are proposed in comparison with existing systems (i) new features have been considered to train machine learning algorithms, (ii) a solution based on sub-models is proposed to tackle various interrelated VFOA behaviour generation problems, (iii) the proposed solution is independent of the number of participants.

The next section formalizes the problem, details the task and describes the models developed for speaker and listener VFOA behaviour generation.

6.4 Problem Formalization and Methodology

This section formalizes the sub-tasks that the VFOA behaviour generation models performs

This section explains the tasks identified for end-to-end VFOA behaviour generation for speakers and listeners. VFOA is generated at the level of a DA; in the following the term “utterance” is used for a single DA.

Task 1: Number of VFOA turns per DA

During an utterance, a meeting participant can change its VFOA target multiple times. For instance, a speaker can look at one or multiple participants or objects, especially if the addressee is a group. A *VFOA turn* for a participant refers to a stretch of VFOA towards a single target which can be a person or an object. As soon as the target of VFOA changes, new turn starts. The task of predicting the number of turns has been framed as a regression problem, since the number of VFOA turns can be any number and is not limited to a set of predefined labels.

Task 2: Target of VFOA per turn

A participant can look at different person(s) or object(s) during an utterance. Therefore, planning several VFOA targets, with a unique person or object by VFOA turn, is important to exhibit a natural behaviour. Since the target of VFOA can be a fixed number of persons or objects, VFOA target prediction has been framed as a multi-label classification problem where output vector consists of one or more VFOA targets.

Task 3: Duration of each VFOA Target per DA

A participant may look at a certain person or object for a longer duration than to others. The problem of VFOA duration per target prediction is framed as a regression problem since the output is a vector of key-value pairs where keys are the targets of the VFOA, i.e. the person(s) or object(s) that is (are) in focus of attention, and values contain the fraction of the overall duration.

Task 4: Scheduling of the VFOA turns

The final step is to schedule the VFOA turns. The VFOA schedule refers to the order of the VFOA turns during an utterance. The schedule generation is also framed as a classification problem because the number of VFOA scheduling pattern, no matter how large, is technically finite and hence it cannot be framed as a regression problem.

6.5 VFOA Behaviour Generation Model

During a mixed interaction, involving humans and intelligent agents, an intelligent agent can speak and listen to other agents as well as human speakers.

All the features are available for the speaker VFOA behaviour generation and listener VFOA behaviour generation when an agent speaks. However a difficult arises when a human speaks because in such a case utterance duration, DA and DA end time can only be estimated at the end of an utterance. When the speaker is an agent, the response of the agent (verbal or non-verbal), is known and therefore, the values for utterance duration, DA and DA end time can be estimated. These values can also be communicated to other agents present in the interaction field. Hence, for speaker VFOA behaviour generation and for listener VFOA generation when an agent speaks, all the features can be exploited. On the other hand, for listener VFOA behaviour generation when a human speaks, listener agents cannot estimate the values of utterance duration, DA and DA end time before the utterance ends. Hence, the utterance duration, DA and DA end time cannot be exploited for listener VFOA generation when a human speaks.

Therefore, participant VFOA behaviour generation model is divided into three sub-models: speaker VFOA behaviour generation model abbreviated as SVFOA, and listener VFOA behaviour generation model abbreviated as LVFOA. LVFOA model is further divided into two types: listener VFOA behaviour generation model when an agent is speaker (LVFOA-AS) and listener VFOA behaviour generation model when a human speaks (LVFOA-HS).

6.5.1 Speaker (SVFOA) and Listener VFOA Generation when an Agent Speaks (LVFOA-AS)

Since the problem is divided into 4 inter-related tasks, a modular architecture whose components are dedicated to each sub-problem is proposed. There are two advantages to a modular approach: (i) it can support a hybrid architecture where each module is implemented via different models depending upon the characteristics of the individual problem, and (ii) performances of the models can be evaluated individually.

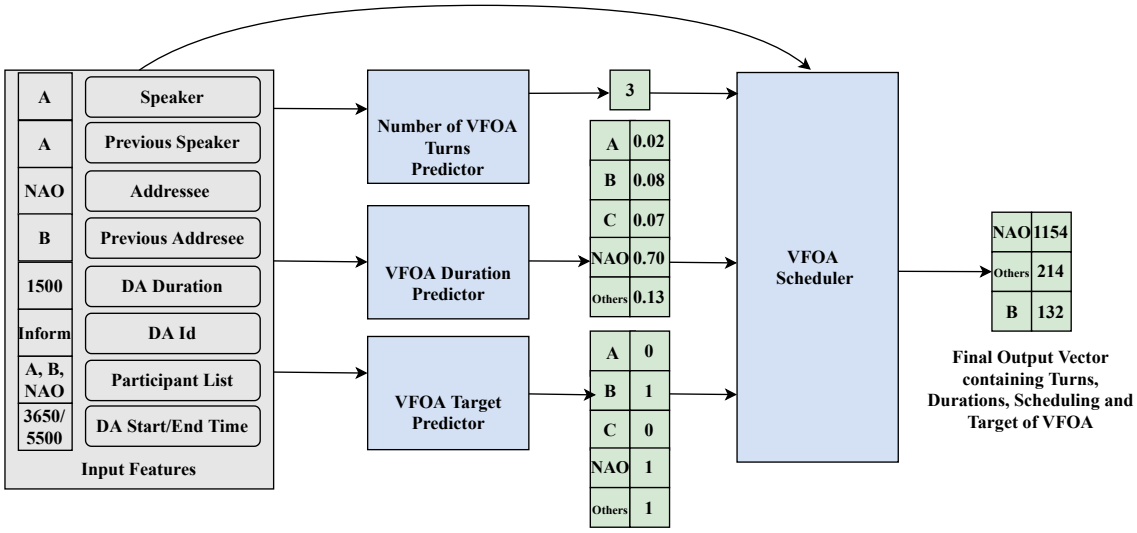

Figure 6.2 – SVFOA and LVFOA-AS Behaviour Generation Model (Values taken from MPR) [Funakoshi, 2018]

Figure 6.2 depicts the architecture of the proposed VFOA behaviour generation model and illustrates an example of how data flows between the elements in the experimental conditions of MPR [Funakoshi, 2018]. Similarly, concerning the experimental conditions of AMI [Carletta et al., 2005], the participants A, B, C and NAO are replaced by PM, UI, ID, and ME and an additional VFOA target towards *Objects* is added.

The model has four main sub-models: Number of VFOA Turns Predictor, VFOA target Predictor, VFOA Duration per Target Predictor, and VFOA Scheduler.

Number of VFOA Turns Predictor

Number of VFOA Turns Predictor is a machine learning based sub-model that predicts the number of VFOA turns during the current DA. The sub-model requires as input the features described in section 6.3. Number of VFOA Turns Predictor learns the relationship between these features and the number of VFOA turns. The output is half

rounded up to zero decimal places. For example in Figure 6.2, the output from Number of VFOA Turns Predictor is 2.8, then rounded to the next integer i.e. 3.

VFOA Target Predictor

VFOA target Predictor predicts the persons and/or objects in VFOA. Similarly to VFOA Duration Predictor, VFOA target Predictor outputs a vector of size 5 in MPR conditions and 6 in AMI conditions. Figure 6.2 depicts an example of the output vector from VFOA target Predictor. The output vector is a one-hot encoded vector where 1 corresponds to predicted VFOA target. The predicted VFOA target in Figure 6.2 is B, NAO and Others.

VFOA Duration Predictor

VFOA Duration Predictor is also a machine learning based sub-model which predicts the overall duration of VFOA per target during a DA For example in MPR, the VFOA target can be towards participants A, B, C, NAO or Other, so the length of the output vector of probabilities is 5. Figure 6.2 shows an example of output vector of VFOA Duration Predictor. This vector conveys that the probability of VFOA duration turn directed towards A is 2%, the probability of VFOA turn directed towards NAO is 70% and so on. Similarly in AMI, the length of the output vector for VFOA Duration Predictor is 6 (PM, UI, ID, ME, Objects, Others).

VFOA Scheduler: Turn Scheduling and Final VFOA behaviour

VFOA turns seems not follow any specific pattern. For instance in MPR, there is a total of 10,214 utterances that contain more than 1 VFOA turn. For these 10,214 utterances the total number of unique VFOA turn patterns is 470. This shows that machine learning approaches are not suitable to schedule VFOA turns. A heuristic approach is therefore proposed, to exibit models that is specific to the speaker and to the listeners.

For the SVFOA scheduling, the scheduler model follows the set of rules described in Algorithm 1, and for the LVFOA-AS scheduling, the scheduler model for LVFOA-AS follows the set of rules described in Algorithm 2.

Speaker VFOA Scheduling

The algorithm is designed as follows: if Number of VFOA Turns Predictor predicts a single person/object, then that person/object is the only item in the VFOA target (lines 1-2). If VFOA Turns Predictor predicts multiple targets, then the VFOA turns have to be scheduled. The algorithm first checks whether for the current DA, the speaker and addressee(s) are the same as those of the previous DA, meaning that the speaker is

Algorithm 1 VFOA Scheduling for Speaker

Result: Final VFOA target and duration
Input : Num-VFOA-Turns (Output from Number of VFOA Turns Predictor), VFOA-Target-Vector (dictionary that maps the VFOA durations with VFOA targets), Current-Speaker, Previous-Speaker, Current-Addressee, Previous-Addressee, Last-VFOA-PDA (Last VFOA target of previous DA)
Output : Predicted-VFOA (dictionary where key: target / value: duration)

```
   // If the Num-VFOA-Turns contains only 1 item
1  if size of Num-VFOA-Turns == 1 then
2  |  Predicted-VFOA.add(the target and duration of the only object/person in VFOA-Target-Vector)
3  else
   |  // If VFOA-Target-Vector contains more than 1 item
4  |  if Current-Speaker == Previous-Speaker and Current-Addressee == Previous-Addressee and Last-VFOA-PDA in VFOA-Target-Vector then
5  |  |  i) Add the target and duration of item Last-VFOA-PDA from VFOA-Target-Vector to Predicted-VFOA
6  |  |  ii) Remove the item with the same target as Last-VFOA-PDA from VFOA-Target-Vector
7  |  end
8  |  foreach for each item i in VFOA-Target-Vector do
9  |  |  if p(duration) of i > 0.05 then
10 |  |  |  i) Add i to Predicted-VFOA
11 |  |  end
12 |  end
13 |  if size of VFOA-Target-Vector < Num-VFOA-Turns and VFOA-Target-Vector contains outlier duration then
14 |  |  foreach item j in VFOA-Target-Vector[outlier durations] do
15 |  |  |  i) New duration = j[outlier duration] - j[duration]
16 |  |  |  ii) Select one of the items from VFOA-Target-Vector in round robin way
17 |  |  |  iii) Create new turn by assigning new duration to the item selected in step ii
18 |  |  |  iv) Add newly created turn to Predicted-VFOA
19 |  |  end
20 |  end
21 |  Return Predicted-VFOA
22 end
```

basically continuing its utterance with the same addressee(s). In this case, if the VFOA target of the last VFOA turn (for the previous DA) exists in the list of selected targets for current DA, that VFOA turn is scheduled first (lines 4-6). The rationale is to mimic the behaviour illustrated in Figure 6.1 where VFOA target towards participant B overlaps between DA1 and DA2 since B is the last focus direction during DA1 and first focus direction during DA2.

Then, all the VFOA targets with predicted probability of VFOA duration lower than 5% of the total DA duration are removed (lines 8-10), in order to avoid very quick VFOA target changes from the speaker. Finally, if the number of turns predicted by Number of VFOA Turns Predictor is greater than the number of unique VFOA targets predicted by the algorithm and any of the VFOA duration is greater than an upper outlier limit, the remaining VFOA duration is added as a new VFOA turn where the direction is chosen

in a round-robin way from the list of VFOA targets (lines 13-18). In this way, very long VFOA durations are avoided and the behaviour where a speaker looks at a particular person or object multiple times during a DA is replicated.

The sum of probabilities of the VFOA targets predicted by the VFOA Duration Predictor and not included in the output of the VFOA target Predictor are distributed over the directions already predicted. For example in Figure 6.2, the predicted directions are B, NAO and Others. Since A and C are not included in the output, the sum of their probabilities, i.e. 0.09, is divided and added to the duration probabilities of NAO, B and Others. Hence, the final probabilities of NAO, Others, and B become respectively 0.769, 0.142 and 0.088. Final VFOA duration is calculated by predicting DA duration according to the probability of duration towards an object. For instance for NAO, the duration is 1500 x 0.769 = 1154.

Listener VFOA Scheduling when an Agent Speaks

Algorithm 2 contains rules for listener VFOA Scheduling when an agent speaks and is quite similar to the Algorithm 1. VFOA Turns Predictor predicts a single person/object, then that person/object is the only item in the VFOA target (lines 1-2).

Then, all the VFOA targets with predicted probability of VFOA duration lower than 5% of the total DA duration are removed (lines 4-6), in order to avoid very quick VFOA target changes.

Finally, if the number of turns predicted by Number of VFOA Turns Predictor is greater than the number of unique VFOA targets predicted by the algorithm and any of the VFOA duration is greater than an upper outlier limit, the remaining VFOA duration is added as a new VFOA turn where the direction is chosen in round robin way from the list of VFOA targets (lines 9-14).

6.5.2 Listener VFOA Behaviour Generation when a Human Speaks (LVFOA-HS)

LVFOA-HS behaviour generation model consists of two sub-models: VFOA Target Predictor and a heuristic-based sub-model that predicts the number of turns, duration per VFOA target and scheduling for VFOA. Figure 6.3 depicts the architecture of the proposed LVFOA-HS behaviour generation model and illustrates an example of how data flows between the elements in the experimental conditions of MPR [Funakoshi, 2018].

Algorithm 2 Listener VFOA Scheduling when speaker is an Agent

Result: Final VFOA target and duration

Input : Num-VFOA-Turns (Output from Number of VFOA Turns Predictor), VFOA-Target-Vector (dictionary that maps the VFOA durations with VFOA targets)

Output : Predicted-VFOA (dictionary where key: target / value: duration.)

```
// If the Num-VFOA-Turns contains only 1 item
if size of Num-VFOA-Turns == 1 then
    Predicted-VFOA.add(target and duration of the only object/person in VFOA-Target-Vector)
else
    // If VFOA-Target-Vector contains more than 1 item
    foreach for each item i in VFOA-Target-Vector do
        if p(duration) of i > 0.05 then
            i) Add i to Predicted-VFOA
        end
    end
    if size of VFOA-Target-Vector < Num-VFOA-Turns and VFOA-Target-Vector contains outlier duration then
        foreach item j in VFOA-Target-Vector[outlier durations] do
            i) new duration = j[outlier duration] - j[duration]
            ii) Select one of the items from VFOA-Target-Vector in round robin way
            iii) Create new turn by assigning new duration to the item selected in step ii
            iv) Add newly created turn to Predicted-VFOA
        end
    end
    Return Predicted-VFOA
end
```

VFOA Target Predictor

VFOA target Predictor predicts a list of participants or objects that a participant may look at to during an utterance. The LVFOA-HS is similar to SVFOA and LVFOA-AS. The only difference is that LVFOA-HS does not exploit utterance duration, DA and DA end times. Figure 6.3 depicts an example of the output vector from VFOA target Predictor. The output vector is a one-hot encoded vector where 1 corresponds to predicted VFOA target. The predicted VFOA target in Figure 6.3 are NAO and Others.

