



**HAL**  
open science

# Modelling interruptions in human-agent interaction

Liu Yang

► **To cite this version:**

Liu Yang. Modelling interruptions in human-agent interaction. Human-Computer Interaction [cs.HC]. Sorbonne Université, 2023. English. NNT : 2023SORUS611 . tel-04606501

**HAL Id: tel-04606501**

**<https://theses.hal.science/tel-04606501v1>**

Submitted on 10 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Modelling interruptions in human-agent interaction

École Doctorale Informatique, Télécommunications et Électronique

**THÈSE DE DOCTORAT  
DE SORBONNE UNIVERSITÉ**

*présentée et soutenue publiquement par*

**Liu YANG**

le 8 Décembre 2023

Directrice de thèse: **Catherine Pelachaud**  
Co-encadrante: **Catherine Achard**

devant le jury composé de :

Chloé CLAVEL, Directrice de Recherche INRIA  
Elisabeth ANDRÉ, Professeure University of Augsburg  
Elisabetta BEVACQUA, Maîtresse de Conférences ENIB  
Frédéric BEVILACQUA, Directeur de Recherche IRCAM, Sorbonne Université  
Philippe BLACHE, Directeur de Recherche CNRS, Aix-Marseille Université  
Catherine PELACHAUD, Directrice de Recherche CNRS, Sorbonne Université  
Catherine ACHARD, Professeure Sorbonne Université

Rapportrice  
Rapportrice  
Examinatrice  
Examineur  
Examineur  
Examinatrice  
Examinatrice



---

## Abstract

Interruptions play a significant role in shaping human communication, occurring frequently in everyday conversations. They serve to regulate conversation flow, convey social cues, and promote shared understanding among speakers. Human communication involves a range of multimodal signals beyond just speech. Verbal and non-verbal modes of communication are intricately intertwined, conveying semantic and pragmatic content while tailoring the communication process. The vocal mode incorporates acoustic features, such as prosody, while the visual mode encompasses facial expressions, hand gestures, and body language.

The rise of virtual and online communication has necessitated the development of expressive communication for human-like embodied agents, including Embodied Conversational Agents (ECA) and social robots. To foster seamless and natural interactions between humans and virtual agents, it is crucial to equip virtual agents with the ability to handle interruptions during interactions.

This manuscript focuses on studying interruptions in human-human interactions and enabling ECAs to interrupt human users during conversations. The primary objectives of this research are twofold: (1) in human-human interaction, analysis of acoustic and visual signals to categorise interruption type and detect when interruptions occur; (2) endow ECA with the capability to predict when to interrupt and generate its multimodal behaviour. To achieve these goals, we propose an annotation schema for identifying and classifying smooth turn exchanges, backchannels, and different interruption types. We manually annotate exchanges in two corpora, a part of the AMI corpus and the French section of the NoXi corpus. After analysing multimodal non-verbal signals, we introduce MIC, an approach to classify the interruption type based on selected non-verbal signals (facial expression, prosody, head and hand motion) from both interlocutors (the interruptee and the interrupter). We also introduce One-PredIT, which utilises a one-class classifier to identify potential interruption points by monitoring the real-time non-verbal behaviour of the current speaker (only interruptee). Additionally, we propose AI-BGM, a generative model to compute the facial expressions and head rotations of ECAs when it is interrupting. Given the limited amount of data at our disposal, we employ transfer learning technology to train our interruption behaviour generation model using the well-trained Augmented Self-Attention Pruning neural network model. One-PredIT and AI-BGM are evaluated with subjective studies.

**Keywords:** Interruption, Turn-taking, Multimodality, Human Behavior Modelling, Embodied Conversational Agents

---

## Acknowledgment

I would like to extend my heartfelt appreciation to my research supervisors, Professor Catherine Pelachaud and Professor Catherine Achard. Their unwavering support and belief in my abilities have been instrumental in enabling me to undertake this project. Their guidance and mentorship throughout this research endeavour have been invaluable.

Working under their supervision has been a tremendous privilege and honour. They have consistently displayed remarkable mentorship qualities, offering guidance and support during both the triumphs and challenges of this journey.

I am indebted to them for their scientific expertise, invaluable advice, and constant motivation. Their unwavering belief in my potential has inspired me to strive for excellence and will continue to guide me long after my PhD. I am deeply grateful for the opportunities they have provided and the impact they have made on my academic journey.

I would like to express my sincere gratitude to our esteemed collaborators from the TAPAS and PANORAMA projects, representing LISN, NAIST, JAIST, SeiKei University, and Augsburg University.

Their invaluable contributions and collaborative efforts have been instrumental in the success of our projects. Working alongside such talented individuals has been an enriching experience, as they have brought diverse perspectives and expertise to the table. I am truly thankful for their dedication, support, and the knowledge they have shared throughout our joint endeavours.

I would also like to thank Ecole Doctorale Informatique, Télécommunications et Electronique (EDITE) for the three-year full scholarship that made this research possible, and Institut des Systèmes Intelligents et de Robotique (ISIR) for providing offices, equipment and a creative research environment.

I consider myself incredibly fortunate to have crossed paths with a multitude of talented collaborators who have not only imparted invaluable knowledge but also made this project an engaging and enjoyable endeavour. I extend my deepest gratitude to Jiyeon Woo, Mireille Fare, Michele Grimaldi, Fabien Boucaud, Lucie Galland, Nezh Yousi, Fajrian Yunus, for their exceptional spirit, boundless inspiration, and unwavering motivation.

I want to thank my three lovely little parrots for their endless companionship and comfort, especially during the covid pandemic.

Lastly, my deepest thanks go to my parents, my mother Wenping, and my father Jianhua, for their unwavering support and encouragement throughout my studies and the six years I have spent in France. Their love and guidance have been instrumental in my journey, and I am forever grateful for their presence in my life.

# Résumé en Français

Ce qui distingue les humains des autres animaux, c'est notre capacité à créer et à utiliser des outils pour simplifier le travail et améliorer la productivité, ainsi que notre puissante capacité de pensée logique et de communication complexe mais cohérente. Alors que certains animaux utilisent des outils, les humains ont poussé l'utilisation des outils à un niveau sans précédent. L'histoire de la civilisation humaine peut être vue comme une histoire du développement technologique, allant de l'utilisation de simples outils en pierre à la métallurgie, puis à la puissance de la vapeur, et enfin à l'ère actuelle des technologies numériques et en réseau (Bogin and Varea [2020]). Les avancées rapides dans la technologie ont propulsé l'humanité dans l'ère virtuelle, et le "humanisme numérique", comme pont entre les humains et les humanoïdes, a émergé ces dernières années et gagné en popularité (Davies [2016], Wagner et al. [2020]). En raison de notre préférence inhérente pour la communication interpersonnelle, les Agents Conversationnels Incarnés (ACI) ont émergé comme une interface émergente de l'Interaction Humain-Ordinateur (IHO). En simulant des formes humaines, les ACI peuvent transmettre des informations par la voix, les gestes, les expressions faciales et les mouvements du corps, rendant la communication plus précise et plus humaine par rapport au simple texte (Lugrin [2021]). Avec les avancées significatives dans la technologie 3D, la technologie XR et la disponibilité commerciale des appareils, les applications des ACI se sont diversifiées. Sans aucun doute, ils font leur chemin des laboratoires dans la vie quotidienne des gens, offrant une assistance dans divers aspects du travail.

## 0.1 Agents Conversationnels Incarnés

Les ACI sont conçus pour simuler une intelligence et un comportement social semblables à ceux des humains, comblant efficacement le fossé entre les humains et les machines (Ruttkay and Pelachaud [2004]). La Figure 1.1 fournit une représentation visuelle de divers exemples d'ACI. La polyvalence des ACI est mise en avant par leur déploiement dans une multitude de domaines. Dans le domaine du service client, ils peuvent assumer le rôle de représentants virtuels sur les sites Web et au sein des applications, aidant habilement les utilisateurs avec des questions,

## 0.2. PRISE DE PAROLE ET INTERRUPTION DANS L'INTERACTION ENTRE AGENTS HUMAINS

---

des problèmes de dépannage et la récupération d'informations. Dans les environnements de soins de santé, les ACI se révèlent précieux en fournissant des informations médicales, en surveillant à distance les patients, en menant des séances de thérapie et en promouvant des choix de mode de vie sain. Dans l'industrie du jeu, les ACI endossent des rôles de personnages interactifs, de guides ou de compagnons, rehaussant le gameplay en le rendant plus captivant et immersif. En outre, à l'instar des assistants virtuels vocaux populaires tels que Siri ou Alexa, les ACI offrent une interface conversationnelle pour effectuer des tâches telles que la définition de rappels, la réponse à des requêtes et le contrôle d'appareils intelligents. L'utilisation des ACI représente un effort concerté des chercheurs pour établir des interactions plus naturelles et engageantes entre les humains et les machines. Ces agents ne sont pas simplement des outils, mais des facilitateurs dynamiques de la communication et de l'engagement, enrichissant divers aspects de notre vie par leurs attributs et capacités semblables à ceux des humains.

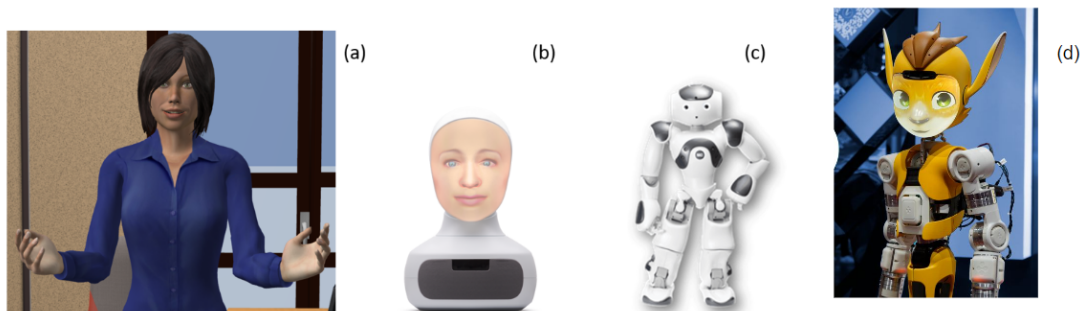


Figure 1 Illustration de (a) Greta, un agent conversationnel incarné (Pelachaud [2015]), (b) le robot social Furhat (Al Moubayed et al. [2013]), (c) un robot humanoïde NAO (Shamsuddin et al. [2011]) et (d) le robot-personnage MIROKI développé par la société Enchanted Tools.