VFOA Turn & Duration Predictor and Scheduling

For number of turns, duration per turn and turn scheduling for LVFOA-HS, a heuristic based model is proposed since if the utterance duration is not known before the end of

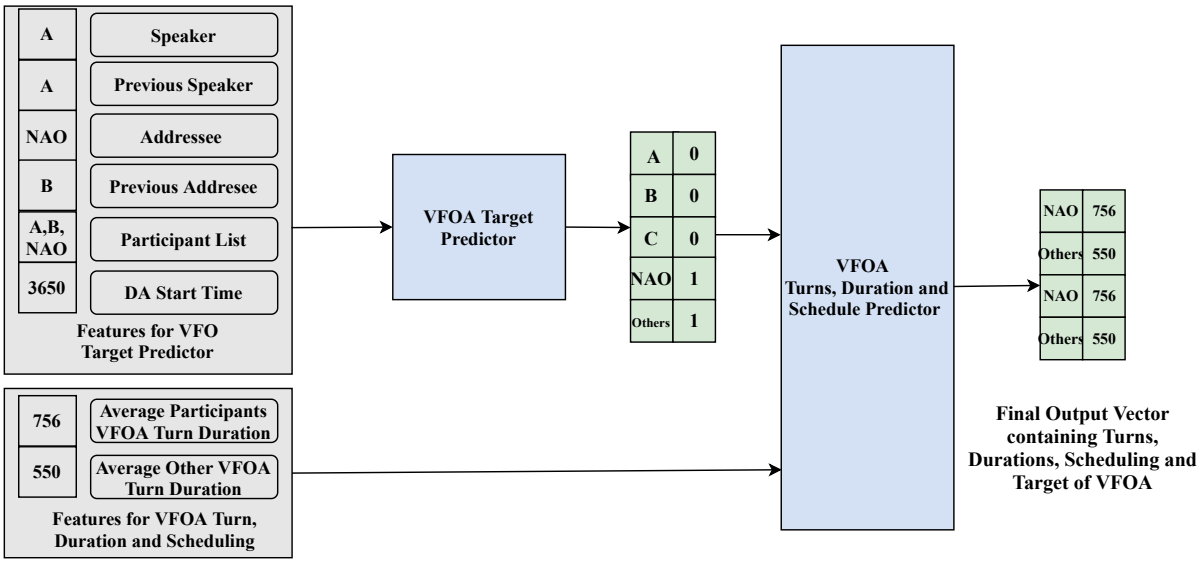

Figure 6.3 – LVFOA-HS Behaviour Generation Model

utterance, predicting a fixed value for number of turns may result in very long or very short turn durations. For example, if the number of turns is predicted as 1, and human speakers continues speaking for a very long duration, that 1 turn duration becomes very long. Therefore instead of employing machine learning models that outputs a fixed value, a heuristic approach is proposed that dynamically increments the number of VFOA turns. For the overall duration per VFOA target in an utterance, the average values for turn duration when VFOA target is human, objects or others, are used from corresponding datasets. Algorithm 3 explains the process of VFOA turn prediction, duration per turn and scheduling for listeners when a human speaks.

The algorithm is designed as follows: If the VFOA Target predictor predicts a single person/object, that person is the only item in the VFOA target. The number of turn will be 1 whose duration is equal to the duration of the whole utterance (lines 1-4).

If the VFOA target predictor predicts more than 1 target then while the human continues speaking, for each turn a person/object is selected from the list of targets. The duration of turn is selected from the list of durations depending upon the target type i.e. a person or object (lines 8-11).

An example of number of turns, direction and scheduling prediction for an utterance in MPR is depicted in figure 6.3. The VFOA turns, duration and scheduler predictor model takes as input VFOA target list and average turn duration values. In MPR, the average turn duration when target the is a person is 756 milliseconds, which decreases to 550 milliseconds when the target is an object. The VFOA scheduler then simply assigns the durations to corresponding VFOA targets in a round robin way as long as human

Algorithm 3 VFOA Scheduling for Listeners when speaker is a Human

Result: Final VFOA target and duration

Input : VFOA-Targets (Output from VFOA Target Predictor), Average-VFOA-Durations (vector containing average turn durations for participants, objects and others from dataset)

Output: Predicted-VFOA (list that contains VFOA targets and durations per target)

```
// If the number of VFOA-Targets contains only 1 item
if VFOA-Targets == 1 then
    Predicted-VFOA.add:
    i) the only item from VFOA-Target list,
    ii) the whole utterance duration
else
    // If VFOA-Targets contains more than 1 target
    while Human continues speaking
    (
    foreach target in VFOA-Targets do
        Predicted-VFOA.add:
        i) the target from VFOA-Targets in round-robin way
        ii) and corresponding duration of the object/person/others from Average-VFOA-Durations
    end
    )
    Return Predicted-VFOA
end
```

continues speaking. The number of turns, and duration per turn can be calculated from the final VFOA. For instance in Figure 6.3, the number of turns is 4 and the duration when the focus is NAO is 756 x 2 = 1512. If a human continues speaking for longer durations the turns and corresponding duration per participant or object is also increased dynamically.

The next section explains the experimental evaluation process and the results received.

6.6 Experimental Evaluation

This section briefly describes the datasets, and explains the procedure followed to perform the experimental evaluation, followed by the results obtained for different experiments.

6.6.1 Datasets

The datasets used to train and test the VFOA generation sub-models are AMI [Carletta, 2007] and MPR [Funakoshi, 2018]. These are the only datasets that contain annotated data for all the features

The AMI Dataset

For the experiments performed on AMI, the VFOA target is divided into three categories: participants, objects and other entities. Participants are identified by their roles in the AMI dataset i.e. PM, ID, ME or UI. A meeting participant can look at other participants (PM, ID, ME or UI), or objects (table, slide-screen, etc) grouped into a single category and other entities.

6.6.1.1 The MPR Dataset

For experiments performed on MPR, VFOA target can be participant and other entities. Participants are identified by their IDs in the MPR dataset i.e. A, B, C or robot NAO. A participant can look at other participants (A, B, C or NAO robot), other entities classified into one category. There are no objects to look at in the MPR.

Further details of the AMI and MPR datasets are already presented in Section 2.6.2.

6.6.2 Baselines

To the best of our knowledge, none of the existing research works propose solutions to the four VFOA tasks. To evaluate our approach, experiments are performed using two custom baselines for comparison.

Baseline 1: Random VFOA Behaviour Generation (Baseline-random)

In baseline-random, the speaker VFOA is generated randomly. The numbers of turns per utterance are randomly selected according to the overall probability of the number of turns per utterance in the corresponding datasets. The VFOA duration per turn is calculated by dividing the number of turns per utterance by the total utterance duration. The turn direction is selected randomly from the list of listeners and objects in a round-robin way according to the default distribution of VFOA targets in the corresponding dataset.

Baseline 2: Rule-based VFOA Behaviour Generation (Baseline-rule-based)

In the baseline-rule-based, VFOA is generated on the basis of a set of rules:

1. The number of VFOA turns is generated as a linear function of VFOA duration using a linear regression algorithm.

2. The VFOA duration per turn is calculated by dividing the number of turns by the total utterance duration.

3. **Turn target for speaker**: (i) if the addressee of the turn is Group then the participant or object for VFOA target is randomly selected from the list of addressees or objects in a round-robin way, (ii) if the addressee is an individual, the VFOA target for the first turn is the individual that is being addressed. For the remaining turns, participants or objects for VFOA target are randomly selected from the list of addressees or objects in a round-robin way.

   Turn target for listener: (i) if an agent is directly addressed, set the speaker as the VFOA target for the agent, (ii) if the addressee of the utterance is a group or if an agent is addressed indirectly, set speaker as VFOA target of for first turn, then randomly select VFOA targets for remaining turns from the list of addressees or objects in a round-robin way.

6.6.3 Experiments for Speaker VFOA (SVFOA) and Listener VFOA when Agents Speak (LVFOA-AS)

For SVFOA and LVFOA-AS behaviour generation, experiments are performed using the two baselines and the proposed approach to evaluate the performances of Number of VFOA Turns Predictor, VFOA Duration Predictor, VFOA target Predictor and VFOA Scheduler detailed in Section 6.5.1. The experiments are performed using the AMI and MPR datasets. Concerning AMI, the 14 meetings [1] that contain all the features are used for training and testing the machine learning based sub-models. For MPR, all the meetings from the two scenarios[2] are employed for training and testing. To evaluate the stability of the results, data is divided into training and test sets via five fold cross-validation.

[1] *IS1003d, IS1006b, IS1001b, IS1003b, IS1008b, TS3005a, TS1000a, IS1001c, IS1008c, ES2008a, IS1008a, IS1001a, IS1006d, 'IS1008d', IS1008a*

[2] except meeting id 24 from scenario 1 due to missing data

Experiments for Number of VFOA Turns Predictor

Since the number of turns is tackled as a regression problem, seven of the most commonly used supervised learning algorithms for regression are trained to evaluate the best model performance: Extreme Gradient Boosting Regression (XGBR) [Sheridan et al., 2016], Random Forest Regression (RFR) [Liaw et al., 2002], Support Vector Machines Regression (SVR) [Tong et al., 2009], Multilayer Perceptron (MLPR) [Murtagh, 1991], K Nearest Neighbour Regression (KNNR) [Devroye et al., 1994], Linear Regression (LRR) [Montgomery et al., 2012], Bayesian Ridge Regression (BRR) [Tipping, 2001]. The results are presented in terms of Mean Absolute Error (MAE), that indicates the average error for all the test samples.

Experiments for VFOA Duration Predictor

The VFOA distribution prediction is also tackled as a regression problem but with multiple outputs since the goal is to predict the probability of VFOA distribution for all the participants which can be in VFOA. The regression algorithms and the distribution of training and test sets are the same as for the number of VFOA Turns Predictor. Since VFOA distribution prediction is a regression problem, the results are also presented in terms of Mean Absolute Error (MAE), averaged on outputs.

Experiments for VFOA Target Predictor

VFOA target predictor performs a multi-label classification where predicted output contains one or multiple objects or persons. The algorithms used for supervised multiclass classification are: XGboost (XGB) [Chen and Guestrin, 2016], Multilayer Perceptron (MLP) [Kruse et al., 2013], Random Forest (RF) [Liaw et al., 2002], Logistic Regression (LR) [Hosmer Jr et al., 2013], Support Vector Machines (SVM) [Hearst et al., 1998], Naive Bayes (NB) [Rish et al., 2001] and K-Nearest Neighbours (KNN) [Zhang and Zhou, 2005]. Micro-average F1 is the metric used to evaluate the performance of a VFOA target predictor.

Experiments for VFOA Scheduler and Final VFOA Prediction

VFOA Scheduler is a heuristics based sub-model that predicts the scheduling of all the VFOA turns predicted by the Number of VFOA Turns Predictor, the VFOA Duration Predictor, and the VFOA target Predictor. Since AMI contains 1,048 unique combinations of VFOA Schedules while MPR has 497 unique combinations of schedules. It is highly unlikely that the whole schedule matches exactly the actual VFOA targets and order.

However, for the sake of quantifying, the results for VFOA scheduling are evaluated in terms of accuracy.

6.6.4 Experiments for Listener VFOA when a Human Speaks (LVFOA-HS)

For LVFOA-HS behaviour generation, experiments are performed using the two baselines and the proposed approach to evaluate the performances of Number of VFOA Target Predictor and VFOA turns, duration and scheduling predictor. The datasets and the and cross validation approach used for experimentation are the same as the ones used for the experimentation of SVFOA and LVFOA-AS behaviour generation.

Experiments for VFOA Target Predictor

Experiments for VFOA target predictor for LVFOA-HS behaviour generation are exactly the same as experiments performed for SVFOA and LVFOA-AS. The main difference is that the former does not exploit utterance duration, DA and DA end time in the feature set.

Experiments for VFOA Turns, Durations, and Scheduling Predictor

VFOA Turns, Durations, and Scheduling Predictor is a heuristics based sub-model that dynamically predicts and updates the number of VFOA turns, the VFOA duration per turn, and the turn scheduling while humans speak. Mean absolute error is the metrics used to evaluate the performance of number of VFOA turns and durations prediction. Accuracy is used as a metric to measure the performance of VFOA scheduling.

6.7 Results

This section presents the results of experiments for speaker SVFOA, LVFOA-AS and LVFOA-HS behaviour generation.

6.7.1 Results for Speaker (SVFOA) & Listener VFOA when Agents Speak (LVFOA-AS)

The experimental results obtained to evaluate the Number of VFOA Turns Predictor, the VFOA Duration Predictor, the VFOA target Predictor and the VFOA Scheduler for

SVFOA and LVFOA-AS are presented in Tables 6.2-6.5. The results depicts the average values and standard deviations using five fold cross-validation.

Results for the Number of VFOA Turns Predictor

The results for the experiments performed for speaker (SVFOA) and listener (LVFOA) number of VFOA turns prediction are presented in Table 6.2. The table depicts that, for AMI, the minimum MAE for number of VFOA turns prediction for speaker (1.08) is obtained via LR and BRR algorithms. This value is slightly greater than the 1.04 achieved via the baseline-rule-based. For listener, a minimum MAE value of 0.35 is obtained via the SVR algorithm which is higher than the 0.27 obtained via baseline-rule-based. For both speakers and listeners in MPR, minimum MAE values of 0.49 are returned via the SVR algorithm. These values are lower than the MAE values for the baselines.

	AMI Dataset		**MPR Dataset**	
Algorithm	**Speaker**	**Listener**	**Speaker**	**Listener**
XGBR	1.14 ±0.08	0.38 ±0.09	0.54 ±0.01	0.53 ±0.05
RFR	1.09 ±0.05	0.36 ±0.10	0.55 ±0.02	0.54 ±0.06
MLPR	1.39 ±0.20	0.40 ±0.04	0.61 ±0.05	0.61 ±0.09
SVR	**1.08 ±0.09**	**0.35 ±0.04**	**0.49 ±0.02**	**0.49 ±0.07**
KNNR	1.20 ±0.04	0.39 ±0.04	0.56 ±0.01	0.56 ±0.06
LRR	1.07 ±0.06	0.39 ±0.05	0.60 ±0.13	0.58 ±0.09
BRR	**1.08 ±0.05**	0.39 ±0.04	0.54 ±0.02	0.51 ±0.05
Baseline-Random	1.66 ±0.13	0.41 ±0.02	0.74 ±0.02	0.72 ±0.05
Baseline-Rule-based)	**1.04** ±(0.03)	0.27 ±0.07	0.50 ±0.02	0.51 ±0.05

Table 6.2 – Results for number of VFOA turns prediction for speaker VFOA (SVFOA) and listener VFOA when an agent speaks (LVFOA-AS). Values represent MAE.

The higher MAE values for speaker for AMI than for MPR can be attributed to two possible reasons. (i) The possible number of VFOA targets is higher in AMI, (ii) for speaker, the average number of turns for AMI is higher than MPR which leads to a higher probability of error. For listener, the MAE values obtained on MPR are higher compared to AMI which can again be attribute to the average of number of VFOA turns which is higher for MPR compared to the AMI dataset. In addition, in case of AMI, baseline-ruleb-ased returns best results for VFOA turns prediction for speakers and listeners. Similarly, for MPR, the results obtained via baseline-rule-based for speakers and listeners (respectively, 0.50 and 0.51) are close to the minimum MAE values of 0.49 obtained via SVR. This shows that the number of VFOA turns can also be calculated as a linear function of gaze duration.

For speakers, the results are significant at $p < 0.05$ (p=0.0002 on AMI and p=0.0001 on MPR) with respect to baseline-random, but not significant at (p=0.1193 on AMI and p=0.0539 on MPR) with respect to baseline-rule-based. For listeners, the results are significant at $p < 0.05$ (p=0.017 on AMI and p=0.00038 on MPR) with respect to baseline-random, but not significant at (p=0.102 on AMI and p=0.067 on MPR) with respect to baseline-rule-based.

Results for the VFOA Duration Predictor

The Table 6.3 depicts results for VFOA duration predictions for speakers (SVFOA) and listeners (LVFOA-AS) in the AMI and the MPR datasets.

The results for the experiments for VFOA distribution prediction show that for AMI the minimum MAE value of 0.14 is obtained via RFR for the speaker model. For listeners in AMI, a minimum MAE value of 0.07 is achieved via RFR. For MPR, a minimum MAE value of 0.11 obtained via the XGBR for speaker. For listener, XGBR returns a minimum MAE value of 0.21. The results show that for both speakers and listeners in the AMI and MPR, the proposed approach returns better results than the baselines, for this task.

For speakers, the results are significant at $p < 0.05$ (p=0.0007 on AMI and p=0.0006 on MPR for baseline-random, and p=0.00009 on AMI and p=0.00005 on MPR for baseline-rule-based). The results are also significant for listeners at $p < 0.05$ (p=0.000068 on AMI and p=0.000016 on MPR for baseline-random, and p=0.000058 on AMI and p=0.000027 on MPR for baseline-rule-based).

	AMI Dataset		**MPR Dataset**	
Algorithm	**Speaker**	**Listener**	**Speaker**	**Listener**
XGBR	0.16 ±0.02	0.08 ±0.02	**0.11 ±0.01**	0.22 ±0.01
RFR	**0.14 ±0.01**	**0.07** ±0.01	0.12 ±0.01	**0.21 ±0.01**
MLPR	0.20 ±0.01	0.22 ±0.06	0.16 ±0.02	0.23 ±0.004
SVR	0.16 ±0.01	0.10 ±0.01	0.14 ±0.01	0.22 ±.001
KNNR	0.26 ±0.22	0.08 ±0.01	0.12 ±0.01	0.20 ±0.01
LRR	0.16 ±0.01	0.08 ±0.01	0.12 ±0.01	0.30 ±0.14
BRR	0.16 ±0.01	0.08 ±0.02	0.12 ±0.01	0.22 ±0.01
Baseline-Random	0.50 ±0.04	0.51 ±0.07	0.22 ±0.03	2.51 ±0.12
Baseline-Rule-based)	0.45 ±0.03	0.52 ±0.07	0.20 ±0.02	2.37 ±0.13

Table 6.3 – Results for VFOA duration prediction for speaker VFOA (SVFOA) and listener VFOA when an agent speaks (LVFOA-AS). Values represent MAE.

Results for the VFOA Target Predictor

Table 6.4 shows results for VFOA Target Prediction for speakers (SVFOA) and listeners (LVFOA) in the AMI and MPR datasets.

The results from Table 6.4 show that concerning the prediction of speaker VFOA targets on AMI, a maximum micro-average F1 value of 0.64 is achieved with the XGB, RF and SVM algorithms. For listeners in the AMI, a maximum micro-average F1 value of 0.87 is obtained via the XGB algorithm. In the MPR, for speakers, a maximum micro F1-Average of 0.77 is observed via XGB and LR algorithms. For listeners in MPR, the XGB algorithm achieves a maximum micro-average F1 value of 0.66. The results show that for both speakers and listeners in the AMI and MPR datasets, the proposed approach returns better results than the baselines.

For speaker VFOA target prediction, the results are significant at $p < 0.05$ (p=0.0010 on AMI and p=0.0006 on MPR for baseline-random, and p=0.0008 on AMI and p=0.0003 on MPR for baseline-rule-based). For listeners, the results are also significant at $p < 0.05$ (p=0.000015 on AMI and p=0.000062 on MPR for baseline-random, and p=0.0000015 on AMI and p=0.000035 on MPR for baseline-rule-based)

For speakers, the performance VFOA Target predictor is better on MPR since in MPR the number of possible VFOA target labels (5) is lower compared to those in AMI (6). Furthermore, in the MPR, persons can join and leave the meeting at any time which further decreases the number of possible VFOA target labels. For listeners, the results for the AMI dataset are better compared to MPR. The possible reason can be the sitting arrangement in AMI where all the participants sit around a table and can directly look at the speaker, resulting in lower number of head movements. In MPR, two or three speakers stand side by side looking at the robot, hence if one of the humans speaks, the other participants have to move their heads by 90 degree in order to look at the speaker. In some cases listeners in MPR cannot look at speaker at all because there is a human standing between them.

Results for the VFOA Scheduler

The results concerning the VFOA scheduler for speakers (SVFOA) and listeners (LVFOA-AS) in the AMI and MPR datasets are depicted in Table 6.5.

The results show that for AMI, a maximum accuracy of 18.07% is achieved with LR for speaker VFOA. For listeners in the AMI, XGB achieves a maximum accuracy of 68.51%. For MPR, a maximum accuracy of 50.34% is returned via XGB algorithm for speakers. For listeners in MPR, this value decreases to 45.32% obtained via the SVM algorithm. These results show that the absence of frequent patterns observed in the

	AMI Dataset		**MPR Dataset**	
Algorithm	**Speaker**	**Listener**	**Speaker**	**Listener**
XGB	**0.64 ±0.02**	**0.87 ±0.03**	**0.77 ±0.01**	0.65 ±0.01
RF	**0.64 ±0.04**	0.86 ±0.02	0.76 ±0.01	0.64 ±0.01
MLP	0.61 ±0.03	0.86 ±0.03	0.74 ±0.01	0.60 ±0.01
SVM	**0.64 ±0.04**	0.85 ±0.02	0.76 ±0.01	0.65 ±0.02
KNN	0.60 ±0.03	0.84 ±0.01	0.75 ±0.01	0.63 ±0.01
LR	0.63 ±0.04	0.85 ±0.02	**0.77 ±0.01**	**0.66 ±0.02**
NB	0.62 ±0.01	0.85 ±0.01	0.38 ±0.02	0.32 ±0.04
Baseline-random	0.45 ±0.02	0.26 ±0.01	0.30 ±0.11	0.48 ±0.01
Baseline-rule-based	0.53 ±0.02	0.23 ±0.01	0.70 ±0.01	0.32 ±0.01

Table 6.4 – Results for VFOA target prediction for speaker VFOA (SVFOA) and listener VFOA when an agent speaks (LVFOA-AS). Values represent Micro Average F1.

	AMI Dataset		**MPR Dataset**	
Algorithm	**Speaker**	**Listener**	**Speaker**	**Listener**
XGB	15.97 ±5.18	**68.51 ±9.43**	**50.34 ±2.17**	40.86 ±4.56
RF	17.84 ±7.05	64.16 ±9.16	47.04 ±2.07	38.14 ±4.05
MLP	11.11 ±5.80	63.41 ±9.97	44.87 ±3.72	32.85 ±5.98
SVM	15.34 ±6.42	57.05 ±7.33	50.09 ±1.65	**45.32 ±5.54**
KNN	15.98 ±6.22	57.90 ±6.98	47.34 ±1.20	37.79 ±4.17
LR	**18.07 ±6.87**	55.00 ±9.20	50.26 ±2.23	43.72 ±5.14
NB	11.24 ±6.76	54.63 ±9.29	13.95 ±2.35	0.02 ±0.01
Baseline-random	3.32 ±3.37	12.43 ±2.26	19.68 ±2.35	21.79 ±1.89
Baseline-rule-based)	4.78 ±1.35	27.84 ±7.39	43.92 ±2.94	10.95 ±1.17

Table 6.5 – Results for VFOA Scheduler for speaker VFOA (SVFOA) and listener VFOA when an agent speaks (LVFOA-AS). Values represent accuracy.

datasets does not allow to replicate exactly the VFOA turn schedules. Nevertheless, the results show that the proposed heuristic approach outperforms both baselines. The results are significant at $p < 0.05$ (p=0.0036 on AMI and p=0.0006 on MPR for baseline-random, and p=0.00002 on AMI and p=0.0034 on MPR for baseline-rule-based) for speakers. The results are also significant for listeners at $p < 0.05$ (p=0.000043 on AMI and p=0.00081 on MPR for baseline-random, and p=0.000043 on AMI and p=0.00039 on MPR for baseline-rule-based) These results show that the absence of frequent patterns observed in the dataset does not allow to replicate exactly the VFOA turns schedules, hence the accuracy values achieved for VFOA scheduler are not very high.

6.7.2 Results for Listener VFOA when a Human Speaks (LVFOA-HS)

The section presents the results for VFOA target predictor, and VFOR turns, duration and scheduling predictor models for LVFOA-HS.

Results for the AMI Dataset

Table 6.6 depicts the results achieved on AMI for VFOA target prediction, number of VFOA turn prediction, VFOA duration prediction and VFOA scheduling for LVFOA-HS.

Results show that for prediction of listener VFOA targets on AMI, the SVM algorithm returns a maximum micro-average F1 value of 0.88. The results are significant at $p < 0.05$ (p=0.00000010 for baseline-random, and p=0.00000051 for baseline-rule-based).

Concerning the number of VFOA turns prediction, minimum MAE value of 0.29 is achieved with the LR algorithms. The results are significant at $p < 0.05$ with respect to both baselines (p=0.002 for baseline-random, p = 0.0016 for baseline-rule-based).

Concerning VFOA duration prediction, minimum MAE value of 0.05 is achieved with the LR algorithm. The results are significant at $p < 0.05$ with respect to both baselines (p=0.000078 for baseline-random, and p=0.000062 for baseline-rule-based).

Finally, for VFOA scheduling prediction, maximum accuracy of 71.60% is obtained via the RF algorithm. For all algorithms, the results are significant at $p < 0.05$ with respect to both baselines (p=0.000011 for baseline-random, and p=0.0000021 for baseline-rule-based).

Algorithm for VFOA Target Prediction	**VFOA Target Prediction (Micro Average F1)**	**Number of VFOA Turns (MAE)**	**VFOA Duration (MAE)**	**VFOA Scheduling (Accuracy)**
XGB	0.86 ±0.01	0.37 ±0.10	0.06 ±0.01	68.88 ±7.59
RF	0.84 ±0.01	0.56 ±0.14	0.06 ±0.01	63.41 ±8.20
MLP	0.85 ±0.01	0.36 ±0.10	0.06 ±0.01	69.08 ±8.00
SVM	**0.88 ±0.03**	0.31 ±0.09	0.06 ±0.01	71.26 ±7.95
KNN	0.85 ±0.01	0.43 ±0.12	0.06 ±0.01	66.18 ±7.35
LR	0.87 ±0.02	**0.29 ±0.08**	**0.05 ±0.01**	**71.60 ±7.68**
NB	0.84 ±0.01	0.73 ±0.27	0.06 ±0.01	58.37 ±12.01
Baseline-random	0.26 ±0.01	0.41 ±0.02	0.52 ±0.07	12.43 ±2.26
Baseline-rule-based	0.23 ±0.01	0.27 ±0.07	0.53 ±0.07	27.84 ±7.39

Table 6.6 – Results for listeners (LVFOA-HS) on AMI

Results for the MPR Dataset

Table 6.7 contains the results achieved on MPR for VFOA target prediction, number of VFOA turn prediction, VFOA duration prediction and VFOA scheduling for LVFOA-HS.

Results depict that concerning the prediction of speaker VFOA targets on MPR, the SVM algorithm returns a maximum micro-average F1 value of 0.65. The results are significant at $p < 0.05$ (p=0.0001 for baseline-random, and p=0.0000086 for baseline-rule-based).

Algorithm for VFOA Target Prediction)	**VFOA Target Prediction (Micro Average F1)**	**Number of VFOA Turns (MAE)**	**VFOA Duration (MAE)**	**VFOA Scheduling (Accuracy)**
XGB	0.64 ±0.02	0.52 ±0.08	0.18 ±0.18	43.96 ±5.39
RF	0.58 ±0.01	0.56 ±0.08	0.23 ±0.01	34.03 ±5.17
MLP	0.64 ±0.02	0.53 ±0.08	0.18 ±0.01	42.98 ±5.12
SVM	**0.65 ±0.02**	0.49 ±0.07	**0.17 ±0.01**	**46.51 ±5.43**
KNN	0.64 ±0.01	0.53 ±0.08	0.19 ±0.01	41.39 ±4.58
LR	**0.65** ±0.02	**0.51 ±0.08**	**0.17 ±0.01**	46.19 ±5.56
NB	0.40 ±0.11	0.64 ±0.05	0.39 ±0.02	5.78 ±2.11
Baseline-random	0.48 ±0.01	0.73 ±0.05	2.51 ±0.12	21.79 ±1.89
Baseline-rule-based	0.32 ±0.01	0.52 ±0.05	2.37 ±0.13	10.95 ±1.17

Table 6.7 – Results for listeners (LVFOA-HS) on MPR

For number of VFOA turns prediction, minimum MAE value of 0.51 is achieved with the LR algorithms. The results are significant at $p < 0.05$ (p=0.002 for baseline-random) but not significant for baseline-rule-based (p = 0.18).

Concerning VFOA duration prediction, minimum MAE value of 0.17 is achieved with the SVM and LR algorithms. The results are significant at $p < 0.05$ (p=0.0000014 for baseline-random, and p=0.0000023 for baseline-rule-based).

Finally, for VFOA scheduling prediction, maximum accuracy of 46.51% is obtained via the SVM algorithm. The results are significant at $p < 0.05$ (p=0.00071 for baseline-random, and p=0.000060 for baseline-rule-based).

6.7.3 Comparison for listener VFOA between Agent (LVFOA-AS) and Human Speakers (LVFOA-HS

The Table 6.8 contains results for the significance test conducted for the comparison of different tasks performed LVFOA-AS and LVFOA-HS behaviour generation models.

LVFOA-HS do not use utterance duration, DA and DA end time in the feature set for training machine learning models.

Task	**AMI**	**MPR**
VFOA Target Prediction	0.292	0.404
VFOA Duration Predictor	0.038	0.046
VFOA Turns Prediction	0.150	0.432
VFOA Scheduling	0.032	0.050

Table 6.8 – Results for significance test for performance comparison of LVFOA-AS and LVFOA-HS

The results show that concerning VFOA target prediction for both AMI and MPR datasets, the performance difference is not significant at $p < 0.05$ (p=0.292 for AMI, and p= 0.404 for MPR). Therefore, utterance duration, DA and DA end time do not significantly improve the performance of VFOA target prediction model.

Concerning VFOA duration prediction, in LVFOA-AS, fixed values i.e. average durations for person/objects from AMI and MPR datasets are used. The results show that in LVFOA-HS that uses average duration values from datasets significantly outperform machine learning based VFOA duration prediction model exploited in LVFOA-AS at $p < 0.05$ (p=0.038 for AMI, and p= 0.046 for MPR). The reason for the better performance of LVFOA-HS is that actual average duration values from datasets are being exploited instead of estimated values predicted by machine learning based prediction model. However a downside to this approach is that average VFOA durations vary by dataset, e.g. in AMI average turn duration value for listener when VFOA target is a person is 2,370 milliseconds which decreases to 441 milliseconds for objects/other targets. On the other hand in MPR, average turn duration for listeners when VFOA target is a person is 756 milliseconds and for objects is 550 milliseconds. Hence, VFOA duration seems to depend on the dataset and cannot be generalized.

For VFOA turn prediction, no significant performance difference is observed between LVFOA-AS and LVFOA-HS. For LVFOA-HS the number of turns depend upon the average turn duration values for participants and objects. The performance difference between the LVFOA-AS and LVFOA-HS is not significant at $p < 0.05$ (p=0.150 for AMI, and p= 0.432 for MPR).

The performance difference for VFOA scheduling algorithms for two listener VFOA model ((i)LVFOA-AS, (ii) LVFOA-HS) is significant for both AMI and MPR datasets at $p <= 0.05$ (p=0.032 for AMI, and p= 0.050 for MPR). VFOA scheduling depends upon number of turns and VFOA targets, and the number of VFOA turns depend upon the VFOA duration. A possible reason for better performance of LVFOA-HS compared to

LVFOA-AS can be that the VFOA duration prediction for LVFOA-HS is significantly better than LVFOA-AS, the error from the VFOA duration prediction of LVFOA-AS propagates to the VFOA scheduling for LVFOA-AS.

6.8 Discussion & Perspectives

The results provide some useful insight into the performance of individual machine learning and heuristic models for generating VFOA turns, assigning duration to each VFOA target in an utterance and setting target for speaker and listener VFOA. The models can be combined to create end-to-end VFOA behaviour generation models for speakers and listeners in multimodal, multiparty human-agent interaction.

The results depict that machine learning based models can be exploited to create VFOA behaviour generation systems that outperform the baselines. It is further observed that in some cases, machine learning algorithms are not suited to perform specific VFOA sub-tasks. For instance, for VFOA scheduling, owing to a high number of unique sequences, the scheduling patterns cannot be learned via machine learning models. In such cases heuristic approaches can be employed to perform scheduling.