## 0.2 Prise de parole et interruption dans l'interaction entre agents humains

Afin de rendre l'interaction entre les ACI et les utilisateurs humains plus fluide et plus conforme aux habitudes humaines ainsi qu'à la logique comportementale, les chercheurs ont commencé à explorer diverses directions. Ces orientations comprennent la simulation d'apparences réalistes, la garantie de mouvements naturels et fluides, l'amélioration de la reconnaissance et de l'expression émotionnelles, ainsi que l'amélioration de la gestion du dialogue, entre autres. Notre objectif est de rendre les ACI plus naturels, leur permettant de comprendre les besoins des utilisateurs et de s'exprimer aussi richement que dans la communication humaine à humaine. Cela nécessite que les ACI traitent les informations utilisateur en temps réel et prennent des décisions et des réponses appropriées. Surtout pendant les conversations, des décisions telles que quand prendre la parole, quand fournir des

signaux de retour, quand interrompre un utilisateur, et des considérations similaires sont d'une importance capitale.

Les chercheurs travaillent activement au développement d'ACI pouvant s'intégrer harmonieusement dans les environnements humains et offrir une expérience conversationnelle plus engageante et semblable à celle des humains. Ces efforts englobent un large éventail de domaines, de l'intelligence artificielle et de l'apprentissage automatique à l'interaction homme-machine et à la psychologie. L'objectif ultime est de créer des ACI capables de comprendre, de s'adapter et d'améliorer la communication humaine dans divers contextes.

Converser en temps réel, à la fois écouter et parler simultanément, est une tâche complexe. Dans nos interactions humaines quotidiennes, nous avons développé l'habitude de laisser une seule personne parler à la fois. Cette pratique implique des changements de rôle fréquents et rapides entre le locuteur et l'auditeur. Juger du moment opportun pour prendre la parole peut sembler simple pour les humains, mais cela reste un défi de taille pour les Agents Conversationnels Incarnés (ACI).

Pour les ACI, le défi principal réside dans le discernement précis de savoir si un utilisateur vient de conclure son énoncé ou non, afin de prendre des décisions sur le moment où prendre la parole. Pour garantir une prise de parole précise, certains ACI utilisent ce qu'on appelle une configuration "Wizard-of-Oz". Dans cette approche, un opérateur humain caché contrôle le système et prend les décisions cruciales sur la prise de parole. Bien que cette méthode offre une expérience utilisateur positive, elle n'est pas automatisée, exigeant une intervention humaine, et est donc impraticable pour les applications à grande échelle.

Par la suite, certains systèmes de dialogue ont incorporé la Détection d'Activité Vocale (DAV) pour détecter la conclusion du discours d'un utilisateur en fonction de seuils de durée de silence. Cependant, fixer la longueur appropriée de ces seuils représente un défi important. Si le seuil est fixé trop court, il risque de mal évaluer les pauses à l'intérieur d'un tour de parole de l'utilisateur. À l'inverse, s'il est trop long, cela peut avoir un impact néfaste sur l'expérience utilisateur, donnant l'impression que la conversation est disjointe.

Les chercheurs et les développeurs travaillent activement sur des mécanismes de prise de parole plus sophistiqués et automatisés pour les ACI. Ils s'appuient sur des techniques issues de domaines tels que la reconnaissance vocale, le traitement du langage naturel et l'apprentissage automatique. L'objectif ultime est de doter les ACI de la capacité à reconnaître les indices subtils indiquant la conclusion d'un tour de parole de l'utilisateur, améliorant ainsi la fluidité et la naturalité des interactions humain-ACI. Cela implique de répondre à la fois aux aspects techniques et à l'expérience utilisateur pour garantir que les ACI peuvent engager efficacement les utilisateurs dans la conversation.

Pendant ce temps, la plupart des recherches concernant la prise de parole dans les interactions humain-agent visent à minimiser les chevauchements, un phénomène assez courant dans les interactions humain-humain. De même, les interruptions dans les dialogues humain-agent sont souvent perçues comme des erreurs système. Cependant, dans les conversations humaines quotidiennes, les

interruptions sont assez courantes, et nous ne les catégorisons pas simplement comme un comportement inapproprié. Au contraire, des interruptions bien synchronisées peuvent ajuster le rythme d'une conversation et engager l'interaction. Une interaction complètement ininterrompue pourrait sembler monotone, laissant les utilisateurs avoir l'impression que l'attention de l'agent n'est pas centrée sur eux. D'un autre côté, trop d'interruptions peuvent être très perturbantes. Permettre des interruptions lorsque nécessaire dans les interactions humain-agent pourrait potentiellement améliorer la qualité de l'interaction et, par conséquent, améliorer l'expérience utilisateur.

Bien sûr, les interruptions discutées ici ne sont pas dues à des erreurs de décision système conduisant à une mauvaise initiation de prise de parole. Au lieu de cela, elles sont semblables au jugement indépendant d'un auditeur dans les conversations humain-humain. Elles impliquent que les auditeurs s'insèrent dans la conversation avant que le locuteur ait fini, essayant de prendre la parole, que ce soit de manière coopérative ou compétitive, en fonction de l'intention de l'interrupteur. Ces interruptions peuvent servir à diverses fins, telles que demander des éclaircissements, exprimer de l'enthousiasme ou même remettre en question les affirmations du locuteur. Bien que de telles interruptions puissent dévier des normes conventionnelles de prise de parole, elles font partie de la riche toile de fond de la communication humaine, contribuant à la nature dynamique et engageante des conversations. Dans les interactions humain-agent, permettre stratégiquement de telles interruptions, lorsque cela est approprié, pourrait potentiellement conduire à des échanges plus dynamiques et naturels, améliorant ainsi l'expérience utilisateur globale.

## 0.3 Comportement non verbal dans l'interaction entre agents humains

La communication ne se limite jamais aux mots prononcés. Une communication efficace est une interaction complexe entre le langage verbal et le comportement non verbal multimodal. Alors que la communication verbale transmet le sens explicite des mots, le comportement non verbal joue un rôle crucial en complément, enrichissant et renforçant le message global (Giles [2016]). Par conséquent, lors de la prise de décisions éclairées, il est essentiel de considérer si les schémas de comportement non verbal correspondent au message voulu. Cela devient particulièrement crucial dans le contexte des agents virtuels, où les indices non verbaux réalistes sont essentiels pour créer une expérience de communication immersive et semblable à celle des humains. Dans de telles interactions, l'harmonie entre les décisions et les éléments non verbaux influence considérablement l'efficacité et l'authenticité de l'échange.

Dans le contexte des Agents Conversationnels Incarnés (ACI), l'intégration d'expressions faciales, de regards, de postures et de gestes corporels semblables à ceux des humains a un impact significatif sur leur expressivité et leur niveau d'engagement (Lugrin [2021]). Un agent virtuel capable d'afficher des indices



non verbaux appropriés et pertinents peut établir un sentiment de rapport et de familiarité, favorisant une conversation plus naturelle et interactive avec les utilisateurs.

Les indices non verbaux chez les agents virtuels peuvent illustrer visuellement divers aspects du message verbal, renforçant ou clarifiant le sens voulu. Par exemple, un hochement de tête en signe d'accord peut renforcer la compréhension de l'agent par rapport à la déclaration de l'utilisateur, tandis qu'une expression perplexe peut indiquer la confusion de l'agent, encourageant ainsi l'utilisateur à fournir des éclaircissements supplémentaires. De plus, les indices non verbaux peuvent transmettre des émotions et des nuances difficiles à exprimer uniquement à travers des mots, enrichissant la richesse émotionnelle de l'interaction.

En exploitant le pouvoir de la communication non verbale, les ACI ont le potentiel de combler le fossé entre les interactions humain-humain et humain-machine, rendant ainsi le processus de communication plus fluide et plus efficace.

## 0.4 Questions de recherche

Comme mentionné précédemment, les interruptions peuvent effectivement améliorer l'expérience utilisateur lors des interactions entre humains et agents. Accorder aux ACI la capacité de gérer les interruptions est crucial. En plus de pouvoir gérer les interruptions des utilisateurs humains, les ACI doivent également avoir la capacité d'interrompre les utilisateurs humains de manière appropriée, de manière compréhensible et acceptable pour les utilisateurs, évitant ainsi la perception d'erreurs résultant de décisions système.

À ce jour, quelques études se sont concentrées sur le premier aspect, qui implique de répondre à une interruption de l'utilisateur pendant l'interaction. Nous approfondirons cet aspect dans le chapitre 3. Cependant, il existe actuellement un manque de recherche sur les ACI initiants des interruptions avec les utilisateurs humains, qui forme le thème central de cette thèse. L'objectif est de doter les ACI de la capacité d'initier de manière appropriée des interruptions lors des conversations.

Pour permettre aux ACI d'interrompre efficacement les utilisateurs humains, plusieurs aspects clés doivent être abordés. Premièrement, les ACI doivent être en mesure d'évaluer en temps réel les opportunités d'interruption appropriées, leur permettant de s'interposer dans la conversation lorsque cela est motivé ou nécessaire. Deuxièmement, les ACI devraient générer des comportements non verbaux correspondants pour accompagner leurs décisions d'interruption.

Pour rendre les interruptions dans les interactions humain-agent plus naturelles, notre recherche est fondée sur la compréhension des interruptions qui se produisent de manière organique dans les dialogues humain-humain. Ensuite, nous présenterons une introduction détaillée à chacune des questions de recherche abordées dans cette thèse.

### 0.4.1 Identification et classification des interruptions à travers l'interaction humain-humain

Un des objectifs principaux de cette recherche est d'identifier et d'étudier divers types d'interruptions qui se produisent lors des conversations humain-humain. Comprendre comment les humains identifient différentes situations d'interruption est essentiel. Le défi réside dans la capacité des ACI à reconnaître automatiquement et à répondre à ces différents scénarios d'interruption, étant donné que les conversations sont souvent fluides et dynamiques. Pour atteindre cet objectif, la thèse cherche à caractériser les interruptions à travers des signaux multimodaux, en considérant la combinaison de différentes modalités de communication telles que les gestes de tête, les expressions faciales et le langage corporel.

#### Questions de recherche

Pour répondre à l'objectif ci-dessus, les questions de recherche suivantes sont explorées :

1. Comment les humains perçoivent-ils et catégorisent-ils différentes situations d'interruption lors de leurs interactions ?
2. Pouvons-nous tirer parti du comportement multimodal humain pour catégoriser automatiquement les types d'interruption ? Quelles sont les modalités clés qui aident dans cette classification, et comment ces modalités sont-elles exprimées ?
3. Comment pouvons-nous concevoir des modèles informatiques capables d'identifier différents types d'interruption grâce à l'analyse de diverses modalités ?

### 0.4.2 Prédiction du timing des interruptions

Comprendre pourquoi et quand les humains choisissent d'interrompre pendant les conversations est un autre aspect significatif de cette recherche. Les humains possèdent la capacité d'identifier les moments appropriés pour interrompre sans perturber significativement le flux de la conversation. Ce processus de prise de décision est influencé par la volonté de l'auditeur de prendre la parole et la volonté du locuteur de céder la parole lorsqu'il est interrompu, comme indiqué par divers signaux non verbaux. La thèse vise à concevoir des modèles informatiques capables de prédire avec précision les moments propices à l'interruption sur la base des données d'interruption existantes provenant des interactions humain-humain. De plus, elle explore comment les interruptions initiées par les ACI sont perçues par les humains et si elles sont considérées comme acceptables dans les contextes conversationnels.