Another difficult arises when values for features used for offline training of machine learning models are not available or are consistently updating during real interactions. For instance, the value of utterance duration is consistently updated during an utterance and the final value is not known before an utterance ends. Furthermore, the values for DA and DA end time can also not be predicted before the utterance ends. In such cases either there can be three solutions: (i) an estimated value can transmitted to the model, (ii) model can be trained without the features, and (iii) a heuristic approach can be used. The three options are exploited for various sub-tasks in SVFOA, LVFOA-AS, and LVFOA-HS

For speaker VFOA (SVFOA), the values for utterance duration, DA and DA end time are estimated from agent response when an agent is a speaker. These estimated values are transmitted to other agents involved in an interaction and acting as listeners (LVFOA-AS). For the scenarios where a human speaks and agents listen (LVFOA-HS), listening agents cannot obtain estimated values for utterance duration since the human response is not known hence the length of the duration of human utterance cannot be estimated before the utterance ends. The estimated values for DA and DA end time are also not present for LVFOA-HS. In the LVFOA-HS model, machine learning based VFOA-target prediction sub-model is trained without utterance duration, DA and DA end-time. For the number of VFOA turns, VFOA duration and VFOA scheduling models

in LVFOA-HS, a heuristic approach is used. The results show that with the proposed solutions, all the three VFOA behaviour generation models, i.e. SVFOA, LVFOA-AS, and LVFOA-HS outperform baselines.

The results presented in this chapter only show the performance of individual models and do not evaluate the whole VFOA generation system. In order to evaluate the performance of the combined models the proposed VFOA behaviour generation model has to be implemented in agents participating in real or virtual interactions. The next chapter explains the process of system implementation and the evaluation of overall VFOA behaviour generation model via user surveys.

CHAPTER 7

EVALUATION OF THE VISUAL FOCUS OF ATTENTION GENERATION MODEL

The objective evaluation of the performance of individual sub-models for VFOA behaviour generation i.e. number of VFOA turns predictor, VFOA duration predictor, VFOA target predictor and VFOA scheduler is presented in the previous chapter. However, to evaluate the performance of the overall VFOA behaviour generation model i.e. how it could be actually perceived when implemented in real interactions, none of the traditional performance metrics can be used. In this chapter, overall VFOA behaviour generation models is implemented and user surveys are performed to evaluate the user perception of the proposed model compared with the baseline models and real VFOA behaviour.

The chapter is divided into 4 main sections. Section 7.1 explains experimental protocol while section 7.2 discusses model implementation and the steps performed to generate videos for user evaluation. Section 7.3 analyses the results obtained. Finally, discussion and perspectives are presented in section 7.4

7.1 Experimental Protocol

This section contains information on hypotheses, and scenarios used to implement the overall VFOA behaviour generation model along with the user surveys designed to evaluate the overall perception of the VFOA behaviour generation models for speakers and listeners.

7.1.1 VFOA Generation Tasks

In Section 6.4, we identified four tasks that are required to be performed for end-to-end VFOA behaviour generation for speakers and listeners. The tasks are labelled as: (i) number of VFOA turns prediction, target of VFOA per turn, (ii) duration of each VFOA target per turn, (iii) scheduling of the VFOA turns

7.1.2 VFOA Behaviour Generation Models

Four independent models are proposed to solve the tasks. The models for speakers and listener VFOA behaviour generation have been detailed in Section 6.5. For listener VFOA behaviour generation, two models are proposed: listener VFOA behaviour generation when an agent speaks (LVFOA-AS) and listener VFOA behaviour generation model when a human speaks (LVFOA-HS). Due to COVID 19, the access to the laboratories were limited and during the evaluation period, enough number of human participants were not available. Therefore experiments are proposed via intelligent agents and no human participant is directly involved. Hence, only the listener VFOA behaviour generation model when the speaker is an agent (LVFOA-AS) is implemented in this experiment, in addition to speaker VFOA behaviour generation model (SVFOA). As a reminder, the sub-models for VFOA behaviour generation model are:

Number of VFOA Turns prediction. Number of VFOA Turns Predictor is a machine learning based sub-model that predicts the number of VFOA turns during the current DA. A VFOA turn refers to a stretch of VFOA towards a single target which can be either a person or an object.

VFOA Target Predictor. VFOA target predictor selects the participants or objects in focus of an interaction participant per DA.

VFOA Duration Predictor. VFOA Duration Predictor is also a machine learning based sub-model which predicts the overall duration for VFOA target in focus per DA.

VFOA Scheduler. Turn Scheduling and Final VFOA behaviour. VFOA schedular is a heuristic-based model which schedules the VFOA turns determining the final VFOA for a meeting participant.

7.1.3 Experiments for Speakers and Listeners

During an interaction, an intelligent agent can either speak or listen to the other meeting participants. Therefore, two models have to be evaluated.

Speaker VFOA Behaviour Generation

In speaker VFOA generation experiments, VFOA is generated for an intelligent agent that pronounces an utterance. An utterance can be addressed to a single addressee or a group. A speaker can look at different participants or objects during an utterance. In addition, the duration for which a speaker looks at a particular target vary between participants and objects. Though an utterance can contain multiple dialogue acts (DAs) and vice versa. The VFOA is generated at the level of a DA.

Listener VFOA Behaviour Generation

In listener VFOA behaviour generation attention, VFOA behaviour is generated for intelligent agents acting as direct or indirect listener. A direct listener is a listener to which an utterance is directly addressed whereas indirect listeners, also known as over-hearers are the participants not directly addressed by a speaker but that are able to hear the utterance. The VFOA behaviour for listeners is also generated at the level of a dialogue act. As a reminder, for listener VFOA generation, only the listener VFOA behaviour generation model when an agent speaks (LVFOA-AS) is implemented.

7.1.4 Baselines

The baselines used for VFOA generation for speakers and listeners are the same as those exploited in Section 6.6.2.

Baseline 1: Baseline-Random

In baseline-random, the speaker VFOA behaviour is generated randomly. The number of turns per utterance is randomly selected according to the overall probability of the number of turns per utterance in the corresponding datasets. The VFOA duration per turn is calculated by dividing the number of turns per utterance by the total utterance duration. The turn direction is selected randomly from the list of listeners and objects in a round-robin way according to the default distribution of VFOA directions in the corresponding dataset.

Baseline 2: Baseline-Rule-based

In baseline-rule-based, VFOA behaviour is generated on the basis of a set of rules:

1. The number of VFOA turns is generated as a linear function of VFOA duration using a linear regression algorithm.

2. The VFOA duration per turn is calculated by dividing the number of turns by the total utterance duration.

3. **Turn direction for speakers**: (i) if the addressee of the turn is Group then the participant or object for VFOA target is randomly selected from the list of addressees or objects in a round-robin way, (ii) if the addressee is an individual, the VFOA target for the first turn is the individual that is being addressed. For the remaining turns, participants or objects for VFOA target are randomly selected from the list of addressees or objects in a round-robin way.

   Turn direction for listener: (i) if an agent is directly addressed, set the speaker as the VFOA target for the agent, (ii) if the addressee of an utterance is a group or if an agent is addressed indirectly, set speaker as VFOA target of for first turn, then for the remaining VFOA turns randomly select VFOA target in a round-robin way.

Real VFOA behaviour generation model which is based on real values from the actual meeting scenarios is also implemented in order to compare the proposed VFOA behaviour generation model with the actual VFOA behaviour of meeting participants.

7.1.5 Hypotheses

The overall goal of this experiment is to evaluate how the proposed VFOA behaviour generation models is perceived when compared to baseline-random, baseline-rule-based, and real VFOA behaviour generation model, in a virtual environment, in multiparty interaction. In this regard, we define a total of 4 hypotheses to evaluate how each of the VFOA behaviour generation sub-models are perceived in the overall generated behaviour. Each hypothesis is divided into two parts: s and l for the experiments where an intelligent agent acts respectively as a speaker or direct/indirect listener. Sub-hypothesis s and l are further divided into two parts: bl and rb where the proposed VFOA behaviour generation model is compared respectively with baselines and real VFOA behaviour.

The hypotheses are summarized in table 7.1. The hypothesis h1(s,l) is proposed to evaluate the performance of the Number of VFOA turn predictor sub-model for speakers

Hypothesis	Description
h1(s)	For speakers, the number of VFOA turns predicted by the proposed model should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour
h1(l)	For listeners, the number of VFOA turns predicted by the proposed model should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour
h2(s)	The VFOA duration predicted by the proposed model should should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour, while the agent is speaking
h2(l)	The VFOA duration predicted by the proposed model should should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour, while the agent is listening
h3(s)	The VFOA targets predicted by the proposed model should should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour, while the agent is speaking
h3(l)	The VFOA targets predicted by the proposed model should should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour, while the agent is listening
h4(s)	For speakers, the overall VFOA generation behaviour should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour
h4(l)	For listeners, the overall VFOA generation behaviour should be perceived more natural than baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour

Table 7.1 – Hypotheses for comparing proposed approach with baselines

and listeners. The hypotheses h2(s,l), and h3(s,l) evaluate the VFOA duration predictor and VFOA target prediction models for speakers and listeners. Finally, the hypothesis h4(s,l) estimates the overall performance of VFOA behaviour generation including VFOA scheduling for both speakers and listeners. Each of the hypotheses for speakers and listeners are further divided into two parts: (bl) which states that the proposed models should be perceived as more natural than the baselines, and (rb) the proposed should be perceived as natural as the real VFOA behaviour.

7.1.6 Scenarios and Videos

The proposed VFOA behaviour generation models can be implemented in intelligent virtual agents as well as robots with which humans can interact at real time. However, due to COVID 19, the access to the logistics and human resources required for real-time experiments was limited. Hence, the VFOA models are implemented only in the form

of intelligent virtual agents and interaction videos are recorded and evaluated via user surveys.

The users who evaluate the proposed VFOA behaviour generation model with respect to baseline-random, baseline-rule-based, and real VFOA behaviour are presented with pairs of videos. In each pair, one video has VFOA behaviour generated via the proposed VFOA model while the VFOA behaviour in the other video is generated via either baseline-random, baseline-rule-based or contains real VFOA behaviour from meeting scenarios.

Each video in a pair contains one of 6 meeting scenarios. AMI [Carletta, 2007] and MPR [Funakoshi, 2018] are the only datasets that contain all the features used to train the proposed VFOA behaviour generation models. However, scenarios are chosen from AMI because (i) they have a higher number of participants (4) than in MPR (3) (ii) for the AMI, the meeting scripts or the text of the meetings are available in English whereas in MPR the meeting scripts are in Japanese and are not publicly available. The recorded meeting scenarios are different than those used to train the VFOA behaviour generation models. The details of the 6 meeting scenarios along with the meeting ids from the AMI corpus, and the index numbers of the utterances are presented in Appendix C. All the indexes that only contain a non-verbal information such as laughter, cough, etc. without any textual information are not included in the experiments.

In each video, one of the 6 meeting scenarios is played by a set of virtual agents. In each pair the scenario is the same for the two videos. The only difference among the videos in a pair is the VFOA behaviour generation model. In one of the videos in the pair, VFOA behavior is generated via the proposed VFOA behaviour generation model, whereas in the other video VFOA behaviour is generated via one of the baselines or real VFOA behaviour values. The sequence in which the two videos are played is counter-balanced in order to avoid evaluator's bias towards a particular baseline or towards the proposed approach.

A total of six sets of meeting scenarios are played. To generate VFOA behaviour for each scenario, proposed VFOA behaviour generation model, baseline-random, baseline-rule-based and real VFOA behaviour values are used. Therefore, there can be three pairs for comparison: (i) proposed VFOA model versus baseline-random, (ii) proposed VFOA model versus baseline-rule-based, and (iii) proposed VFOA model versus real VFOA behaviour values. Since the positions of each pair is shuffled, there are 6 pairs per scenario and a total of 36 pairs for 6 scenarios. An example of 6 pairs of videos for one scenario is presented in Table 7.2. The Video *A* corresponds to the first video while the Video *B* is the video which is played in the second position.

Pair No	Video A	Video B
1	Proposed VFOA Model	baseline-random
2	Proposed VFOA Model	baseline-rule-based
3	Proposed VFOA Model	Real VFOA Values
4	baseline-random	Proposed VFOA Model
5	baseline-rule-based	Proposed VFOA Model
6	Real VFOA Values	Proposed VFOA Model

Table 7.2 – An example of 6 pairs of videos for one scenario

The length of the video for each scenarios is limited to approximately 2 minutes to maintain the attention of the evaluators. A user can evaluate as many pairs of videos as s(he) wants, up to 36 pairs of videos (6 per scenario).

7.1.6.1 Survey Questionnaires

For each pair of videos, four questions are asked. Questions are designed to validate the hypotheses proposed in Table 7.1.

A Seven-point (1-7) likert scale [Harpe, 2015] has been used to record answers for each question: Fully Unnatural, Very Unnatural, Unnatural, Acceptable, Natural, Very Natural, Fully Natural. From Fully Unnatural to Fully Natural, options are assigned points from 1 to 7 respectively. The questionnaire are summarized in Table 7.3.

Since there are 36 pair of videos in total and 4 questions are asked per pair, the maximum number of questions is 144.

7.2 Model Implementation and Video Generation Steps

The experiment organization process concerns the development of interaction videos containing VFOA behaviour generated via the proposed VFOA model, baseline-random, baseline-rule-based and real VFOA behaviour values. In each interaction video, intelligent agents play role of the actual participants in the AMI meeting scenarios.

The overall model implementation and video generation process is depicted in Figure 7.1. The experimental steps performed to generate VFOA in each video are explained in the next section.

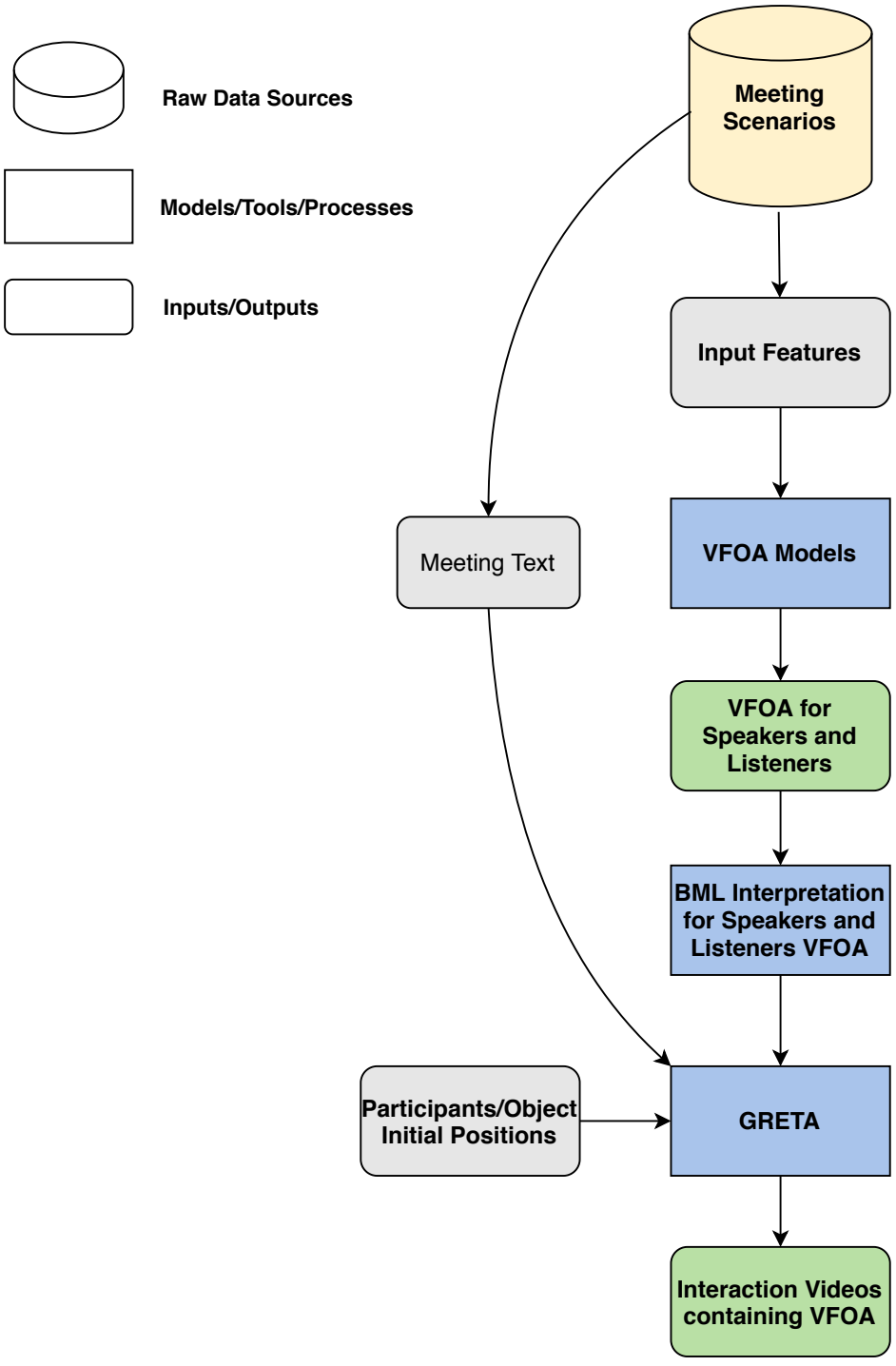

Figure 7.1 – VFOA Behaviour Generation process for experiments

S.No	Question	Hypotheses Validation
1	How natural does the overall visual focus of attention look for the speakers and Listeners in the VIDEO-A and VIDEO-B? (Visual Focus of attention refers to the head and eye movements and the places where the participants look at during an utterance)	h4(s,l)
2	For speakers and listeners in VIDEO-A and VIDEO-B, how natural does the target of the visual focus of attention looks? (Visual focus of attention target is the participants or objects that the speakers and listeners look at during interaction)	h3(s,l)
3	For speakers and listeners in the VIDEO-A and VIDEO-B, how natural does the number of changes in the visual focus of attention look? (Visual focus of attention changes refers to head shifts and gaze shifts (change in target) during an utterance)	h1(s,l)
4	For speakers and listeners in the VIDEO-A and VIDEO-B, how natural does the duration for the visual focus of attention looks? (Visual focus of attention duration refers to the duration for meeting participants when they look at other participants or objects without changing VFOA target.)	h2(s,l)

Table 7.3 – Questions and related hypotheses

7.2.1 Meeting Scenarios

Meeting scenarios contain raw data that is used to train the VFOA models. For the experiments in this research work, meeting scenarios in the form of CSV files from AMI are exploited. The details of meeting scenarios is mentioned in Appendix C.

7.2.2 Input Features

Input features consist of the data used by the proposed VFOA model, baseline-random, and baseline-rule-based. For real VFOA behaviour generation, this step is skipped since real VFOA values are already available in the dataset. For the experimentation, the input features are extracted from CSV files. The input features consist of current and previous speakers, current and previous addressee, dialogue act, utterance duration, DA start time and participants lists.

7.2.3 Generating VFOA via Proposed and Baseline Models

Input features are transmitted to the proposed VFOA model, baseline-random, and baseline-rule-based. VFOA models exploit the input features to predict the VFOA for

speakers as well as for listeners. The output from VFOA model is a CSV file that contains VFOA for speakers and listeners. This step is also skipped for real VFOA, since are already available.

7.2.4 BML Interpretation of VFOA

VFOA for speakers and listeners are converted to BML (Behaviour Markup Language) instructions before it can be transmitted to generate VFOA for intelligent agents in GRETA [Niewiadomski et al., 2009]. GRETA is a comprehensive virtual reality environment developed by Telecom Paristech and is available online [1]. In this research work, the output of VFOA model is stored in a CSV file. The BML interpreter takes this CSV file as input and generates BML values for speaker and listener focus.

7.2.5 GRETA for VFOA Behaviour Generation

VFOA for speakers and listeners generated is integrated into intelligent agents using GRETA. GRETA can play human behaviour such as facial expressions, head and gaze movements, hand gestures, and torso movements in virtual agents. In addition, GRETA is integrated in Unity [2] and OGRE [3] which allows to create different environments and settings for interactions including objects such as laptops, tables, sofa, or lamps etc. that can be used to enhance interaction experience. We used GRETA with OGRE to create virtual agents and interaction environment.

The inputs of GRETA are the text from meeting scenarios, the BML instructions for VFOA behaviour generation and the initial position of participants and objects in the interaction. Four virtual agents are created for these experiments in each video, corresponding to the four participants (PM, ME, ID, UI) in AMI. Laptops are used as objects for each participant. A threshold value of 700 milliseconds is used for turn duration since gaze turns for durations lower than 700 milliseconds result in very abrupt VFOA changes that do not look very natural to human eye.

A view of the interaction video [4] created via GRETA is depicted in Figure 7.2.

[1] https://github.com/isir/greta
[2] https://unity.com/
[3] https://www.ogre3d.org/
[4] https://youtu.be/qpgcaLDy2d8

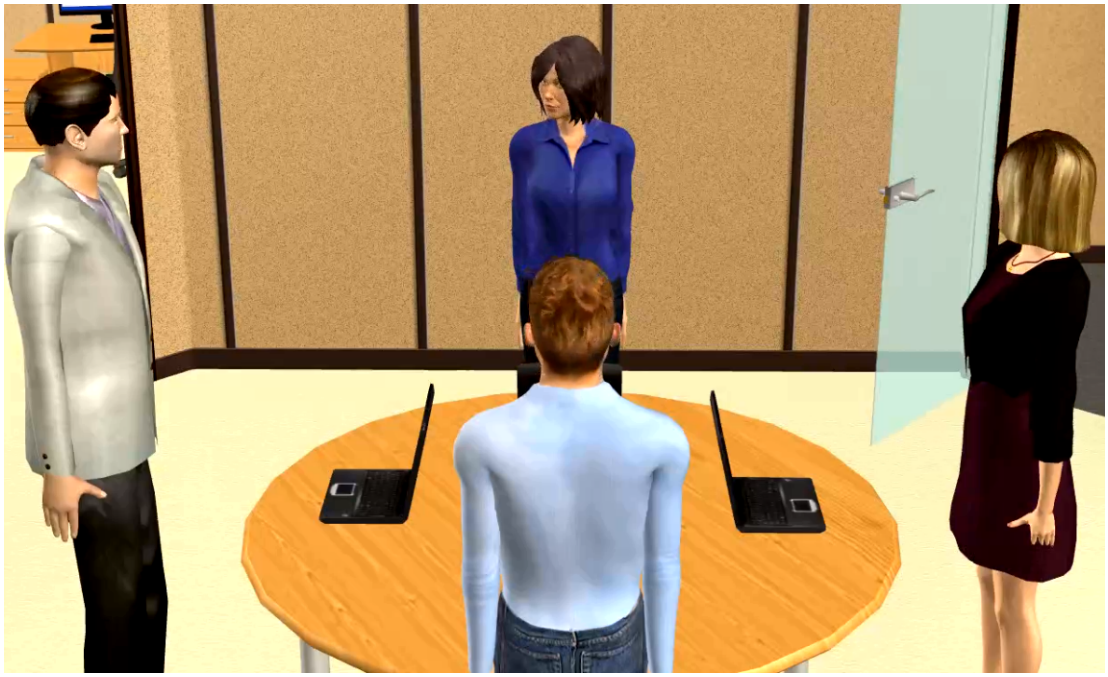

Figure 7.2 – System Implementation via the proposed VFOA generation model

7.3 Results and Analysis

This section presents results of the user evaluation of the proposed VFOA generation model, compared to baseline-random, baseline-rule-based, and real VFOA behaviour values.

A total of 144 participants participated in the survey with a survey completion rate of 28%. On average, each pair of videos is evaluated in 58 surveys.

7.3.1 Results for Speaker VFOA Behaviour Generation

Proposed Speaker VFOA Model vs Baseline-Random

Table 7.4 shows that in the four questions related to speaker VFOA behaviour generation, i.e. overall VFOA generation (question 1), VFOA target (question 2), VFOA turn changes (question 3), and VFOA duration (question 4), the proposed speaker VFOA model outperforms the baseline-random.

For overall VFOA, VFOA targets, VFOA turn changes and VFOA duration, the proposed VFOA model returns average user ratings of respectively 5.56, 5.36, 5.62 and 5.61, which outperform the corresponding random VFOA values of 2.46, 2.38, 2.63, and 2.59. The results are significant at $p < 0.05$ ($p=1.17e^{-11}$ for overall VFOA, $p=1.07e^{-12}$ for VFOA targets, $p=3.31e^{-12}$ for VFOA turn changes, and $p=5.13e^{-13}$ for VFOA durations).

	Overall VFOA (Question 1)	**VFOA Targets (Question 2)**	**VFOA Turn Changes (Question 3)**	**VFOA Duration (Question 4)**
Proposed VFOA Model	5.56 ±0.19	5.36 ±0.12	5.62 ±0.13	5.61 ±0.15
baseline-random	2.46 ±0.20	2.38 ±0.13	2.63 ±0.14	2.59 ±0.13
P-Values	1.17e-11	1.07e-12	3.31e-12	5.13e-13

Table 7.4 – Results of user surveys for the comparison of proposed VFOA behaviour generation model with baseline-random.

	Overall VFOA (Question 1)	**VFOA Targets (Question 2)**	**VFOA Turn Changes (Question 3)**	**VFOA Duration (Question 4)**
Proposed VFOA Model	5.60 ±0.18	5.38 ±0.17	5.62 ±0.13	5.57 ±0.13
baseline-rule-based	3.42 ±0.13	3.29 ±0.11	3.40 ±0.11	3.37 ±1.10
P-Values	6.64e-11	1.42e-11	1.24 e-11	1.29e-12

Table 7.5 – Results of user surveys for the comparison of proposed speaker VFOA behaviour generation model with baseline-rule-based.

Proposed Speaker VFOA Model vs Baseline-Rule-based

Table 7.5 depicts that for speaker VFOA generation, the proposed VFOA model achieves better user ratings compared to rule-based VFOA model for overall VFOA generation (question 1), VFOA target (question 2), VFOA turn changes (question 3), and VFOA duration (question 4). The proposed VFOA model received average user ratings of 5.60, 5.38, 5.62 and 5.57, respectively for overall VFOA, VFOA targets, VFOA turn changes and VFOA duration, which outperform the corresponding rule-based VFOA values of 3.42, 3.29, 3.40, and 3.37. The results are significant at $p < 0.05$ ($p=6.64e^{-11}$ for overall VFOA, $p=1.42e^{-11}$ for VFOA targets, $p=1.24e^{-11}$ for VFOA turn changes, and $p=1.29e^{-12}$ for VFOA durations).

Proposed Speaker VFOA Model vs Real VFOA Behaviour

The proposed speaker VFOA behaviour generation model is also compared with VFOA generated using real values for VFOA targets, VFOA turn changes and VFOA durations as depicted by the results in table 7.6. The results show that user ratings for real VFOA values and the proposed VFOA model are very close to each other. For the proposed VFOA model, user ratings of 5.69, 5.43, 5.66, and 5.70 are obtained respectively for overall VFOA generation (question 1), VFOA target (question 2), VFOA turn changes (question 3), and VFOA duration (question 4), in comparison to 5.73, 5.66, 5.72, and 5.69 achieved using real VFOA. Though slightly better results are obtained with real

	Overall VFOA (Question 1)	**VFOA Targets (Question 2)**	**VFOA Turn Changes (Question 3)**	**VFOA Duration (Question 4)**
Proposed VFOA Model	5.69 ±0.20	5.43 ±0.15	5.66 ±0.14	5.70 ±0.15
Real VFOA Behaviour	5.73 ±0.11	5.66 ±0.08	5.72 ±0.10	5.69 ±0.13
P-Values	0.297	0.368	0.176	0.468

Table 7.6 – Results of user surveys for the comparison of the proposed speaker VFOA behaviour generation model with real VFOA behaviour.

	Overall VFOA (Question 1)	**VFOA Targets (Question 2)**	**VFOA Turn Changes (Question 3)**	**VFOA Duration (Question 4)**
Proposed VFOA Model	4.80 ±0.14	4.51 ±0.12	4.77 ±0.13	4.81 ±0.07
baseline-random	1.86 ±0.19	1.86 ±0.19	1.87 ±0.14	1.86 ±0.18
P-Values	1.94e-12	2.23e-11	1.02e-13	9.38e-14

Table 7.7 – Results of user surveys for the comparison of proposed listener VFOA behaviour generation model with baseline-random

VFOA behaviour values but the overall results are not significant at $p < 0.05$ (p=0.297 for overall VFOA, p=1.42e^{-11} for VFOA targets, p=1.24e^{-11} for VFOA turn changes, and p=1.29e^{-12} for VFOA durations).

7.3.2 Results for Listener VFOA Generation

Proposed Listener VFOA Model vs Baseline-Random

Table 7.7 depicts that for all the four survey questions related to speaker, the proposed listener VFOA model outperforms the baseline-random. For overall VFOA, VFOA targets, VFOA turn changes and VFOA duration, the proposed VFOA model receives average user ratings of 4.80, 4.51, 4.77 and 4.81 respectively, which outperform the corresponding baseline-random VFOA values of 1.86, 1.86, 1.87 and 1.86. The results are significant at $p < 0.05$ (p=1.94e^{-12} for overall VFOA, p=2.23e^{-11} for VFOA targets, p=1.02e^{-13} for VFOA turn changes, and p=9.38e^{-14} for VFOA durations).

Proposed Listener VFOA Model vs Baseline-Rule-based

Table 7.8 shows that for listener VFOA generation, the proposed VFOA model obtains better user ratings compared to rule-based VFOA model for all the four survey questions. The proposed VFOA model gets average user ratings of 4.79, 4.53, 4.77 and 4.74 for overall VFOA, VFOA targets, VFOA turn changes and VFOA duration, which outperform

	Overall VFOA (Question 1)	**VFOA Target (Question 2)**	**VFOA Turn Changes (Question 3)**	**VFOA Duration (Question 4)**
Proposed VFOAModel	4.79 ±0.15	4.53 ±0.16	4.77 ±0.12	4.74 ±0.15
baseline-rule-based	2.77 ±0.12	2.69 ±0.24	2.67 ±0.10	2.69 ±0.10
P-Values	1.61e-11	1.38e-09	4.15e-13	1.37e-12

Table 7.8 – Results of user surveys for the comparison of proposed listener VFOA behaviour generation model with baseline-rule-based

	Overall VFOA (Question 1)	**VFOA Target (Question 2)**	**VFOA Turn Changes (Question 3)**	**VFOA Duration (Question 4)**
Proposed Model	4.82 ±0.13	4.71 ±0.20	4.91 ±0.16	4.69 ±0.14
Real VFOA Behaviour	4.92 ±0.09	4.66 ±0.28	4.91 ±0.10	4.90 ±0.11
P-Values	0.460	0.206	0.449	0.431

Table 7.9 – Results of user surveys for the comparison of proposed listener VFOA behaviour generation model with real VFOA behaviour values

the corresponding rule-based VFOA values of 2.77, 2.69, 2.67, and 2.69, respectively. The results are significant at $p < 0.05$ (p=1.61e^{-11} for overall VFOA, p=1.38e^{-09} for VFOA targets, p=4.15e^{-13} for VFOA turn changes, and p=1.37e^{-12} for VFOA durations).

Proposed Listener VFOA Model vs Real VFOA Behaviour

The results from Table 7.9 show that for listener, user ratings for real VFOA behaviour and the behaviour generated via the proposed VFOA model are very close to each other. For the proposed listener VFOA model, user ratings of 4.82, 4.71, 4.91, and 4.69 are obtained in comparison to 4.92, 4.66, 4.91, and 4.90 achieved using real VFOA, The difference of results at $p < 0.05$ (p=0.460 for overall VFOA, p=0.206 for VFOA targets, p=0.449 for VFOA turn changes, and p=0.431 for VFOA durations).