### Questions de recherche

Pour explorer les subtilités du timing des interruptions, les questions de recherche suivantes sont abordées :

4. Comment les modèles informatiques peuvent-ils identifier les moments appropriés pour les interruptions sur la base des données d'interruption existantes provenant des interactions humain-humain ?
5. Comment les humains perçoivent-ils les interruptions initiées par les ACI ? Sont-elles considérées comme acceptables ou perturbatrices ?
6. Quels sont les principaux facteurs qui impactent la perception des interruptions ? Comment les concepteurs d'ACI peuvent-ils améliorer l'acceptation des ACI capables d'interruption ?

### 0.4.3 Génération du comportement d'interruption

En plus de comprendre quand interrompre, il pourrait exister des signaux spécifiques qui précèdent une interruption. Les humains peuvent rapidement identifier les interruptions et répondre ou ajuster leur comportement en conséquence. Par conséquent, il est essentiel pour les ACI non seulement de savoir quand interrompre, mais aussi de générer un comportement cohérent avec les décisions d'interruption. Nous cherchons à identifier des signaux spécifiques indiquant les interruptions imminentes juste avant leur début et visons à concevoir des modèles informatiques capables de générer des signaux multimodaux appropriés pendant les interruptions.

### Questions de recherche

Pour aborder cet aspect, les questions de recherche suivantes sont étudiées :

7. Existe-t-il des signaux spécifiques qui anticipent les interruptions imminentes juste avant leur début ?
8. Comment les modèles informatiques peuvent-ils générer des signaux multimodaux appropriés pendant les interruptions pour maintenir la cohérence conversationnelle ?
9. Comment les humains perçoivent-ils les interruptions générées par les ACI, et comment le comportement des ACI impacte-t-il la perception des utilisateurs ?

## 0.5 Contributions

Le principal objectif de cette thèse est d'équiper les ACI de la capacité d'initier des interruptions de manière appropriée pendant les conversations. Les objectifs de la recherche sont les suivants :

## 0.5. CONTRIBUTIONS

---

1. Identifier et étudier divers types d'interruption qui se produisent pendant les conversations humain-humain. Cela implique de comprendre comment les humains reconnaissent différentes situations d'interruption grâce à des signaux multimodaux.
2. Concevoir des modèles informatiques capables de prédire avec précision les moments propices aux interruptions sur la base de données d'interruption existantes provenant des interactions humain-humain.
3. Développer des modèles informatiques capables de générer des signaux multimodaux appropriés pendant les interruptions.

Pour atteindre ces objectifs et répondre aux diverses limitations et défis techniques, la thèse propose différents modèles et ensembles de données, qui sont discutés en détail ci-dessous.

### 0.5.1 Annotation des corpus NoXi et AMI & classification des types d'interruption (Chapitre 4, 5 et 6)

La première contribution de la thèse implique de proposer un nouveau schéma d'annotation pour l'annotation manuelle des interruptions. Ce schéma couvre diverses situations d'interruption rencontrées dans les conversations quotidiennes, et il tient également compte d'autres cas tels que les rétroactions et les échanges de tours fluides. En utilisant ce schéma, les corpus NoXi et AMI sont annotés, permettant l'utilisation de ces ensembles de données dans diverses études liées à l'analyse multimodale, la prédiction du timing des interruptions et la génération du comportement des interruptions.

- *Corpus NoXi* : Ce corpus est fondamental pour analyser les interruptions dans l'interaction humain-humain (questions de recherche Q1 et Q2), entraîner, tester et valider des modèles qui classifient différents types d'interruptions (question de recherche Q3), prédire le timing d'interruption possible (question de recherche Q4), et générer des gestes faciaux pendant la période d'interruption (questions de recherche Q7 et Q8).
- *Corpus AMI* : Initialement construit par Carletta [2007], pour développer une technologie de navigation dans les réunions, ce corpus est étendu dans cette thèse pour inclure des fonctionnalités multimodales supplémentaires liées aux expressions faciales et aux caractéristiques acoustiques de bas niveau. Il est utilisé pour entraîner, tester et valider le modèle de classification des types d'interruption (question de recherche Q3).

### 0.5.2 Classification à une classe pour la prédiction du timing des interruptions possibles (Chapitre 7)

La deuxième contribution implique de proposer une nouvelle approche pour prédire le timing possible de l'initiation des interruptions dans les interactions dyadiques.

Ce modèle est conçu pour être appliqué à un agent virtuel, en tenant compte des différences de comportement potentielles par rapport aux humains réels. L'approche est basée sur un modèle de classification à une classe entraîné sur le corpus NoXi.

- *Prédiction du timing possible d'initiation des interruptions* : Le modèle de classification à une classe est entraîné pour détecter les interruptions sur la base des échantillons positifs existants en utilisant des expressions faciales, des mouvements de tête et des caractéristiques acoustiques de bas niveau (question de recherche Q4).
- *Étude de la perception des interruptions* : Pour évaluer le modèle de prédiction du timing, une étude perceptuelle est menée, comparant les interruptions prédites par le modèle avec les données de référence et les interruptions aléatoires, en tenant compte de diverses variables indépendantes comme le timing des interruptions, la parole de l'interrupteur, la voix audio de l'interrupteur et le type d'interruption (questions de recherche Q5 et Q6).

### 0.5.3 Génération du comportement d'interruption (Chapitre 8)

La dernière contribution de la thèse est de générer un comportement non verbal pendant les interruptions. En raison de la disponibilité limitée des données d'interruption, un modèle génératif pré-entraîné est adapté à cette fin. Le modèle génère des expressions faciales et des rotations de tête en temps réel.

- *Génération du comportement d'interruption* : Différentes technologies d'apprentissage par transfert sont comparées pour enseigner au modèle à générer spécifiquement le comportement pendant les périodes d'interruption. Le modèle apprend à partir des caractéristiques acoustiques, des expressions faciales et des mouvements de tête des deux interlocuteurs pour générer le prochain cadre pour l'interrupteur (question de recherche Q8).
- *Évaluation du comportement d'interruption* : Une étude perceptuelle est menée pour évaluer le comportement d'interruption généré, le comparant à la vérité de terrain et au comportement généré par le modèle général avant l'apprentissage par transfert (questions de recherche Q9).

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Embodied Conversational Agents . . . . .	2
1.2	Turn-taking & interruption in human agent interaction . . . . .	3
1.3	Nonverbal behaviour in human agent interaction . . . . .	5
1.4	Research Questions . . . . .	5
1.5	Contributions . . . . .	8
1.6	Publications and Submissions . . . . .	9
1.7	Thesis Outline . . . . .	10
<b>2</b>	<b>Background</b>	<b>12</b>
2.1	Multimodal human communication . . . . .	13
2.2	Human Conversation . . . . .	18
<b>3</b>	<b>Related Works</b>	<b>26</b>
3.1	Turn-taking cues . . . . .	27
3.2	Handling turn-taking . . . . .	29
3.3	Interruption cues . . . . .	32
3.4	Handling interruption in human-agent interaction . . . . .	34
3.5	Positioning . . . . .	38
<b>4</b>	<b>Corpora</b>	<b>40</b>
4.1	Corpora: NoXi / AMI . . . . .	41
4.2	Features . . . . .	42
4.3	Data Cleaning . . . . .	46
4.4	Conclusion . . . . .	48
<b>5</b>	<b>Annotation schema &amp; multimodal analysis</b>	<b>50</b>
5.1	Related Works . . . . .	51
5.2	Annotation Schema . . . . .	53
5.3	Statistical results . . . . .	58
5.4	Multimodal Analysis . . . . .	60
5.5	Conclusion . . . . .	63

## CONTENTS

---

<b>6</b>	<b>MIC: Multimodal interruption classification in humans' interaction</b>	<b>65</b>
6.1	Related works . . . . .	66
6.2	Features . . . . .	67
6.3	The proposed model . . . . .	71
6.4	Result & discussion . . . . .	71
6.5	Conclusion . . . . .	77
<b>7</b>	<b>One-PredIT: Prediction of Interruption Timing in dyadic interaction using one-class classification model</b>	<b>79</b>
7.1	Related works . . . . .	80
7.2	Approach . . . . .	81
7.3	Comparative study . . . . .	82
7.4	Subjective Evaluation . . . . .	82
7.5	Discussion . . . . .	92
7.6	Conclusion . . . . .	93
<b>8</b>	<b>AI-BGM: Agent Interruption Behaviour Generation Model.</b>	<b>95</b>
8.1	Related works . . . . .	96
8.2	Corpus & Features . . . . .	101
8.3	Interruption preparation duration . . . . .	101
8.4	Smooth turn exchange vs. interruption . . . . .	103
8.5	ASAP introduction . . . . .	104
8.6	Proposed model . . . . .	106
8.7	Objective evaluation . . . . .	107
8.8	Subjective evaluation . . . . .	109
8.9	Discussion . . . . .	113
8.10	Conclusion . . . . .	114
<b>9</b>	<b>Conclusion</b>	<b>116</b>
9.1	Summary . . . . .	116
9.2	Contributions . . . . .	117
9.3	Limitations and Future Work . . . . .	118

# List of Figures

1	Illustration de (a) Greta, un agent conversationnel incarné (Pelachaud [2015]), (b) le robot social Furhat (Al Moubayed et al. [2013]), (c) un robot humanoïde NAO (Shamsuddin et al. [2011]) et (d) le robot-personnage MIROKI développé par la société Enchanted Tools. . . . .	v
1.1	Illustration of (a) Greta, an embodied conversational agent (Pelachaud [2015]), (b) Furhat social robot (Al Moubayed et al. [2013]), (c) a humanoid NAO robot (Shamsuddin et al. [2011]) and (d) Robot-character MIROKI developed by Enchanted Tools company. . . . .	3
2.1	Action Units of FACS system. Figures are from (Fac, HAGER [2002])	15
2.2	Combinations of action units. Figures are from (Brahnam et al. [2007], HAGER [2002]) . . . . .	15
2.3	(1) $t_2 - t_1 < 0$ , negative FTO. (2) $t_2 - t_1 > 0$ , positive FTO. . . . .	21
4.1	An example of NoXi dyadic conversation. Figure from (Cafaro et al. [2017]) . . . . .	41
4.2	AMI corpus: four-party English conversation . . . . .	42
4.3	Open Face AU detection . . . . .	45
4.4	This figure is a plot of different median filtering window sizes applied to <i>AU1</i> signal. . . . .	47
4.5	Linear Interpolation is applied on the frames where OpenFace's <i>success score</i> is equal to 0. . . . .	48
5.1	Classification of interruption and smooth speaker exchange (Beattie [1981]). . . . .	52
5.2	Interruption annotation schema . . . . .	54
5.3	Nova annotation interface . . . . .	56
5.4	Overlap length in second for different types of exchanges. Labels: Interruption (interrupt), backchannel (BC), smooth turn exchange (s_turn), cooperative interruption (coop), competitive interruption (comp). . . . .	59
5.5	Relative distance in seconds between the exchange onset point and the start of the speaker's last IPU for different types of exchanges. Labels: Interruption (interrupt), backchannel (BC), smooth turn exchange (s_turn), cooperative interruption (coop), competitive interruption (comp) . . . . .	60



## LIST OF FIGURES

---

5.6	Explanation of data segmentation structure (taken: the last IPU of the speaker before the exchange, taker: the first IPU of the exchange initiator after the exchange onset point). . . . .	61
5.7	Average value of selected features during the corresponding intervals (taken: [t1, t2], taker: [t3,t4]). . . . .	62
6.1	Interruption overlap duration in second (AMI corpus). . . . .	68
6.2	Segmentation of features. . . . .	69
6.3	The long-short-term memory (LSTM) architecture. . . . .	71
6.4	Accuracy & Macro F1-score with different interruption window lengths. . . . .	75
6.5	Percentage of classified interruptions according to the number of frames after the beginning of the overlap, for different thresholds. . . . .	77
7.1	Screenshot of generated video for an interruption. Interrupter on the left side ("Yep, yep."), and interruptee on the right side ("So after coming back from school he has some time to play and ..."). . . . .	84
7.2	Interruption simulation: for the same speaking turn of interruptee, predicted interruption and random interruption may occur at different timing as ground truth. Once interrupted, the interruptee audio was cut off after finishing the current word. . . . .	85
7.3	Error bar plot of 8 groups by question. . . . .	88
8.1	Classification model architecture. . . . .	101
8.2	One-second segments with offset distance (each 0.2s from 0.2s to 1.2s). . . . .	102
8.3	Classification model accuracy with different offset duration (0s, 0.2s, 0.4s, 0.6s, 0.8s, 1.0s, and 1.2s) . . . . .	103
8.4	ASAP model architecture. Image from Woo et al. [2023]. . . . .	105
8.5	Video timeline. . . . .	110
8.6	User perception test video clip examples of interactions between a male/female SIA and a human participant from NoXi database. . . . .	110
8.7	Error bar plot for seven questions rating results. . . . .	113

# List of Tables

4.1	Features utilized in our research. . . . .	43
5.1	Probability distribution according to the 8 types of interruption. . .	57
5.2	Probability distribution of different interruption types for successful interruption. . . . .	57
5.3	Probability distribution of different interruption types for Failed interruption. . . . .	58
6.1	Accuracy and F1 measure for FFNN, SVM and LSTM model with different combinations of modalities for the AMI corpus. . . . .	72
6.2	Ablation study of FFNN, SVM and LSTM model for the AMI corpus with our modalities. . . . .	72
6.3	Accuracy and F1 measure for FFNN, SVM and LSTM model with different combinations of modalities for the NoXi corpus. . . . .	74
6.4	Ablation study of FFNN, SVM and LSTM model for the NoXi corpus with our modalities. . . . .	74
6.5	Mean accuracy & mean reaction time with different thresholds . . .	77
7.1	Accuracy & F1-score for Deep residual learning network and One-class SVM models with different modality combinations. . . . .	82
7.2	Scripted interrupter speech sentences, selected from interruptions in our corpus. . . . .	85
7.3	Evaluation questions. Separated into three question sets. The 11 questions were asked for all the evaluated interruptions, randomly ordered. . . . .	87
7.4	Comparison of different groups for 11 questions are presented in the table. Mean differences are reported, with a positive value indicating that the first group scored higher than the second group. Significant differences ( $p < 0.05$ ) between the first and second groups are highlighted in green colour. . . . .	89
7.5	Comparison of different interruption types for 11 questions are presented in the table (groups 7 and 8). Mean differences are reported, with a positive value indicating that the first group scored higher than the second group. Significant differences between the first and second groups are highlighted in green colour. . . . .	89

## LIST OF TABLES

---

8.1	Comparison of different comparison groups for 7 questions are presented in the table. Mean differences are reported, with a positive value indicating that the first group scored higher than the second group. Significant differences between the first and second groups are highlighted in green colour ( $p < 0.05$ ). . . . .	111
8.2	RMSE values of generated features for the six models. . . . .	112
8.3	KS test values of generated features for the six models. . . . .	112

# Chapter 1

## Introduction

### Contents

---

1.1	Embodied Conversational Agents . . . . .	2
1.2	Turn-taking & interruption in human agent interaction . . . . .	3
1.3	Nonverbal behaviour in human agent interaction . . . . .	5
1.4	Research Questions . . . . .	5
1.4.1	Identifying and classifying interruptions through human-human interaction . . . . .	6
1.4.2	Interruption timing prediction . . . . .	7
1.4.3	Interruption behaviour generation . . . . .	7
1.5	Contributions . . . . .	8
1.5.1	Annotating NoXi and AMI corpora & interruption types classification (Chapter 4, 5 and 6) . . . . .	8
1.5.2	One-class classification for possible interruption timing prediction (Chapter 7) . . . . .	9
1.5.3	Interruption behaviour generation (Chapter 8) . . . . .	9
1.6	Publications and Submissions . . . . .	9
1.7	Thesis Outline . . . . .	10

---

*Man with all his noble qualities, with sympathy which feels for the most debased, with benevolence which extends not only to other men but to the humblest living creature, with his God-like intellect which has penetrated into the movements and constitution of the solar system—with all these exalted powers—Man still bears in his bodily frame the indelible stamp of his lowly origin.*

---

Charles Darwin (1871/1898)

What distinguishes humans from other animals is our ability to create and use tools to simplify labour and enhance productivity, as well as our powerful capacity for logical thinking and complex yet coherent communication skills. While some animals use tools, humans have taken tool use to an unprecedented level. The history of human civilization can be seen as a history of technological development, ranging from the use of simple stone tools to metallurgy, then to steam power, and finally to the current era of digital and network technologies (Bogin and Varea [2020]). The rapid advancements in technology have propelled humanity into the virtual age, and "digital humanism," as a bridge connecting humans and humanoids, has emerged in recent years and gained popularity (Davies [2016], Wagner et al. [2020]).

Due to our inherent preference for interpersonal communication, Embodied Conversational Agents (ECAs) have emerged as a burgeoning Human-Computer Interaction (HCI) interface. By simulating human forms, ECAs can convey information through voice, gestures, facial expressions, and body movements, making communication more accurate and humane compared to plain text (Lugrin [2021]). With the significant advancements in 3D technology, XR technology, and the commercial availability of devices, the applications of ECAs have become more diversified. Undoubtedly, they are making their way from the laboratories into people's daily lives, offering assistance in various aspects of work.

## 1.1 Embodied Conversational Agents

ECAs are designed to simulate human-like social intelligence and behaviour, effectively bridging the gap between humans and machines (Ruttkay and Pelachaud [2004]). Figure 1.1 provides a visual representation of various ECA examples.

The versatility of ECAs is highlighted by their deployment across a multitude of domains. In the realm of customer service, they can assume the role of virtual representatives on websites and within applications, adeptly assisting users with inquiries, troubleshooting issues, and retrieving information. In healthcare settings, ECAs prove invaluable by providing medical insights, remotely monitoring patients, conducting therapy sessions, and promoting healthy lifestyle choices. Within the gaming industry, ECAs take on roles as interactive characters, guides, or companions, elevating gameplay by making it more engaging and immersive. Fur-

thermore, akin to popular voice-activated virtual assistants like Siri or Alexa, ECAs offer a conversational interface for performing tasks such as setting reminders, answering queries, and controlling smart devices.

The utilization of ECAs represents a concerted effort by researchers to establish more natural and engaging interactions between humans and machines. These agents are not merely tools but dynamic facilitators of communication and engagement, enriching various facets of our lives through their human-like attributes and capabilities.

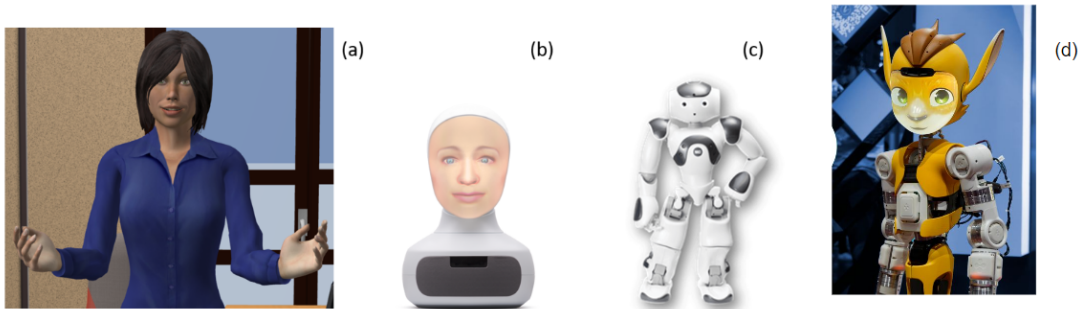


Figure 1.1 Illustration of (a) Greta, an embodied conversational agent (Pelachaud [2015]), (b) Furhat social robot (Al Moubayed et al. [2013]), (c) a humanoid NAO robot (Shamsuddin et al. [2011]) and (d) Robot-character MIROKI developed by Enchanted Tools company.

## 1.2 Turn-taking & interruption in human agent interaction

In order to make the interaction between ECAs and human users smoother and more in line with human habits and behaviour logic, researchers have begun to explore various directions. These directions include simulating realistic appearances, ensuring natural and fluid movements, enhancing emotional recognition and expression, and improving dialogue management, among others. Our goal is to make ECAs appear more natural, enabling them to understand user needs and express themselves as richly as human-to-human communication. This requires ECAs to process user information in real time and make appropriate decisions and responses. Especially during conversations, decisions such as when to take the floor, when to provide backchannels, when to interrupt a user, and similar considerations are of paramount importance.

Researchers are actively working to develop ECAs that can seamlessly integrate into human environments and provide a more engaging and human-like conversational experience. These efforts encompass a wide range of fields, from artificial intelligence and machine learning to human-computer interaction and psychology. The ultimate aim is to create ECAs that can understand, adapt to, and enhance human communication in various contexts.

Conversing in real-time, both listening and speaking simultaneously, is an intricate task. In our day-to-day human interactions, we've developed the habit of having only one person speak at a time. This practice involves frequent and swift role shifts between the speaker and the listener. Judging the opportune moment for turn-taking may appear straightforward for humans, but it remains a formidable challenge for Embodied Conversational Agents (ECAs).

For ECAs, the key challenge lies in accurately discerning whether a user has just concluded their statement or not, in order to make decisions about when to assume the conversational floor. To ensure precise turn-taking, some ECAs employ what's known as a "Wizard-of-Oz setup." In this approach, a concealed human operator controls the system and makes the critical turn-taking decisions. While this method results in a positive user experience, it is not automated, demanding human intervention, and is thus impractical for large-scale applications.