7.3.3 Result Analysis

The results obtained validate the four hypothesis as mentioned in Section 7.1.5. For both speakers and listeners, the proposed VFOA behaviour generation model is perceived as more natural than the baseline-random and baseline-rule-based, and as natural as the real VFOA behaviour for: (i) VFOA turns prediction, which validates hypothesis h1(s,l), (ii) VFOA duration prediction, verifying hypothesis h2(s,l), (iii) VFOA target

prediction, which confirms hypothesis h3(s,l), and (iv) overall VFOA generation which satisfies hypothesis h4(s,l).

For speakers VFOA, the proposed model achieves user ratings between 5 and 6 for overall VFOA, VFOA target, VFOA turn changes, and VFOA duration, which shows that as per users, VFOA behaviour generated by the proposed model is natural to very natural. These results are similar to the user ratings for real VFOA behaviour values. For listeners, the proposed VFOA model achieves average user ratings between 4-5 which corresponds to acceptable to natural VFOA on the likert scale used for experiments. The results are similar to real VFOA behaviour values for listeners. Significance tests also confirm that there is no significant performance difference between the proposed VFOA model and the real VFOA behaviour for speakers and listeners, which shows that the proposed VFOA model can be used to generate behaviours that are perceived as natural as real VFOA behaviour.

The results show that users ratings for speakers for various tasks are slightly greater than for listeners. For example, the comparison of the proposed VFOA model with baseline-random, show that for the proposed VFOA model, on average the user ratings for speakers for all the four survey questions are between 5-6 as depicted in Table 7.4 whereas the average ratings for listeners via the proposed model is between 4-5 as shown in Table 7.7. Similarly for baseline-random, as depicted in Table 7.4 the average ratings for speakers lies between 2 and 3, where as the average rating for listeners fall between 1 and 2 as shown in Table 7.7. The same observations can be made for the comparison between proposed VFOA model and rule-based VFOA generation as shown in Tables 7.5, 7.8, and the comparison between the proposed VFOA model and real VFOA values as depicted in Tables 7.6, 7.9, where on average listener ratings are 1 point less than speaker ratings. The reason for lower user ratings for listener VFOA model can be attributed to the unavailability of the identity of the virtual characters for the viewers. Since users do not know easily who is being addressed by seeing the video, it is sometimes difficult to estimate if a listener should look at the speaker or one of the other listeners or an object Therefore, the overall user ratings are lower compared to speakers.

The results further depict that for both speakers and listeners, the user ratings for various VFOA tasks i.e. overall VFOA, VFOA turn changes, VFOA target prediction, and VFOA duration prediction are very similar. A reason can be that all the VFOA sub-tasks equally contribute to overall VFOA behaviour generation.

7.4 Discussion and Perspective

This chapter has presented the experimentation and evaluation details of the proposed VFOA behaviour generation models and their comparison with the baseline models via user surveys. The results of user surveys show that the proposed VFOA behaviour generation model is perceived as more natural than the baselines. The proposed model significantly outperforms the baseline models in terms of overall VFOA generation, number of VFOA turn prediction, VFOA duration and VFOA scheduling for both listeners and speakers. The results further show that difference between the perceived naturalness of the proposed VFOA behaviour generation model and real VFOA behaviour is not significant, confirming the hypothesis h1(s,l), h2(s,l), h3(s,l), and h4(s,l).

The proposed VFOA behaviour generation model can therefore be exploited to implement multimodal multiparty interaction scenarios where multiple agents and/or humans communicate with each other.

The proposed VFOA behaviour generation model is currently only implemented with intelligent agents without involving any human participants. Hence the listener VFOA behaviour generation model should be evaluated at real time via interactions that involve human speakers.

Currently the proposed VFOA behaviour generation model is only tested at real-time with 4 meeting participants. To test the proposed VFOA models with more than 4 participants, a database is required which contains more than 4 participants and all the necessary features used to train the proposed VFOA model. An experiment that involves human participants in the interaction, to test the model with more than 4 participants and to implement the proposed VFOA behaviour generation models on robots, would also be required to fully validate the model.

CHAPTER 8

CONCLUSIONS

Intelligent agents are being integrated into everyday life in the form of personal assistant, pedagogical agents, health care professionals, autonomous cars, etc. Moreover, multimodal human-agent interaction has enhanced intelligent agent ability to interact with users via multiple verbal and non-verbal channels such as speech, text, gestures, touch, etc. Traditionally, intelligent agents are deployed in dyadic settings, however, with the advancements in artificial intelligence, agents are lately being employed to solve complex tasks in multiparty settings as well. In this thesis, we study and improve three aspects of an intelligent agent interacting in multiparty, multimodal scenarios : (i) agent perceptive ability to identify the addressee of an utterance, (ii) agent decision making capability to detect turn change and next speaker, and (iii) agent behavior generation capability to generate visual focus of attention (VFOA) behaviour.

This chapter summarizes our contributions and briefly discusses limitations and perspectives.

8.1 Summary of Contributions

8.1.1 Addressee Detection

Addressee detection is one of the most important tasks for smooth dialogue management and turn-taking in human agent interaction. Addressee detection involves identifying the participant or group of participants being addressed by a speaker during the current utterance.

In this thesis, we obtained better results for addressee detection than the baselines on two dataset (AMI and MULTISIMO). The proposed addressee detection model is based on machine learning algorithms. In addition, the impact of various focus encoding techniques on addressee detection is also studied.

Our first contribution (C1) relates to finding the most suitable features for addressee detection. We show that machine learning algorithms trained using a feature set that consists of speaker and listener focus, current and previous dialogue acts (DA), current and previous speaker, previous addressee, you usage, the utterance duration and number of words in the utterance outperforms the baseline addressee detection models.

The proposed model exploits previous addressee as a feature to predict the addressee of the current utterance. During real interactions, the ground-truth value of the previous addressee is not available, rather previous addressee is also predicted. In this regard, two solutions are exploited to tackle this problem. Experiments are performed using ground-truth values of the previous addressee as well without the previous addressee. The results show that at real-time, models trained without predicted previous addressee outperform models that use predicted previous addressee in the feature set. The possible explanation is that the error generated due to wrong predictions for the previous addressee propagates to all the next addressee predictions.

Our second contribution (C2) concerns finding the most suitable model for addressee detection. The results show that the models trained via ensemble learning algorithms, particularly XGBoost achieves highest results for addressee detection. It is further observed that deep learning models such as LSTM performe poorly on both AMI and MULTISIMO datasets for which there can be two possible reasons (i) there is not enough data to train deep learning classifiers, and/or (ii) the addressee detection problem cannot be treated as a sequence problem as. The second reason seems sustained by the results further obtained when comparing the predicted previous addressee and no previous addressee, since all models with no previous addressee performed better than the predicted previous addressee.

Our third contribution (C3) is to compare different focus encoding techniques and select the best technique for addressee detection in multiparty interaction. In this regard, two types of focus encoding techniques are proposed to generate feature vectors for speaker and listener focus: shared focus and one-hot encoded focus. In shared focus encoding scheme, speaker and listener focus is shared among the participants, whereas in one-hot encoded focus, the listener and speaker focus is assigned to the participant or object that is in focus during majority of utterance duration. The results depict that shared focus encoding scheme outperform one-hot focus encoding scheme for addressee

detection in both the datasets. The poor performance of one-hot encoding scheme can be attributed to the fact that there is a very high probability of information loss since the focus is being approximated whereas shared focus encoding captures real focus ratio for all the participants and objects.

8.1.2 Turn Change & Next Speaker Prediction

Detecting when to participate in a conversation is a fundamental part of human-agent interaction. In addition, knowing who should speak next can help an agent to adjust its behaviour such as looking at the next expected meeting participant. In this research work we propose machine learning based turn change and next speaker prediction models that significantly outperform the existing baseline on two datasets (AMI and MPR).

We propose independent machine learning models based on smart feature selection, for turn change and next speaker predictions that respectively predict if the speaker change occurs between two successive utterances, and who should be the speaker of the next utterance.

Our fourth contribution (C4) relates to identifying the most important features for turn change and next speaker prediction. The results show that machine learning models trained via speaker focus, DA, speaker and addressee roles, pause duration, DA start and end time, perform better than the baselines. Among the feature set, pause duration between the utterance and addressee role are the two new features we introduced for next speaker prediction.

Experiments are performed using individual and combined turn change and next speaker prediction models. In combined next speaker prediction model, turn change is predicted in the first step which is consequently included in the feature set to predict next speaker. The results show that for AMI the combined model performs better than the individual models for turn change and next speaker prediction. However, the same results differ on MPR. One of the reasons could be that for AMI, the maximum accuracy of predicted turn change is slightly higher as compared to the predicted turn change accuracy in MPR. Hence in case of AMI, the ratio of propagated error from turn change to next speaker is lesser in AMI than in MPR.

The contributions C5 and C6 concern identifying the best models for turn change and next speaker prediction. We found that similarly to addressee detection, for turn change and next speaker prediction, XGboost outperforms the baselines as well as the remaining machine learning algorithms.

Finally, an ablation study shows that turn change and next speaker models trained when addressee role and pause duration features are included in the feature set outper-

form the models that are trained without these features. This confirms that the newly added features of pause duration and addressee are actual markers for turn change and next speaker prediction.

8.1.3 Visual Focus of Attention Behaviour Generation Models

Visual focus of Attention (VFOA) generation is a behaviour generation task that concerns the visual focus of attention for an agent during speaking and listening. An agent can look at a person or an during an interaction.

In this thesis, we propose a hybrid approach for smart feature selection that relies on the selection of features relevant for VFOA behaviour generation. Four tasks are identified for end-to-end VFOA behaviour generation: VFOA turn prediction, VFOA target prediction, VFOA duration per turn, and VFOA scheduling. The proposed VFOA generation model is divided into three parts: speaker VFOA behaviour generation model, listener VFOA behaviour generation model when an agent speaks, and listener VFOA generation model when a human speaks. The results show that the proposed models, when trained on AMI and MPR, outperform two custom baselines.

Our contribution C7 concerns selecting the most suitable features for speaker and listener VFOA behaviour generation. The identified features are: current and previous speaker roles, current and previous addressee roles, utterance duration, DA, participants list and DA start and end time. All the features for speakers are also available for listener VFOA behaviour generation when an agent speaks. However when a human speaks, the utterance duration, DA and DA end time are not known before the human speaker completes it utterance. Hence, for listener VFOA behaviour generation when a human speaks, utterance duration, DA and DA end time are not included in the feature set.

Our 8th and final contribution (C8) concerns the development of VFOA behaviour generation models. In this regard, to perform the four tasks, individual sub-models are proposed that contribute to the overall VFOA behaviour gaze generation. For speaker VFOA behaviour generation and listener VFOA behaviour generation when an agent speaks, the number of VFOA turn predictor, the VFOA target predictor and the VFOA duration per target predictor models are machine learning models while VFOA scheduling is performed via heuristic based algorithm. For listener VFOA behaviour generation when a human speaks, only VFOA target predictor is based on machine learning algorithms while VFOA turn prediction, VFOA duration prediction and VFOA scheduling is performed via heuristic models. The results show that the four proposed sub-models significantly outperform the custom baseline.

Finally, the proposed speaker and listener VFOA generation models along with the baselines are implemented in intelligent virtual agents interacting in multiparty interaction. Interaction videos are recorded and presented to independent viewers for evaluation. The results show that the proposed VFOA behaviour generation models for speakers and listeners outperform the baselines. In addition, the VFOA behaviour generated via the proposed models is perceived as equally natural as the real VFOA behaviour.

8.2 Limitations & Perspectives

This research proposes improved solutions to three interrelated human agent interaction problems i.e. addressee detection, turn change and next speaker prediction, and VFOA behaviour generation. Though the individual models outperform the baselines, integrating the three models to develop an end-to-end system capable of predicting addressee, detecting turn change and next speaker and then generating VFOA behaviour for agents is a difficult task.

One of the difficulties arises when features used by one model are predicted by another and vice versa. For instance, in the proposed research work, addressee detection model exploits VFOA as input and the VFOA behaviour generation model exploits current addressee to generate VFOA behaviour. In such cases, models have to be trained without some of the features, or with approximated values of features.

Another important limitation is the real-time error propagation. In some cases the proposed models depend on the previously predicted value. For instance for addressee detection, previously predicted addressee is used as input feature. Similarly, the next speaker prediction model uses the predicted turn change values. In both cases, the performance of the proposed model is affected by error propagation. In such cases, it should be tested whether the use of a predicted previous feature value or training a model without the concerned feature obtains better results.

The dynamic nature of data is another major limitation during real-time interaction. Machine learning algorithms are trained via fixed set of features. However, some of the features are dynamic. For instance, the turn change and next speaker models use pause duration which is the time between two successive utterances. The pause duration is a dynamic value since the duration keeps increasing until the next utterance starts. Hence a fixed value of pause duration is not available during real interactions. To tackle this problem, the turn change and next speaker models use a thread that calls turn change and next speaker models after a specific time using the updated value of pause duration.

A major drawback of this approach is that threads can slow down response time of turn change and next change prediction model. Hence a better optimization technique should be designed.

Unavailability of datasets containing all relevant features is another limitation. For instance, existing works show that for turn change and next speaker prediction and for addressee detection, prosody and head gestures can also be exploited. Currently, to the best of our knowledge, none of the existing datasets contain all the relevant features for the three problems discussed in this thesis. Hence, more datasets with a higher number of participants containing all the relevant features for addressee detection, turn change and next speaker prediction, and VFOA behaviour generation models, should be developed.

Due to sanitary conditions enforced because of COVID 19, some of the real-time experiments, such as VFOA behaviour generation for listeners when a human speaks, could not be performed. Thus, listener VFOA behaviour generation model when a human speaks is not evaluated in real interactions. Further experiments should be performed where agents and humans interact with each other in multiparty settings and agents VFOA is generated via the proposed listener VFOA behaviour generation model when humans speak.

Current research work only evaluates the proposed VFOA behaviour generation models implemented via virtual agents. Experiments should be conducted to implement and evaluate the proposed VFOA behaviour generation model in robots.

Finally, the current VFOA behaviour generation experiments include 4 participants and the meeting scenarios are task oriented. Experiments should be performed with more than 4 participants and where participants are engaged in open dialogue. However for such experimentation, datasets with necessary features and more than 4 participants should be developed.

APPENDIX A

Results for Dialogue Act Annotation using Manual vs ASR Transcribed Speech Utterances

	N=1,000		N=2,000		N=3,000		N=4,000		N=5,000	
	ASR	**Manual**	**ASR**	**Manual**	**ASR**	**Manual**	**ASR**	**Manual**	**ASR**	**Manual**
Random Forest	62.13	54.10	61.98	53.84	**63.92**	**55.66**	62.10	54.25	62.00	54.22
Logistic Regression	55.04	50.25	55.04	50.25	57.28	50.57	55.04	50.25	55.04	50.25
ANN	56.55	38.70	57.10	48.16	58.85	47.19	57.27	49.04	55.13	48.14
SVM	56.00	49.9	56.00	49.9	57.58	50.32	56.00	49.9	56.00	49.9
Naive Bayes	43.55	43.01	43.55	43.01	44.68	44.64	43.55	43.01	43.55	43.01

Accuracy in % for different ML algorithms for Bag of Words approach. N refers to first N most occurring words.

	N=8,000		N=15,000		N=25,000		N=30,000	
	ASR	**Manual**	**ASR**	**Manual**	**ASR**	**Manual**	**ASR**	**Manual**
Random Forest	65.55	56.25	65.95	56.55	**66.87**	**57.71**	65.64	56.39
Logistic Regression	61.78	49.51	63.05	50.75	64.37	50.74	63.20	51.82
ANN	64.80	45.17	64.75	47.36	65.68	52.91	65.20	48.25
SVM	32.19	32.01	32.19	32.01	35.00	34.99	32.19	32.01
Naive Bayes	56.46	54.59	57.11	55.75	58.12	56.86	57.28	56.66

Accuracy in % for different ML algorithms for N-Gram (2-7). N refers to first N most occurring grams.