Subsequently, certain dialogue systems have incorporated Voice Activity Detection (VAD) to detect the conclusion of a user's speech based on silence duration thresholds. However, setting the appropriate length for these thresholds presents a significant challenge. If the threshold is set too short, it risks misjudging pauses within a user's turn. Conversely, if it's too long, it can have a detrimental impact on the user experience, causing the conversation to feel disjointed.

Researchers and developers are actively working on more sophisticated and automated turn-taking mechanisms for ECAs. They are drawing upon techniques from fields such as speech recognition, natural language processing, and machine learning. The ultimate aim is to equip ECAs with the capability to recognize subtle cues indicating the conclusion of a user's turn, thereby enhancing the fluidity and naturalness of human-ECA interactions. This involves addressing both the technical and user experience aspects to ensure that ECAs can engage users effectively in conversation.

Meanwhile, most research concerning turn-taking in human-agent interactions aims to minimize overlap, a phenomenon quite common in human-human interactions. Similarly, interruptions in human-agent dialogues are often seen as system errors. However, in everyday human conversations, interruptions are fairly common, and we don't simply categorize them as inappropriate behaviour. Instead, well-timed interruptions can adjust the rhythm of a conversation and engage the interaction. A completely uninterrupted interaction might come across as monotonous, leaving users feeling like the agent's attention isn't focused on them. On the other hand, too many interruptions can be highly disruptive. Allowing interruptions when necessary in human-agent interactions could potentially enhance the quality of the interaction and, consequently, improve the user experience.

Of course, the interruptions discussed here are not due to system decision errors leading to incorrect turn-taking initiation. Instead, they are akin to a listener's independent judgment in human-human conversations. They involve listeners inserting themselves into the conversation before the speaker has finished, attempting to grab the conversational floor, whether cooperatively or competitively, depending on the interrupter's intention. These interruptions can serve

various purposes, such as seeking clarification, expressing enthusiasm, or even challenging the speaker's statements. While such interruptions may deviate from the conventional norms of turn-taking, they are part of the rich tapestry of human communication, contributing to the dynamic and engaging nature of conversations. In human-agent interactions, strategically allowing for such interruptions, when appropriate, could potentially lead to more vibrant and natural exchanges, thereby enhancing the overall user experience.

## 1.3 Nonverbal behaviour in human agent interaction

Communication is never solely about the spoken words. Effective communication is a multifaceted interplay between spoken language and multimodal nonverbal behaviour. While verbal communication conveys the explicit meaning of words, nonverbal behaviour serves as a crucial complement, enriching and reinforcing the overall message (Giles [2016]). Therefore, when making informed decisions, it's essential to consider whether the patterns of nonverbal behaviour align with the intended message. This becomes particularly pivotal in the context of virtual agents, where lifelike nonverbal cues are instrumental in crafting an immersive and human-like communication experience. In such interactions, the harmony between decisions and nonverbal elements significantly influences the effectiveness and authenticity of the exchange.

In the context of Embodied Conversational Agents (ECAs), the incorporation of human-like facial expressions, gaze, posture, and body gestures significantly impacts their expressiveness and engagement level (Lugrin [2021]). A virtual agent capable of displaying appropriate and relatable nonverbal cues can establish a sense of rapport and familiarity, fostering a more natural and interactive conversation with users.

Nonverbal cues in virtual agents can visually illustrate various aspects of the spoken message, reinforcing or clarifying the intended meaning. For instance, a nod of agreement can reinforce the agent's understanding of the user's statement, while a puzzled expression can indicate the agent's confusion, encouraging the user to provide further clarification. Additionally, nonverbal cues can convey emotions and nuances that may be challenging to express purely through words, enhancing the emotional richness of the interaction.

By harnessing the power of nonverbal communication, ECAs have the potential to bridge the gap between human-human and human-machine interactions, making the communication process smoother and more effective.

## 1.4 Research Questions

As mentioned earlier, interruption may indeed enhance the user experience during human-agent interactions. Granting ECAs the capability to manage interruptions



is crucial. Besides being able to handle interruptions from human users, ECAs also need the ability to interrupt human users appropriately, in a manner that is understandable and acceptable to users, avoiding the perception of errors resulting from system decisions.

To date, there have been some studies focusing on the first aspect, which involves responding to a user's interruption during interaction. We will delve into this aspect in detail in Chapter 3. However, there is currently a lack of research on ECAs initiating interruptions with human users, which forms the central theme of this thesis. The goal is to empower ECAs with the capability to appropriately initiate interruptions during conversations.

To enable ECAs to interrupt human users effectively, several key aspects need to be addressed. Firstly, ECAs must be able to assess real-time opportunities for appropriate interruptions, allowing them to interject into the conversation when motivated or necessary. Secondly, ECAs should generate corresponding nonverbal behaviours to accompany their interruption decisions.

To make interruptions in human-agent interactions appear more natural, our research is grounded in understanding of interruptions that occur organically in human-human dialogues. Next, we will provide a detailed introduction to each of the research questions addressed in this thesis.

### 1.4.1 Identifying and classifying interruptions through human-human interaction

One of the primary objectives of this research is to identify and study various interruption types that occur during human-human conversations. Understanding how humans identify different interruption situations is essential. The challenge lies in enabling ECAs to automatically recognize and respond to these diverse interruption scenarios, as conversations are often fluid and dynamic. To achieve this objective, the thesis seeks to characterize interruptions through multimodal signals, considering the combination of different communication modalities such as head gestures, facial expressions, and body language.

#### Research Questions

To address the above objective, the following research questions are explored:

1. How do humans perceive and categorize different interruption situations during their interactions?
2. Can we leverage human multimodal behaviour to automatically categorize interruption types? What are the key modalities that aid in this classification, and how are these modalities expressed?
3. How can we design computational models capable of identifying different interruption types through analysis of various modalities?

### 1.4.2 Interruption timing prediction

Understanding why and when humans choose to interrupt during conversations is another significant aspect of this research. Humans possess the ability to identify appropriate moments to interrupt without disrupting the conversation flow significantly. This decision-making process is influenced by the listener's willingness to take over the floor and the speaker's willingness to yield to the floor when interrupted, as signalled through various nonverbal cues. The thesis aims to design computational models that can accurately predict suitable interruption moments based on existing interruption data from human-human interactions. Additionally, to explore how ECA-initiated interruptions are perceived by humans and whether they are deemed acceptable in conversational settings.

#### Research Questions

To delve into the intricacies of interruption timing, the following research questions are addressed:

4. How can computational models identify appropriate moments for interruptions based on existing interruption data from human-human interactions?
5. How do humans perceive interruptions initiated by ECAs? Are they considered acceptable or disruptive?
6. What are the key factors that impact the perception of interruptions? How can ECA designers enhance the acceptance of interruption-capable ECAs?

### 1.4.3 Interruption behaviour generation

In addition to understanding when to interrupt, there might exist specific signals that may precede an interruption. Humans can quickly identify interruptions and respond or adjust their behaviour accordingly. Therefore, it is essential for ECAs to not only know when to interrupt but also to generate coherent behaviour that aligns with interruption decisions. We seek to identify specific signals that indicate upcoming interruptions just before their onset and aim to design computational models capable of generating appropriate multimodal signals during interruptions.

#### Research Questions

To address this aspect, the following research questions are investigated:

7. Are there specific signals that anticipate forthcoming interruptions right before their onset?
8. How can computational models generate appropriate multimodal signals during interruptions to maintain conversational coherence?

9. How do humans perceive interruptions generated by ECAs, and how does the behaviour of ECAs impact user perception?

## 1.5 Contributions

The main focus of this thesis is to equip ECAs with the capability to appropriately initiate interruptions during conversations. The objectives of the research are as follows:

1. Identifying and studying various interruption types that occur during human-human conversations. This involves understanding how humans recognize different interruption situations through multimodal signals.
2. Designing computational models that can accurately predict suitable moments for interruptions based on existing interruption data from human-human interactions.
3. Developing computational models capable of generating appropriate multimodal signals during interruptions.

To achieve these objectives and address the various limitations and technical challenges, the thesis proposes different models and datasets, which are discussed in detail below.

### 1.5.1 Annotating NoXi and AMI corpora & interruption types classification (Chapter 4, 5 and 6)

The first contribution of the thesis involves proposing a new annotation schema for manual interruption annotation. This schema covers various interruption situations encountered in daily conversations, and it also considers other cases such as backchannels and smooth turn exchanges. Using this schema, the NoXi and AMI corpora are annotated, enabling the use of these datasets in various studies related to multimodal analysis, interruption timing prediction, and interruption behaviour generation.

- *NoXi corpus*: This corpus is foundational for analyzing interruptions in human-human interaction (research questions Q1 and Q2), training, testing, and validating models that classify different types of interruptions (research questions Q3), predicting possible interruption timing (research questions Q4), and generating facial gestures during the interruption period (research questions Q7 and Q8).
- *AMI corpus*: Originally built by Carletta [2007], for developing meeting browsing technology, this corpus is extended in this thesis to include additional multimodal features related to facial expressions and low-level acoustic features. It is used to train, test, and validate the interruption type classification model (research questions Q3).

### 1.5.2 One-class classification for possible interruption timing prediction (Chapter 7)

The second contribution involves proposing a novel approach to predict possible interruption initiation timing in dyadic interactions. This model is designed to be applied to a virtual agent, considering potential behavioural differences from real humans. The approach is based on a one-class classification model trained on NoXi corpus.

- *Possible interruption initiation timing prediction:* The one-class classification model is trained to detect interruptions based on existing positive samples using facial expressions, head motion, and low-level acoustic features (research questions Q4).
- *Interruption perception study:* To evaluate the timing prediction model, a perceptual study is conducted, comparing model-predicted interruptions with ground truth data and random interruptions, considering various independent variables like interruption timing, interrupter speech, interrupter audio voice, and interruption type (research questions Q5 and Q6).

### 1.5.3 Interruption behaviour generation (Chapter 8)

The final contribution of the thesis is to generate nonverbal behaviour during interruptions. Due to the limited availability of interruption data, a pretrained generative model is adapted for this purpose. The model generates facial expressions and head rotations in real time.

- *Interruption behaviour generation:* Different transfer learning technologies are compared to teach the model to specifically generate behaviour during interruption periods. The model learns from acoustic features, facial expressions, and head motion from both interlocutors to generate the next frame for the interrupter (research questions Q8).
- *Interruption behaviour evaluation:* A perceptual study is conducted to evaluate the generated interruption behaviour, comparing it with ground truth and behaviour generated by the general model before transfer learning (research questions Q9).

## 1.6 Publications and Submissions

- Liu Yang, Catherine Achard, Catherine Pelachaud. What If I Interrupt You? Proceedings of the 2021 International Conference on Multimodal Interaction. 2021.
- Liu Yang, Catherine Achard, Catherine Pelachaud. Modeling of interruptions in human-agent interaction. WACAI 2021.

- Liu Yang, Catherine Achard, Catherine Pelachaud. Annotating Interruption in Dyadic Human Interaction. Thirteenth Language Resources and Evaluation Conference, LREC, 2022.
- Liu Yang, Catherine Achard, Catherine Pelachaud. Multimodal Analysis of Interruptions. International Conference on Human-Computer Interaction. HCII, 2022.
- Liu Yang, Catherine Achard, Catherine Pelachaud. Multimodal classification of interruptions in humans' interaction. Proceedings of the 2022 International Conference on Multimodal Interaction. ICMI, 2022.
- Jieyeon WOO, Liu Yang, Catherine Achard, Catherine Pelachaud. Are we in sync during turn switch? 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE (SIVA workshop), 2023.
- Jieyeon WOO, Liu Yang, Catherine Achard, Catherine Pelachaud. Is Turn-Shift Distinguishable with Synchrony? International Conference on Human-Computer Interaction. HCII, 2023.
- Liu Yang, Catherine Achard, Catherine Pelachaud. Now or When? Interruption timing prediction in dyadic interaction. Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. IVA, 2023.

## 1.7 Thesis Outline

This thesis is organized into 8 Chapters:

1. Chapter 2 establishes the necessary background knowledge for *multimodal nonverbal communication, human conversation, interruption* and Chapter 3 which gives an overview of the existing works on *human-agent communication system, turn-taking, interruption and backchannel management and analysis approaches*, their underlying principles, and their limitations.
2. We present in Chapter 4 two corpora: (1) *NoXi* and (2) *AMI Corpus* which we have used for our research.
3. Chapter 5, which is related to *interruption annotation and multimodal analysis for different switch types*.
4. In Chapter 6 we present our first model to classify different types of interruptions.
5. Chapter 7 explains our approach to predict interruption and evaluation results.
6. Finally the generation of virtual agents' nonverbal interruption behaviour in Chapter 8 and we finish with a conclusion (Chapter 9).

**The key points of this Chapter:**

*Goal of this thesis*

- The central theme of this thesis revolves around modelling interruptions during human-agent interaction to develop ECAs that can effectively handle interruptions, both initiating and responding to them.
- The main focus is on empowering ECAs with the ability to appropriately initiate interruptions during conversations. Additionally, the thesis aims to evaluate how human users perceive these interruptions and identify the key factors impacting their perception.

*Thesis Research Questions*

- *Identifying and classifying interruptions.* How do humans perceive and categorize different interruption situations during their interactions? How can we design computational models capable of identifying different interruption types through analysis of various modalities?
- *Interruption timing prediction.* How can computational models identify appropriate moments for interruptions based on existing interruption data from human-human interactions? How do humans perceive interruptions initiated by ECAs? What are the key factors that impact the perception of interruptions?
- *Interruption behaviour generation.* How can computational models generate appropriate multimodal signals during interruptions to maintain conversational coherence? How do humans perceive interruptions generated by ECAs, and how does the behaviour of ECAs impact user perception?

# Background

## Contents

---

2.1	Multimodal human communication . . . . .	13
2.1.1	Facial expression . . . . .	14
2.1.2	Prosodic information . . . . .	16
2.2	Human Conversation . . . . .	18
2.2.1	Turn-taking . . . . .	22
2.2.2	Backchannel . . . . .	23
2.2.3	Interruption . . . . .	24

---

Human communication is a multifaceted system that encompasses both spoken and unspoken channels. Information is shared through various means, with nonverbal cues such as vocal tones, hand and body movements, head motions, or facial expressions complementing spoken words. These nonverbal elements are tightly intertwined with a speaker’s verbal message, serving to emphasize and clarify their intentions. Before delving into the central focus of this study, this chapter serves as an introduction to the diverse communication methods employed by humans. Its aim is to highlight the interplay between these modalities and how individuals collaborate to achieve effective communication. We first discuss the nonverbal communication modalities employed in human interaction. These encompass prosodic information, facial expressions, gestures, gaze and body postures. Subsequently, we delve into a discussion on the "conversation mechanisms," including "turn-taking" and "interruption", wherein interlocutors alternate roles as speakers and listeners. Additionally, we explore "backchannel," initiated by listeners to indicate their focus of attention and provide feedback.

### 2.1 Multimodal human communication

Human beings belong to the primate group, along with our relatives: the great apes like chimpanzees, gorillas, orangutans, and monkeys. Human behaviour has its origins rooted in our phylogenetic history (Knapp et al. [2013], Argyle [2013]), displaying similarities to behaviours exhibited by our nonhuman primate counterparts. Charles Darwin's theory of evolution (Darwin [1998]) found support in the observation of shared expressive behaviours among different species. The progression of evolution was exemplified by the increasing utilization of facial expressions, vocalizations, and body movements for communication and the conveyance of emotions. Darwin (Darwin [1998]) regarded "expressivity" as a pivotal component in the evolutionary discourse. He postulated that a diverse range of expressive and signalling behaviours is intricately tied to the intricacy of a species' social structure. The resemblances in behaviour between humans and non-human primates can be attributed to shared biological and social challenges encountered by both groups.

The resemblances in behaviour between humans and non-human primates sparked the inception of research into nonverbal communication (Argyle [2013]). Nonetheless, significant disparities set humans apart, with the foremost divergence residing in the application of language. Language represents an intricate expressive system founded on speech, and its presence or absence stands as the fundamental distinction between animal and human communication frameworks (Levinson and Holler [2014]).

While animal communication revolves primarily around internal intentions and states, human conversations encompass a broader spectrum, involving individuals, events, and temporal dimensions like the past and future (Argyle [2013], Knapp et al. [2013]). The advent of language introduced a fresh set of nonverbal cues and signals, intended to accompany, provide feedback for, and synchronize with speech. Nonverbal communication in humans serves the purpose of conveying emotions and regulating interpersonal dynamics. Remarkably, the use of nonverbal communication has persisted throughout human evolution (Argyle [2013]). Human distinctiveness is further underscored by their unparalleled complexity and expressiveness (Levinson and Holler [2014]), as well as their strategic orchestration of social behaviour through social acts - premeditated behaviours often involving words and directed toward specific objectives (Argyle [2013]). A hierarchical structure encompasses fundamental nonverbal cues within these social acts.

Nonverbal communication constitutes the initial mode of interaction in the human lifecycle. Prior to the development of verbal language, humans relied on visual body gestures as their primary means of communication (Knapp et al. [2013]). Human-human interaction (HHI) encompasses a wide array of nonverbal cues, including body language, facial expressions, vocal tones, appearance, touch, distancing, and other physical signals. These nonverbal behaviours carry a wealth of information for those engaged in communication, serving to accentuate or clarify the intended message. Among the various channels of communication



in HHI, the human face holds particular significance. It serves as a canvas for a diverse array of verbal, emotional, and conversational cues during interactions. Through facial expressions, individuals can convey their desire to transition between speaking turns (Burgoon et al. [2021]).

Nonverbal communication comprises two primary dimensions: "perception" and "production". On one hand, listeners employ nonverbal cues to "perceive" information about the speaker. Conversely, these cues serve as tools for the speaker to "produce" and convey their intentions. These nonverbal signals encompass prosodic elements (such as pitch, volume, and intonation), facial expressions and body movements (hand gestures for example). Facial expressions are employed either consciously or unconsciously to emphasize words or indicate speech pauses. Many facial expressions and head movements align with the syntactic and prosodic structure of speech.

### 2.1.1 Facial expression

The face stands as the quintessential nonverbal conduit for conveying emotions and attitudes (Ekman [1992], Argyle [2013]). Throughout social interactions, facial expressions undergo rapid transformations, allowing insights into various personality traits. Individuals possess a repertoire of diverse facial expressions, including a range of emotional manifestations like happiness, sadness, fear, and surprise.

In an endeavour to articulate facial expressions, Birdwhistell (Birdwhistell [1974]) proposed a set of thirty-two "kinemes" - elemental building blocks of facial expression. Conversely, Ekman and Friesen (Ekman and Friesen [1982]) defined 44 "Action Units" (AUs), aligning with Birdwhistell's work.

A more systematic exploration by Ekman, Friesen, and Tomkins (Ekman et al. [1971]) paved the way for the development of the Facial Affect Scoring Technique (FAST). This method entails the independent assessment of three distinct facial regions, comparing them against reference photographs. Within these regions, there are 8 positions covering the brows and forehead, 17 concerning the eyes and eyelids, and 45 addressing the lower face. The efficacy of this scoring technique lies in its ability to detect nervous system actions and its capacity to facilitate an analysis of the impacts of antagonistic muscle movements. Building upon this foundation, Ekman and Friesen (Ekman and Friesen [1982]) introduced the more intricate Facial Action Coding System (FACS). This system relies on a comprehensive scheme of minuscule facial movements termed "Action Units" (AUs), each grounded in anatomical principles. These movements, observable to onlookers, are distinct from one another and are driven by individual facial muscles. FACS also incorporates a gradation of intensity, providing a measure of the strength of facial muscle activation. Each AU captures a distinct observable movement of a particular facial feature (e.g., eyebrows) orchestrated by facial muscles. Figure 2.1 presents an illustrative example of Action Units.

## 2.1. MULTIMODAL HUMAN COMMUNICATION

Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser *AU 41	Outer Brow Raiser *AU 42	Brow Lowerer *AU 43	Upper Lid Raiser AU 44	Cheek Raiser AU 45	Lid Tightener AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler AU 15	Upper Lip Raiser AU 16	Nasolabial Deepener AU 17	Lip Corner Puller AU 18	Cheek Puffer AU 20	Dimpler AU 22
					
Lip Corner Depressor AU 23	Lower Lip Depressor AU 24	Chin Raiser *AU 25	Lip Puckerer *AU 26	Lip Stretcher *AU 27	Lip Funneler AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 2.1 Action Units of FACS system. Figures are from (Fac, HAGER [2002])

Diverse emotions are conveyed through the utilization of various combinations of action units, yielding an array of facial expressions and intricate motions. For instance, to convey surprise, the simultaneous activation of action units 1 and 2 is observed (Fac). Examples of such expressions are showcased in Figure 2.2.





















				
AU 1+6	AU 6+7	AU 1+2+5+6+7	AU 23+24	AU 9+17
				
AU 9+25	AU 9+17+23+24	AU 10+17	AU 10+25	AU 10+15+17
				
AU 12+25	AU 12+26	AU 15+17	AU 17+23+24	AU 20+25
				

Figure 2.2 Combinations of action units. Figures are from (Brahnam et al. [2007], HAGER [2002])

Strong correlations are observed between facial expressions and speech, especially eyebrow movements. There are three action units for the eyebrow: (AU1) inner brow raised, (2) outer brow raised, (3) brow lowered. Eyebrow movements tend to occur during moments of thinking pauses (Cavé et al. [1996], Ekman [2004, 1992]), as well as to accentuate specific words or sequences. During contemplative phases, eyebrows might either rise or lower. These movements play a pivotal role in conversations, representing the most prevalent and significant facial gestures (Chovil [1991]). Variations in fundamental frequency ( $f_0$ ) and eyebrow movements are closely intertwined during speech (Cavé et al. [1996]). However, it's important to note that these variations aren't directly linked; instead, they stem from linguistic and conversational choices. These movements also serve the purpose of reaffirming the speaker's engagement by ensuring the listener's continued attention. Furthermore, eyebrows serve as a reflection of the listener's level of comprehension, doubling as a nonverbal backchannel (Cavé et al. [1996]).