APPENDIX B

Parameters for Addressee Detection Models

Classifier	AMI Parameters	MULTISIMO Parameters
XGB	learning_rate =0.1, n_estimators=140, max_depth=5, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8, objective= 'multi:softmax', nthread=4, scale_pos_weight=1	learning_rate =0.1, n_estimators=130, max_depth=3, min_child_weight=1, gamma=0, subsample=0.6, colsample_bytree=0.5, objective= 'multi:softmax', nthread=4, scale_pos_weight=1
ET	'bootstrap': True, 'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 1000	'bootstrap': True, 'criterion': 'entropy', 'max_features': 'sqrt', 'n_estimators': 200
ADB	Base_estimtor = 'DecisionTree', 'max_features': 30, 'n_estimators':800	Base_estimtor = 'DecisionTree', 'max_features': 30, 'n_estimators':800
MLP	'activation': 'tanh', 'alpha': 0.05, 'hidden_layer_sizes': (100,), 'learning_rate': 'adaptive', 'solver': 'adam'	activation = 'tanh', alpha = 0.0001, hidden_layer_sizes = (50, 100, 50), learning_rate='constant', solver = 'sgd', max_iter = 100
RF	'bootstrap': False, 'criterion': 'gini', 'max_features': 'auto', 'n_estimators': 200	'bootstrap': True, 'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 100
LR	penalty='l2', C =100	penalty='l2', C =0.1
SVM	'C': 100, 'gamma': 0.01	'C': 10, 'gamma': 0.01
NB	No Parameters	No Parameters
KNN	'n_neighbors': 8	'n_neighbors': 9
LSTM	hidden layer neurons = (100, 50), drop Out = 0.5, hidden_activation = relu, final_Activation = softmax, loss = cateorical_crossentropy, optimizer = adam, Bach_size = 4, epochs = 100, callbacks = early Stopping, patience = 20	hidden layer neurons = (50, 25), drop Out = 0.2, hidden_activation = relu, final_Activation = softmax, loss = cateorical_crossentropy, optimizer = adam, Bach_size = 1, epochs = 100, callbacks = early Stopping, patience = 20
Bi-LSTM	hidden layer neurons = (100, 50), drop Out = 0.5, hidden_activation = relu, final_Activation = softmax, loss = cateorical_crossentropy, optimizer = adam, Bach_size = 4, epochs = 100, callbacks = early Stopping, patience = 20	hidden layer neurons = (50, 25), drop Out = 0.2, hidden_activation = relu, final_Activation = softmax, loss = cateorical_crossentropy, optimizer = adam, Bach_size = 1, epochs = 100, callbacks = early Stopping, patience = 20
1D-CNN	hidden layer neurons = (100, 50), kernel_size(3,3) drop Out = 0.5, hidden_activation = relu, final_Activation = softmax, loss = cateorical_crossentropy, optimizer = adam, Bach_size = 4, epochs = 100, calbacks = early Stopping, patience = 20	hidden layer neurons = (50, 25), kernel_size(3,3) drop Out = 0.2, hidden_activation = relu, final_Activation = softmax, loss = cateorical_crossentropy, optimizer = adam, Bach_size = 1, epochs = 100, callbacks = early Stopping, patience = 20

Algorithms parameters for experimentation for addressee detection models

APPENDIX C

Meeting scenarios for VFOA behavior generation experiments

S.No	Meeting Id	Indexes
1	IS1006d	0-20
2	IS1008a	59-83
3	IS1008a	31-56
4	IS1008d	63-97
5	IS1008d	183-208
6	IS1008d	264-290

Meeting scenarios from AMI, chosen for pair of Videos

BIBLIOGRAPHY

Adkar, P. (2013).
Unimodal and multimodal human computer interaction: a modern overview.
Int. J. Comput. Sci. Inf. Eng. Technol, 2(3):1–8.

Admoni, H. and Scassellati, B. (2014).
Data-driven model of nonverbal behavior for socially assistive human-robot interactions.
In *Proceedings of the 16th international conference on multimodal interaction*, pages 196–199.

Admoni, H. and Scassellati, B. (2017).
Social eye gaze in human-robot interaction: a review.
Journal of Human-Robot Interaction, 6(1):25–63.

Advanced Telecom Research Labs, J. (2005).
Freetalk dataset.
`https://freetalk-db.sspnet.eu/`.
[Online; accessed 3-April-2020].

Akker, H. and Akker, R. (2009).
Are you being addressed?-real-time addressee detection to support remote participants in hybrid meetings.
In *SIGDIAL*, pages 21–28.

Akker, R. O. d. and Traum, D. (2009).
A comparison of addressee detection methods for multiparty conversations.
In *SEMDIAL'09*, pages 99–106.

Aldeneh, Z., Dimitriadis, D., and Provost, E. M. (2018).
Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task.

In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6159–6163. IEEE.

Allwood, J. (2000).
An activity based approach to pragmatics.
Abduction, belief and context in dialogue: Studies in computational pragmatics, pages 47–80.

Alpaydin, E. (2020).
Introduction to machine learning.
MIT press.

Aly, A. and Tapus, A. (2012).
Speech to head gesture mapping in multimodal human-robot interaction.
In *Service Orientation in Holonic and Multi-Agent Manufacturing Control*, pages 183–196. Springer.

Amanova, D., Petukhova, V., and Klakow, D. (2016).
Creating annotated dialogue resources: Cross-domain dialogue act classification.
Inform, 26(11.5):36–0.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991).
The hcrc map task corpus.
Language and speech, 34(4):351–366.

Andrist, S., Tan, X. Z., Gleicher, M., and Mutlu, B. (2014).
Conversational gaze aversion for humanlike robots.
In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 25–32. IEEE.

Argyle, M. and Ingham, R. (1972).
Gaze, mutual gaze, and proximity.
Semiotica, 6(1):32–49.

Baba, N., Huang, H.-H., and Nakano, Y. I. (2011).
Identifying utterances addressed to an agent in multiparty human–agent conversations.
In *International Workshop on IVA'11*, pages 255–261.

Baecker, R. M. (1993).
Readings in groupware and computer-supported cooperative work: Assisting human-human collaboration.
Elsevier.

Bakx, I., Van Turnhout, K., and Terken, J. (2003).
Facial orientation during multi-party interaction with information kiosks.
In: INTERACT 2003 Zurich, Switzerland, pages 163–170.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2017).
Multimodal machine learning: A survey and taxonomy.
arXiv preprint arXiv:1705.09406.

Barros, P., Jirak, D., Weber, C., and Wermter, S. (2015).
Multimodal emotional state recognition using sequence-dependent deep hierarchical features.
Neural Networks, 72:140–151.

Bengio, Y., Courville, A., and Vincent, P. (2013).
Representation learning: A review and new perspectives.
IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828.

Bradbury, J. S., Shell, J. S., and Knowles, C. B. (2003).
Hands on cooking: towards an attentive kitchen.
In *CHI'03 extended abstracts on Human factors in computing systems*, pages 996–997.

Breazeal, C. and Scassellati, B. (1999).
How to build robots that make friends and influence people.
In *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No. 99CH36289)*, volume 2, pages 858–863. IEEE.

Bunt, H. (2009).
The dit++ taxonomy for functional dialogue markup.
In *TSMLEDA@AAMAS09*, pages 13–24.

Buswell, G. T. (1935).
How people look at pictures: a study of the psychology and perception in art.

Can, D., Atkins, D. C., and Narayanan, S. S. (2015).
A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations.
In *Sixteenth Annual Conference of the International Speech Communication Association*.

Capurro, C., Nollet, D., and Pletinckx, D. (2015).
Tangible interfaces for digital museum applications.
In *2015 Digital Heritage*, volume 1, pages 271–276. IEEE.

Carletta, J. (2005).
Ami corpus - annotation.
`http://groups.inf.ed.ac.uk/ami/corpus/annotation.shtml`.
Accessed April 1, 2020.

Carletta, J. (2007).
Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus.
Language Resources and Evaluation, 41(2):181–190.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005).
The ami meeting corpus: A pre-announcement.
In *International Workshop on Machine Learning for Multimodal Interaction*, pages 28–39. Springer.

Carpenter, T. and Fujioka, E. (2011).
The role and identification of dialog acts in online chat.
In *Workshop@AAAI'11*.

Cassel, J., Torres, O., and Prevost, S. (1998).
Machine conversations, chapter turn taking versus discourse structure: how best to model multimodal conversation.

Chen, S. and Jin, Q. (2015).
Multi-modal dimensional emotion recognition using recurrent neural networks.
In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56. ACM.

Chen, T. and Guestrin, C. (2016).
Xgboost: A scalable tree boosting system.
In *in SIGKDD*, pages 785–794. ACM.

Chiu, C.-C. and Marsella, S. (2014).
Gesture generation with low-dimensional embeddings.
In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 781–788.

Choi, W. Y., Song, K. Y., and Lee, C. W. (2018).
Convolutional attention networks for multimodal emotion recognition from speech and text data.
In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 28–34, Melbourne, Australia. Association for Computational Linguistics.

De Kok, I. and Heylen, D. (2009).
Multimodal end-of-turn prediction in multi-party meetings.
In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 91–98.

De Mulder, W., Bethard, S., and Moens, M.-F. (2015).
A survey on the application of recurrent neural networks to statistical language modeling.
Computer Speech & Language, 30(1):61–98.

Devillers, L., Rosset, S., Duplessis, G. D., Sehili, M. A., Béchade, L., Delaborde, A., Gossart, C., Letard, V., Yang, F., Yemez, Y., et al. (2015).
Multimodal data collection of human-robot humorous interactions in the joker project.
In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 348–354. IEEE.

Devroye, L., Gyorfi, L., Krzyzak, A., and Lugosi, G. (1994).
On the strong universal consistency of nearest neighbor regression function estimates.
The Annals of Statistics, pages 1371–1385.

Dey, A., Abowd, G., Brown, P., Davies, N., Smith, M., and Steggles, P. (1999).
Towards a better understanding of context and context-awareness.
In *Handheld and ubiquitous computing*, pages 304–307. Springer.

Dey, A. K., Abowd, G. D., and Salber, D. (2001).
A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications.
Human-computer interaction, 16(2):97–166.

Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., and Avrithis, Y. (2013).
Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention.
IEEE Transactions on Multimedia, 15(7):1553–1568.

Fedotov, D., Perepelkina, O., Kazimirova, E., Konstantinova, M., and Minker, W. (2018).
Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data.
In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, pages 1–9.

Ferstl, Y., Neff, M., and McDonnell, R. (2019).
Multi-objective adversarial gesture generation.
In *Motion, Interaction and Games*, pages 1–10.

Foley, G. N. and Gentile, J. P. (2010).
Nonverbal communication in psychotherapy.
Psychiatry (Edgmont), 7(6):38.

Frintrop, S., Rome, E., and Christensen, H. I. (2010).
Computational visual attention systems and their cognitive foundations: A survey.
ACM Transactions on Applied Perception (TAP), 7(1):1–39.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013).
Devise: A deep visual-semantic embedding model.
In *Advances in neural information processing systems*, pages 2121–2129.

Funakoshi, K. (2018).
A multimodal multiparty human-robot dialogue corpus for real world interaction.
Proceedings of the LREC 2018 Special Speech Sessions, pages 35–39.

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (2004).

Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies.
In *ACL'04*, page 669.

Geurts, P., Ernst, D., and Wehenkel, L. (2006).
Extremely randomized trees.
Machine learning, 63(1):3–42.

Ghahramani, Z. (2001).
An introduction to hidden markov models and bayesian networks.
International journal of pattern recognition and artificial intelligence, 15(01):9–42.

Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G., et al. (2011).
Multiple classifier systems for the classification of audio-visual emotional states.
Affective Computing and Intelligent Interaction, pages 359–368.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992).
Switchboard: Telephone speech corpus for research and development.
In *proc. of ICASSP'92*, volume 1, pages 517–520.

Goffman, E. (1981).
Forms of talk. university of pennsylvania publications in conduct and communication.

Goodrich, M. A. and Schultz, A. C. (2008).
Human-robot interaction: a survey.
Now Publishers Inc.

Guntakandla, N. and Nielsen, R. (2015).
Modelling turn-taking in human conversations.
In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction, Stanford CA*.

Gupta, A., Verma, Y., Jawahar, C., et al. (2012).
Choosing linguistics over vision to describe images.
In *AAAI*, page 1.

Gupta, R. (2012).
Human computer interaction–a modern overview.
International Journal Computer Technology Application, 3(5):1736–1740.

Gupta, S., Niekrasz, J., Purver, M., and Jurafsky, D. (2007).
Resolving "you" in multiparty dialog.
In *In Proc. SIGdial*, pages 227–230.

Gutierrez, M. A., D'Haro, L. F., and Banchs, R. (2016).
A multimodal control architecture for autonomous unmanned aerial vehicles.
In *Proceedings of the Fourth International Conference on Human Agent Interaction*, pages 107–110.

Guy, N., Azulay, H., Kardosh, R., Weiss, Y., Hassin, R. R., Israel, S., and Pertzov, Y. (2019).
A novel perceptual trait: Gaze predilection for faces during visual exploration.
Scientific reports, 9(1):1–12.

Harpe, S. E. (2015).
How to analyze likert and other rating scale data.
Currents in Pharmacy Teaching and Learning, 7(6):836–850.

Harvey Sacks, E. A. S. and Jefferson, G. (1974).
A simplest systematics for the organization of turn-taking for conversation.
Language, 50(4):696–735.

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009).
Multi-class adaboost.
Statistics and its Interface, 2(3):349–360.

Hawkins, D. M. (2004).
The problem of overfitting.
J Chem Inform Comput Sci, 44(1):1–12.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998).
Support vector machines.
Intelligent Systems and their applications, 13(4):18–28.

Hilbrink, E. E., Gattis, M., and Levinson, S. C. (2015).
Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction.
Frontiers in psychology, 6:1492.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012).

Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.
IEEE Signal Processing Magazine, 29(6):82–97.

Hochreiter, S. and Schmidhuber, J. (1997).
Long short-term memory.
Neural computation, 9(8):1735–1780.

Hoffman, M. W., Grimes, D. B., Shon, A. P., and Rao, R. P. (2006).
A probabilistic model of gaze imitation and shared attention.
Neural Networks, 19(3):299–310.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013).
Applied logistic regression, volume 398.

Hossain, M. S. and Muhammad, G. (2019).
Emotion recognition using deep learning approach from audio–visual emotional big data.
Information Fusion, 49:69–78.

Hung, H. and Chittaranjan, G. (2010).
The idiap wolf corpus: exploring group behaviour in a competitive role-playing game.
In *Proceedings of the 18th ACM international conference on Multimedia*, pages 879–882.

Ishi, C. T., Machiyashiki, D., Mikata, R., and Ishiguro, H. (2018).
A speech-driven hand gesture generation method and evaluation in android robots.
IEEE Robotics and Automation Letters, 3(4):3757–3764.

Ishii, R., Kumano, S., and Otsuka, K. (2015a).
Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings.
In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 99–106.

Ishii, R., Kumano, S., and Otsuka, K. (2015b).
Predicting next speaker based on head movement in multi-party meetings.
In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2319–2323. IEEE.

Ishii, R., Otsuka, K., Kumano, S., Higashinaka, R., and Tomita, J. (2019).
Prediction of who will be next speaker and when using mouth-opening pattern in multi-party conversation.
Multimodal Technologies and Interaction, 3(4):70.

Ishii, R., Otsuka, K., Kumano, S., Matsuda, M., and Yamato, J. (2013).
Predicting next speaker and timing from gaze transition patterns in multi-party meetings.
In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 79–86.

Jacob, M., Li, Y.-T., Akingba, G., and Wachs, J. P. (2012).
Gestonurse: a robotic surgical nurse for handling surgical instruments in the operating room.
Journal of Robotic Surgery, 6(1):53–63.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003).
The icsi meeting corpus.
In *Proc. of ICASSP'03*, volume 1, pages I–I.

Jayagopi, D. B., Sheiki, S., Klotz, D., Wienke, J., Odobez, J.-M., Wrede, S., Khalidov, V., Nyugen, L., Wrede, B., and Gatica-Perez, D. (2013).
The vernissage corpus: A conversational human-robot-interaction dataset.
In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 149–150. IEEE.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013).
3d convolutional neural networks for human action recognition.
IEEE transactions on pattern analysis and machine intelligence, 35(1):221–231.