### 2.1.2 Prosodic information

Nonverbal behaviour extends beyond visible signals such as posture, gestures, and facial expressions; it also encompasses auditory cues, specifically the intonation and tone of our speech. In this context, we are referring to "prosody", which goes beyond the content of speech and carries its own set of meaningful information.

The term "prosody" encompasses all suprasegmental aspects of speech, as defined by Xu (Xu [2019]). It extends beyond the mere lexical meaning of an utterance and imparts important supplementary information. Prosody serves to imbue spoken words with additional significance and to maintain the engagement of the listeners. It entails various elements, including emphasizing specific words, employing variations in voice pitch, adjusting voice loudness, modulating intonation patterns, and utilizing different voice timbres.

"Prosody" encompasses a range of sounds with varying frequencies and intensities. Deciphering these sounds reveals that some carry meaningful speech, while others convey emotions or interpersonal attitudes (Argyle [2013]). In the act of speaking, we possess the capacity to modulate these aspects within our voice. We can wield a voice that is "high" or "low" in pitch, "loud" or "soft" in volume, and "fast" or "slow" in speech rate. This intricate interplay involves elements such as rhythm, stress, and intonation, bestowing a musical quality upon speech. Speech prosody is far from mere musicality; it bears a wealth of information, encompassing the speaker's emotional state and the emphasis they wish to convey. These acoustic prosodic cues that emerge during speech are integral components of language.

Beyond syntax and emphasis, prosodic signals effectively communicate emotional nuances. These signals not only offer insights into the speaker but also influence the message's interpretation. The manner in which an individual speaks can provide glimpses into their personality, age, social standing, and identity. This multitude of cues can be systematically classified.

## 2.1. MULTIMODAL HUMAN COMMUNICATION

---

Employing accents or distinctive "intonations" imparts supplementary layers of meaning to transmitted messages, such as using question intonation for statements. The evaluation of prosodic cues offers significant insight into the underlying implications of a speaker's words. A single phrase can hold vastly different meanings within varying contexts, and the employed prosodic features hold substantial sway over these meanings.

Moving forward, let's delve into the definitions of several commonly used terms when discussing prosody:

### **Pitch**

Pitch, as explained by Titze (Titze [1994]), stands out as the most prominent attribute of the voice, characterizing its "highness" or "lowness." The perception of pitch is subject to the influence of several acoustic parameters, including amplitude and resonant (formant) frequencies. Nevertheless, the fundamental frequency ( $f_0$ ) emerges as the primary determinant of pitch.  $f_0$  represents the rate at which vocal folds vibrate during phonation (Titze [1994]). It's noteworthy that pitch and  $f_0$  are frequently treated as if they are almost interchangeable, although pitch serves as a perceptual feature, whereas  $f_0$  pertains to the physical characteristics of the sound waveform.

When it comes to variations in habitual speaking  $f_0$  among individuals, these differences hinge on variations in vocal fold length and thickness. Furthermore, individuals exhibit the capacity to modulate their pitch, whether consciously or unconsciously, in diverse contexts. This modulation allows for the adaptation of pitch to specific situations and communication needs.

### **Tone**

In the realm of linguistics, "tone" refers to a modulation in the pitch of one's voice during speech. This term finds its primary application in languages known as "tone languages," where pitch plays a vital role in distinguishing words and grammatical categories. In such languages, the nuances in pitch serve as a means to differentiate words that are otherwise identical in terms of their consonant and vowel sequences. For instance, in Mandarin Chinese, the word "man" can take on the meaning of either "deceive" or "slow" based on its pitch (Laplante and Ambady [2003]).

It's important to note that in tone languages, the significance lies not in absolute pitch values but in relative pitch distinctions. These languages typically utilize a finite set of pitch contrasts, known as "tones," which operate at the syllabic level. These tones play a crucial role in shaping the meaning of words and expressions within these languages.

### **Stress**

"Stress" denotes the heightened intensity or emphasis imparted to a particular "syllable" or "word" within a spoken expression, resulting in an audibly louder pro-

nunciation. The same phrase can assume diverse meanings contingent upon which words are stressed, thereby guiding the listener's focus (Pierrehumbert [1990]).

### **Loudness**

Loudness is the perception of how intense sounds appear to be, what the audience actually perceives and it correlates with the physical strength (amplitude). The greater the amplitude of the vocal cord vibration, the louder the sound.

### **Speech tempo**

Speech tempo serves as a quantifiable metric that assesses the quantity of specific speech units produced within a specified timeframe. This parameter is recognized for its propensity to fluctuate based on a range of factors, encompassing contextual elements, emotional states, disparities between individual speakers, as well as variances among various languages and dialects.

The precision of speech tempo measurements can be significantly influenced by the presence of pauses and hesitations within spoken discourse. Consequently, it is customary to distinguish between two distinct facets: speech tempo inclusive of pauses and hesitations, termed "speaking rate," and speech tempo that excludes these interruptions, referred to as "articulation rate."

## **2.2 Human Conversation**

Conversation stands as one of the most prevalent applications of human language. It serves as a means through which individuals socialize, foster relationships, and sustain connections with each other. While verbal communication is integral to conversations, the dynamics encompass much more than linguistic coding. Non-verbal aspects like eye gaze, body posture, and the contextual backdrop within which dialogue unfolds hold significant importance.

Conventional discourse has sometimes faced devaluation in scholarly pursuits, with linguists like Chomsky (Chomsky [2014]) characterizing spontaneously occurring instances of communication as flawed and impacted by nonverbal factors. Such perspectives, however, detach the linguistic system from its primary role in human communication. Considering the pivotal role conversation plays in human social interactions, it's essential to recognize it as a linguistic endeavour. Since the 1960s, a growing emphasis has been placed on conversation analysis as an academic domain (Maynard and Clayman [2003], Goodwin and Heritage [1990], Heritage [1989]).

Harold Garfinkel (Garfinkel [1991, 2023, 1964]) introduced Conversation Analysis as a framework for investigating interactive discourse, stemming from the ethnomethodological tradition within sociology. Goffman (Goffman [1964]) underscored the importance of scrutinizing everyday speech instances, a facet he



believed had been neglected. According to Goffman, speech operates within a socially organized framework, representing a system of mutually ratified, ritualistically governed face-to-face interactions. He contended that the study of speech wasn't solely about narrowly focused linguistic descriptions, but about understanding interactions governed by their own set of rules and structures, which aren't intrinsically linguistic in nature. This implies that studying language purely in linguistic terms doesn't adequately capture its practical application.

The groundwork laid by Garfinkel and Goffman spurred the evolution of Conversation Analysis, fostering an interest in exploring the orderliness of daily life. Harvey Sacks, building on these ideas in his lectures on conversation in the early 1960s (Sacks [1992]), devised an approach that investigated social order emanating from everyday talk practices. With the contributions of Harvey Sacks, Emmanuel A. Schegloff, and Gail Jefferson, Conversation Analysis evolved into an independent area of inquiry, influencing multiple social science disciplines dealing with human communication (Lerner [2004]). Rooted in ethnomethodology, Conversation Analysis aimed to comprehend the mechanisms of achieving order in social interactions, employing empirically based, micro-analytic methodologies (Maynard and Clayman [2003]).

Sacks posited that conversation inherently bore an orderly nature, consistently evident across all junctures (Sacks [1992]). This orderliness emerged from the attainment of the same outcomes using analogous methods within comparable contexts. Conversations were shaped through sets of practices, enabling speakers to execute specific actions in distinct contexts, actions recognized as such by fellow participants. Central to the exploration of Conversation Analysis was the delineation and elucidation of the skills ordinary speakers wield and rely upon in participating in coherent, socially structured interactions. In its most basic form, this aspiration seeks to delineate the procedures through which conversationalists navigate their behaviours and interpret those of others (Heritage [2013]).

To facilitate a better comprehension of the upcoming content, let's take a moment to outline the definitions of several frequently used terms.

### **Turn, turn-taking, turn-yielding, turn-holding and interruption**

A speaking turn refers to a continuous segment of speech during which a speaker communicates without significant interruptions.

Turn-taking is an organizational pattern in conversations and discourse where participants take their speaking roles one after the other, promoting a seamless transfer of conversational initiative in interpersonal communication (Sacks et al. [1978]).

Turn-yielding involves the use of audible or visible signals to indicate that the current speaker is concluding, allowing the next person to begin speaking.

Turn-holding indicates that the speaker is not yet finished with their current topic and intends to retain the speaking role for the ongoing turn.

An interruption is a speech action when one person breaks in to interject while another person is talking (Bennett [1978]).

### Overlap

In instances of overlap, participants typically employ an "Overlap Resolution Device" to swiftly address the situation: either one participant withdraws after a beat or two (e.g., syllables), or they employ tactics to vie for the turn (e.g., speaking faster or louder) (Jefferson and Schegloff [1975], Schegloff [2000, 2001]).

The turn-taking system also minimizes overlapping speech to uphold the norm of "one-speaker-at-a-time" (Sacks et al. [1978]). However, this doesn't negate the occurrence of overlapping speech, which frequently happens. Some types of overlap convey strong affiliative tendencies, such as recognition overlap, which can signal profound agreement (Vatanen [2018]). Yet, participants often view overlap as a deviation from the norm and actively address it through interactional strategies (Drew [2009], Jefferson and Schegloff [1975], Schegloff [2000, 2001]).

Overlap can arise at various structural junctures in turn construction (Drew [2009], Jefferson [1973, 1984, 1986], Jefferson and Schegloff [1975], Vatanen [2018]): (i) when a subsequent speaker enters at a point where they anticipate the ongoing speaker's content (termed recognitional onset); (ii) when a subsequent speaker initiates a turn simultaneously as the current speaker continues (transition space onset); (iii) when a subsequent speaker begins a turn immediately after the current speaker continues (post-transition onset); and (iv) when a subsequent speaker begins a turn at a point where the current speaker is evidently not near completion (interjacent onset).

Overlap occurring at these distinct points carries distinct implications. For instance, in cases of recognitional onset, the completion of a turn offers a natural opportunity for the ongoing speaker to step back, thus resolving overlap. However, when a participant launches a turn at an interjacent position, they actively assert their claim to the turn when another already holds the right, potentially necessitating more competitive approaches (Schegloff [2000]).