Jia, J., Wu, Z., Zhang, S., Meng, H. M., and Cai, L. (2014).
Head and facial gestures synthesis using pad model for an expressive talking avatar.
Multimedia Tools and Applications, 73(1):439–461.

Johnston, M., Chen, J., Ehlen, P., Jung, H., Lieske, J., Reddy, A., Selfridge, E., Stoyanchev, S., Vasilieff, B., and Wilpon, J. (2014).
Mva: The multimodal virtual assistant.

In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 257–259.

Jokinen, K. and Wilcock, G. (2014).
Multimodal open-domain conversations with the nao robot.
In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 213–224. Springer.

Jovanovic, N. (2007).
To whom it may concern-addressee identification in face-to-face meetings.

Jovanovic, N., Akker, R. o. d., and Nijholt, A. (2006).
A corpus for studying addressing behaviour in multi-party dialogues.
LREC'06, 40(1):5–23.

Jovanovic, N. and op den Akker, R. (2004).
Towards automatic addressee identification in multi-party dialogues.
In *SIGdial@HLT-NAACL'04*.

Kaiser, M., Klingspor, V., and Friedrich, H. (1997).
Human-agent interaction and machine learning.
In *European Conference on Machine Learning*, pages 345–352. Springer.

Kanade, T., Cohn, J. F., and Tian, Y. (2000).
Comprehensive database for facial expression analysis.
In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE.

Karray, F., Alemzadeh, M., Abou Saleh, J., and Arab, M. N. (2008).
Human-computer interaction: Overview on state of the art.

Kawahara, T., Iwatate, T., and Takanashi, K. (2012).
Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations.
In *Thirteenth Annual Conference of the International Speech Communication Association*.

Kepuska, V. and Bohouta, G. (2018).
Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home).

In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 99–103. IEEE.

Khatib, O., Yeh, X., Brantner, G., Soe, B., Kim, B., Ganguly, S., Stuart, H., Wang, S., Cutkosky, M., Edsinger, A., et al. (2016).
Ocean one: A robotic avatar for oceanic discovery.
IEEE Robotics & Automation Magazine, 23(4):20–29.

Kim, J., Truong, K. P., Charisi, V., Zaga, C., Evers, V., and Chetouani, M. (2016).
Multimodal detection of engagement in groups of children using rank learning.
In *International Workshop on Human Behavior Understanding*, pages 35–48. Springer.

Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., and Gabbouj, M. (2019).
1-d convolutional neural networks for signal processing applications.
In *ICASSP'19*, pages 8360–8364.

Kleinerman, A., Rosenfeld, A., and Kraus, S. (2018).
Providing explanations for recommendations in reciprocal environments.
In *Proceedings of the 12th ACM conference on recommender systems*, pages 22–30.

Koutsombogera, M. and Vogel, C. (2018).
Modeling collaborative multimodal behavior in group dialogues: the multisimo corpus.
In *LREC-2018*.

Kozima, H. and Ito, A. (1998).
Towards language acquisition by an attention-sharing robot.
In *New methods in language processing and computational natural language learning*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In *Advances in neural information processing systems*, pages 1097–1105.

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., and Held, P. (2013).
Multi-layer perceptrons.
In *Computational Intelligence*, pages 47–81.

Krzywinski, M. and Altman, N. (2013).
Points of significance: Significance, p values and t-tests.

Le Minh, T., Shimizu, N., Miyazaki, T., and Shinoda, K. (2018).
Deep learning based multi-modal addressee recognition in visual scenes with utterances.
In *IJCAI*, pages 1546–1553.

Lee, J., Marsella, S., Traum, D., Gratch, J., and Lance, B. (2007).
The rickel gaze model: A window on the mind of a virtual human.
In *International workshop on intelligent virtual agents*, pages 296–303. Springer.

Li, B., Si, X., Lyu, M. R., King, I., and Chang, E. Y. (2011).
Question identification on twitter.
In *Proc. of CIKM'11*, pages 2477–2480.

Liaw, A., Wiener, M., et al. (2002).
Classification and regression by randomforest.
R news, 2(3):18–22.

Link, S., Barkschat, B., Zimmerer, C., Fischbach, M., Wiebusch, D., Lugrin, J.-L., and Latoschik, M. E. (2016).
An intelligent multimodal mixed reality real-time strategy game.
In *2016 IEEE Virtual Reality (VR)*, pages 223–224. IEEE.

Liu, C., Ishi, C. T., Ishiguro, H., and Hagita, N. (2012).
Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction.
In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 285–292. IEEE.

Makula, P., Mishra, A., Kumar, A., Karan, K., and Mittal, V. (2015).
Multimodal smart robotic assistant.
In *2015 International Conference on Signal Processing, Computing and Control (ISPCC)*, pages 18–23. IEEE.

Malmaud, J., Huang, J., Rathod, V., Johnston, N., Rabinovich, A., and Murphy, K. (2015).
What's cookin'? interpreting cooking videos using text, speech and vision.
arXiv preprint arXiv:1503.01558.

Manawadu, U. E., Kamezaki, M., Ishikawa, M., Kawano, T., and Sugano, S. (2017).

A multimodal human-machine interface enabling situation-adaptive control inputs for highly automated vehicles.
In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1195–1200. IEEE.

Mansimov, E., Parisotto, E., Ba, J. L., and Salakhutdinov, R. (2015).
Generating images from captions with attention.
arXiv preprint arXiv:1511.02793.

Mao, X., Li, Z., and Xue, Y. (2009).
Emotional gaze behavior generation in human-agent interaction.
In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 3691–3696.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005).
The ami meeting corpus.
In *Proc. of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100.

McNeill, D. (1992).
Hand and mind: What gestures reveal about thought.
University of Chicago press.

Melamud, O., Goldberger, J., and Dagan, I. (2016).
context2vec: Learning generic context embedding with bidirectional lstm.
In *20th SIGNLL conference on computational natural language learning*, pages 51–61.

Meshorer, T. and Heeman, P. A. (2016).
Using past speaker behavior to better predict turn transitions.
In *Interspeech*, pages 2900–2904.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In *Advances in neural information processing systems*, pages 3111–3119.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012).
Introduction to linear regression analysis, volume 821.
John Wiley & Sons.

Moon, S., Kim, S., and Wang, H. (2014).
Multimodal transfer deep learning for au-dio visual recognition.
arXiv preprint arXiv:1412.3121.

Morvant, E., Habrard, A., and Ayache, S. (2014).
Majority vote of diverse classifiers for late fusion.
In *Joint IAPR Int. Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 153–162. Springer.

Mroueh, Y., Marcheret, E., and Goel, V. (2015).
Deep multimodal learning for audio-visual speech recognition.
In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2130–2134. IEEE.

Murtagh, F. (1991).
Multilayer perceptrons for classification and regression.
Neurocomputing, 2(5-6):183–197.

Nakano, Y. I. and Ishii, R. (2010).
Estimating user's engagement from eye-gaze behaviors in human-agent conversations.
In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 139–148.

Newell, A. and Card, S. K. (1985).
The prospects for psychological science in human-computer interaction.
Human-computer interaction, 1(3):209–242.

Niewiadomski, R., Bevacqua, E., Mancini, M., and Pelachaud, C. (2009).
Greta: an interactive expressive eca system.
In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400.

Nikola, T. (1898).
Method of and apparatus for controlling mechanism of moving vessels or vehicles.
US Patent 613,809.

O'Connell, D. C. and Kowal, S. (2012).
Dialogical genres: Empractical and conversational listening and speaking.
Springer Science & Business Media.

Ortega, D. and Vu, N. T. (2017).
Neural-based context representation learning for dialog act classification.
In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 247–252, Saarbrücken, Germany. Association for Computational Linguistics.

Park, S. (2018).
Virtual pedagogical agents for english language teaching and learning.
The TESOL Encyclopedia of English Language Teaching, pages 1–9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011).
Scikit-learn: Machine learning in python.
Journal of machine learning research, 12(Oct):2825–2830.

Pelachaud, C. and Bilvi, M. (2003).
Modelling gaze behavior for conversational agents.
In *International Workshop on Intelligent Virtual Agents*, pages 93–100. Springer.

Perera, C., Zaslavsky, A., Christen, P., and Georgakopoulos, D. (2014).
Context aware computing for the internet of things: A survey.
IEEE Communications Surveys & Tutorials, 16(1):414–454.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., and Poggi, I. (2005).
A model of attention and interest using gaze behavior.
In *International Workshop on Intelligent Virtual Agents*, pages 229–240. Springer.

Petukhova, V. and Bunt, H. (2009).
Who's next? speaker-selection mechanisms in multiparty dialogue.
In *Workshop on the Semantics and Pragmatics of Dialogue*.

Poggi, I., Pelachaud, C., and De Rosis, F. (2000).
Eye communication in a conversational 3d synthetic agent.
AI communications, 13(3):169–181.

Ranganathan, H., Chakraborty, S., and Panchanathan, S. (2016).
Multimodal emotion recognition using deep learning architectures.
In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE.

Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. (2015).
Where are they looking?
In *Adv. in Neural Information Processing Systems*, pages 199–207.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013).
Introducing the recola multimodal corpus of remote collaborative and affective interactions.
In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

Rish, I. et al. (2001).
An empirical study of the naive bayes classifier.
In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Ronzhin, A. and Karpov, A. (2005).
Assistive multimodal system based on speech recognition and head tracking.
In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE.

Rosenfeld, A., Agmon, N., Maksimov, O., and Kraus, S. (2017).
Intelligent agent supporting human–multi-robot team collaboration.
Artificial Intelligence, 252:211–231.

Rosenfeld, A., Bareket, Z., Goldman, C. V., LeBlanc, D. J., and Tsimhoni, O. (2015).
Learning drivers' behavior to improve adaptive cruise control.
Journal of Intelligent Transportation Systems, 19(1):18–31.

Sacks, H., Schegloff, E., and Jefferson, G. (1974).
(1974). a simplest systematics for the organization of turn-taking in conversation. language, 50, 696-735.

Salem, M., Lakatos, G., Amirabdollahian, F., and Dautenhahn, K. (2015).
Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust.
In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1–8. IEEE.

Sanchez-Cortes, D., Aran, O., and Gatica-Perez, D. (2011).
An audio visual corpus for emergent leader analysis.

In *Workshop on multimodal corpora for machine learning: taking stock and road mapping the future, ICMI-MLMI*.

Sargin, M. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2008).
Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(8):1330–1345.

Scassellati, B. (1996).
Mechanisms of shared attention for a humanoid robot.
In *Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium*, volume 4, page 21.

Schneider, F. (2007).
European land-robot trial 2006 (elrob).

Scholtz, J., Theofanos, M., and Antonishek, B. (2006).
Development of a test bed for evaluating human-robot performance for explosive ordnance disposal robots.
In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 10–17.

Searle, J. (1969).
Speech Acts: An Essay in the Philosophy of Language.

Sheh, R. K.-M. (2017).
" why did you do that?" explainable intelligent robots.
In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., and Gifford, E. M. (2016).
Extreme gradient boosting as a method for quantitative structure–activity relationships.
Journal of chemical information and modeling, 56(12):2353–2360.

Shutova, E., Kiela, D., and Maillard, J. (2016).
Black holes and white rabbits: Metaphor identification with visual features.
In *HLT-NAACL*, pages 160–170.

Sisbot, E. A. and Alami, R. (2012).
A human-aware manipulation planner.

IEEE Transactions on Robotics, 28(5):1045–1057.

Smit, S. K. and Eiben, A. E. (2009).
Comparing parameter tuning methods for evolutionary algorithms.
In *CEC'09*, pages 399–406.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000).
Dialogue act modeling for automatic tagging and recognition of conversational speech.
Computational linguistics, 26(3):339–373.

Takeuchi, M., Kitaoka, N., and Nakagawa, S. (2003).
Generation of natural response timing using decision tree based on prosodic and linguistic information.
In *Eighth European Conference on Speech Communication and Technology*.

Ten Bosch, L., Oostdijk, N., and De Ruiter, J. P. (2004).
Turn-taking in social talk dialogues: temporal, formal and functional aspects.
In *9th International Conference Speech and Computer (SPECOM'2004)*, pages 454–461.

Thórisson, K. R. (1994).
Face-to-face communication with computer agents. aaai spring symposium on believable agents working notes.

Tipping, M. E. (2001).
Sparse bayesian learning and the relevance vector machine.
Journal of machine learning research, 1(Jun):211–244.

Tong, H., Chen, D.-R., and Peng, L. (2009).
Analysis of support vector machines regression.
Foundations of Computational Mathematics, 9(2):243–257.

Trafton, J. G., Bugajska, M. D., Fransen, B. R., and Ratwani, R. M. (2008).
Integrating vision and audition within a cognitive architecture to track conversations.
In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 201–208.

Tran, Q. H., Zukerman, I., and Haffari, G. (2017).
A hierarchical neural model for learning sequences of dialogue acts.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437, Valencia, Spain. Association for Computational Linguistics.

Traum, D., Rickel, J., Gratch, J., and Marsella, S. (2003).
Negotiation over tasks in hybrid human-agent teams for simulation-based training.
In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 441–448.

Traum, D. R., Robinson, S., and Stephan, J. (2004).
Evaluation of multi-party virtual reality dialogue interaction.
In *LREC'04*, pages 1699–1702.

Traum, D. R., Robinson, S., and Stephan, J. (2006).
Evaluation of multi-party reality dialogue interaction.
Technical report, University of Southern California Marina Del Rey CA Inst For Creative Technologies.

Vertegaal, R. (1998).
Look who's talking to whom.
Mediating Joint Attention in multiparty.

Vertegaal, R., Slagter, R., Van der Veer, G., and Nijholt, A. (2001).
Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes.
In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308.

Vertegaal, R., Van der Veer, G., and Vons, H. (2000).
Effects of gaze on multiparty mediated communication.
In *Graphics interface*, pages 95–102.

Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009).
Canal9: A database of political debates for analysis of social interactions.
In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–4. IEEE.

Wang, W., Yang, Y., Wang, X., Wang, W., and Li, J. (2019).

Development of convolutional neural network and its application in image classification: a survey.
Optical Engineering, 58(4):040901.

Webb, N., Hepple, M., and Wilks, Y. (2005).
Dialogue act classification based on intra-utterance features.
In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, volume 4, page 5. Citeseer.

Weller, C. (2017).
Meet the first-ever robot citizen, a humanoid named sophia that once said it would destroy humans.
Business Insider Nordic. Haettu, 30:2018.

Wells, P. and Deguire, D. (2005).
Talon: A universal unmanned ground vehicle platform, enabling the mission to be the focus.
In *Unmanned Ground Vehicle Technology VII*, volume 5804, pages 747–757. International Society for Optics and Photonics.

Werry, C. C. (1996).
Internet relay chat.
Computer-mediated communication: Linguistic, social and cross-cultural perspectives, pages 47–63.

Wicaksono, I. and Paradiso, J. A. (2017).
Fabrickeyboard: multimodal textile sensate media as an expressive and deformable musical interface.
In *NIME*, volume 17, pages 348–353.

Wiener, N. (1964).
God and Golem, Inc: A Comment on Certain Points where Cybernetics Impinges on Religion, volume 42.
MIT press.

Yarbus, A. L. (1967).
Eye movements during perception of complex objects.
In *Eye movements and vision*, pages 171–211. Springer.

Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. (1989).
Integration of acoustic and visual speech signals using neural networks.
IEEE Communications Magazine, 27(11):65–71.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018).
Human multimodal language in the wild: A novel dataset and interpretable dynamic fusion model.
In *Association for Computational Linguistics*.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009).
A survey of affect recognition methods: Audio, visual, and spontaneous expressions.
IEEE transactions on pattern analysis and machine intelligence, 31(1):39–58.

Zhang, M.-L. and Zhou, Z.-H. (2005).
A k-nearest neighbor based algorithm for multi-label classification.
In *GRC'05*, volume 2, pages 718–721. ACM.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015).
Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.
In *Proc. of the IEEE international conference on computer vision*, pages 19–27.