### Pauses, gaps, and lapses

Sacks et al. (Sacks et al. [1978]) distinguished between three kinds of acoustic silences in conversations: pauses, gaps, and lapses based on what's before and after the silence and the length of the silence. Pauses referred to silences within turns; gaps referred to shorter silences between turns, lapses referred to longer silences between turns. The statistical study of turn-taking began early, prompted by developments in telephony (Norwine and Murphy [1938]). It has become standard to represent overlaps and gaps on a single time scale called 'the Floor Transfer Offset (FTO)', in which positive values correspond to gaps, and negative values represent overlaps. (See Figure 2.3)

## 2.2. HUMAN CONVERSATION

---

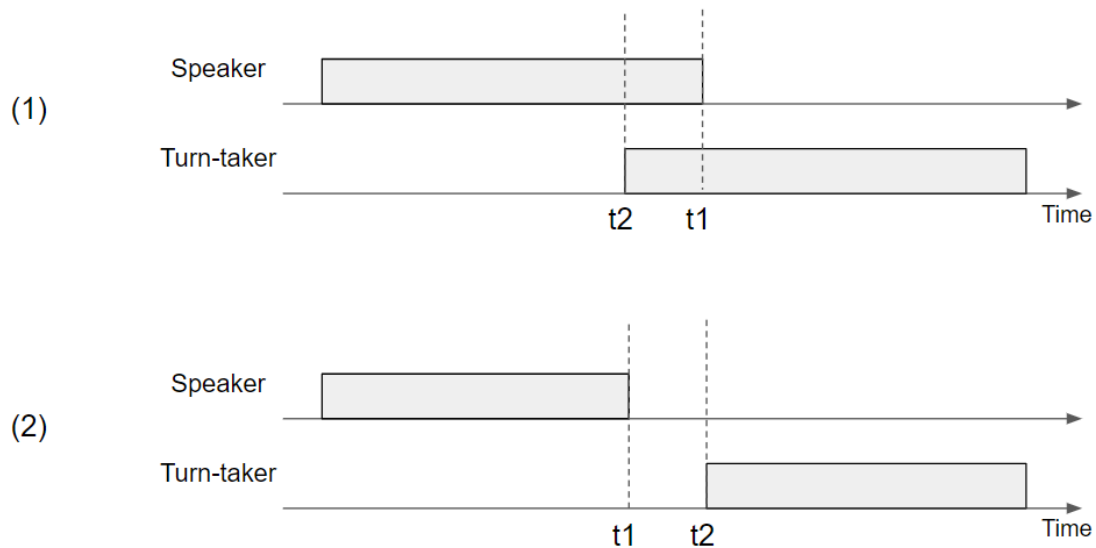


Figure 2.3 (1)  $t_2 - t_1 < 0$ , negative FTO. (2)  $t_2 - t_1 > 0$ , positive FTO.

### Inter pausal unit (IPU)

Inter pausal units are defined as speech units from a single speaker without pauses. IPUs are often defined as continuous speech delimited by silence lasting more than 200ms.

### Transition Relevance Place (TRP) & Turn Construction Unit (TCU)

A Turn Construction Unit (TCU) stands as a foundational speech segment within the realm of conversation analysis. This concept was initially introduced in the work of Sacks, Schegloff, and Jefferson in 1974 (Sacks et al. [1978]) with the aim of delineating the constituent portions of dialogue that could encompass a full turn. The identification of a TCU hinges on the search for a potential point of completion within an utterance. Three key criteria govern the delineation of what constitutes a TCU:

- **Intonationally Complete:** An utterance exhibits potential completion when accompanied by a falling intonation that serves as a signal of its conclusion.
- **Pragmatically Complete:** An utterance demonstrates possible fulfilment within the contextual framework of the ongoing conversation.
- **Grammatically Complete:** An utterance achieves possible syntactic completeness in terms of its grammatical structure.

The conclusion of a TCU is marked by what's termed as a Transition Relevance Place (TRP). This juncture serves as a pivotal point where the speaking turn can transition either to another participant in the conversation or allow the current speaker to continue with another TCU. The notion of TCU illuminates the intricate way in which conversational turns are constructed, not only considering their



structural attributes but also their alignment with the dynamics of intonation, syntax, and pragmatic context. This analytical concept provides a finer understanding of the building blocks that underlie the ebb and flow of dialogue.

### 2.2.1 Turn-taking

As mentioned previously, Conversation Analysis serves as a tool for deciphering communication patterns in social settings. It centres on conversations involving one or more speakers who alternate speaking roles. Turn-taking, a fundamental aspect of communication, refers to the exchange of speaking turns among participants. This exchange encompasses the switch between a speaker and a listener or the transition to a new speaker within a simultaneous conversation (Ghilzai and Baloch [2015]).

In this context, a "turn" or "turn-at-talk" denotes an utterance issued by a speaker with the right to speak. Investigating the dynamics of turn-taking involves delving into the linguistic and other communicative tools employed in constructing these turns-at-talk. Additionally, it entails an examination of the social mechanisms regulating the distribution and allocation of the privilege to speak.

A "turn" signifies the instance where one speaker commences speaking prior to the next speaker. In this context, the initial speaker either initiates or retains the turn for the subsequent speaker. The ensuing speaker needs to discern the expectations of the first speaker. In essence, conversation analysis aims to comprehend how participants discern and respond to fellow speakers within a discourse. As cited in Sari, Adnyani and Paramarta (Sari et al. [2021]), observed that "taking the turn can be tricky." This concept holds interest due to its influence on the overall organization of conversations.

Sacks et al. (Sacks et al. [1978]) put forth the idea that conversations are a fundamental component of social organization and are governed by social norms. Their proposed model comprises turn units and rules that are applied to these units. These turn units vary in TRP (size and can be indicated as full turns through prosody. The conclusion of such a unit is labelled as a transition relevance place). The rules they outlined are as follows:

1. If the current speaker C chooses the next speaker N, then C must cease speaking, and N should start speaking. This selection process may involve address terms, gaze, or, in the case of dyadic conversation, defaults to the other participant.
2. If C does not select N, then any participant has the opportunity to self-select, with the first person to do so gaining the right to the next unit.
3. If no other participant self-selects, C may continue speaking.

A seminal study by Emanuel A. Schegloff (Schegloff [2000]) delved into conversational analysis of turn-taking strategies, including aspects like overlapping, interruption, and prosody. The findings underscored that interruptions in conversation serve as indicators for addressing other utterances. Another study by

Young (Young et al. [2015]) delved into the conversational analysis of turn-taking within an English discussion class, revealing that students employed turn-taking strategies while acquiring a second language.

Turn-taking stands as a foundational organizational principle within human social interactions (Levinson [2016], Sacks et al. [1978]). While the specifics of turn-taking systems may differ across linguistic cultures, the fundamental rules outlined by Sacks, Schegloff, and Jefferson bear wide cross-cultural applicability (Dingemanse and Floyd [2014], Stivers et al. [2009]). Participants in conversations structure their dialogues through sequences of actions, following the norm of one-speaker-at-a-time. The turn-taking system is arguably among the few true universals in communication. Organizing discourse systematically and coherently serves as a basic prerequisite for shared understanding to thrive (Moerman and Sacks [1988], Sacks et al. [1978], Schegloff [1992]).

Complementing the turn-taking system, sequence of organization principles (Schegloff [2007]) have implications for comprehending both speech and the absence of speech at junctures where it's anticipated (i.e., non-talk) (Lerner [2019], Sacks et al. [1978]). While certain silences are permissible (Stivers et al. [2009]), the completion of a turn in which a participant designates another participant to perform a specific action (e.g., posing a question necessitating an answer) often interprets silence as the discernible omission of that action (Bolden et al. [2012], Goodwin [1979], Lerner [2003], Pomerantz [1983], Schegloff [2007], Stivers and Rossano [2010]). Such silences might be interpreted as a refusal to contribute further or as a sign of a forthcoming less favourable or intricate response (Davidson [1985], Kendrick and Torreira [2015], Robinson [2020]). Even in scenarios where a next speaker hasn't been designated, silences amid turns expected to be occupied by speech are viewed as the noticeable absence of dialogue: participants are expected to "self-select" and offer a turn (Hoey [2020]).

### 2.2.2 Backchannel

Backchannels, characterized as brief and soft vocalizations like "mm hm", "uh huh", "yeah", or nonverbal signs like head nod, smile and eyebrow movements, serve as communication cues from the listener, indicating ongoing attention and possibly conveying attitudes and uncertainties (Ward [2004]). This phenomenon has been labelled as "backchannels" (Yngve [1970]), "listener responses" (Dittmann and Llewellyn [1967]), and "accompaniment signals" (Kendon [1967]). In face-to-face interactions, backchannels can also manifest in the visual domain through actions like nodding or facial expressions. Backchannels have a unique role in the realm of turn-taking. They are relatively common, yet they don't typically qualify as a full "turn." Consequently, mechanisms need to be in place to accommodate backchannels in these analyses. Similar to how turn transitions often follow specific cues, the timing of backchannels is also connected to cues that invite backchannel responses. As indicated in (Hjalmarsson and Oertel [2012]), during face-to-face interactions, gaze direction serves as a notable cue for inviting backchannels. In these scenarios, a considerable portion of backchannel communication is nonver-

bal. Gaze, in particular, plays a crucial role in conveying various communicative functions, such as indicating objects, expressing intimacy, conveying dominance, and revealing feelings of embarrassment. These cues arise when the speaker seeks signs of comprehension from the listener (Clark [1996]). Because the listener's goal isn't to seize the conversational turn when producing a backchannel (or other forms of cooperative overlaps), it's vital for the current speaker to be able to distinguish these instances from attempts to take over the speaking turn.

### 2.2.3 Interruption

Interruption, as a distinctive manifestation of turn-taking dynamics, represents a natural occurrence in conversations and holds a significant place in the analysis of conversational structures. It reflects the interrupter's effort to seize the conversational floor before the ongoing speaker completes their utterance.

Numerous researchers have interpreted interruptions as indications of power dynamics and dominance, given their contravention of the principle of "one person speaks at a time" (Ferguson [1977], Tannen et al. [1991]). However, the interpretation of interruptions can vary based on the context and the response of the interrupted party. Some interruptions are not necessarily indicative of power dynamics or discomfort; rather, they can serve as cooperative signs aimed at assisting the speaker, such as providing cooperative completions or enhancing engagement and rhythm in the conversation (Hutchby [1996]).

Schegloff (Schegloff [2001]) distinguished between problematic and unproblematic interruptions in conversational dynamics. Problematic interruptions occur when a listener disrupts the speaker's speech with the intention of taking the floor, preventing the speaker from completing their turn. Conversely, unproblematic overlap involves a brief period of simultaneous speech where one speaker finishes their turn while another starts prematurely.

Goldberg's (Goldberg [1990]) taxonomy categorizes interruptions into two overarching strategies: competitive and cooperative interruptions. These strategies serve to shape the content and redirection of conversational exchanges. Although both strategies share certain local discourse characteristics, they play distinct roles in the broader context of interlocutor interaction. In contrast to cooperative interruptions, competitive interruptions manifest when the listener interrupts with the intention of exerting control over the interaction (Goldberg [1990]).

Distinguishing between overlaps and interruptions is a crucial distinction to make. Bennett (Bennett [1978]) emphasizes that overlaps can be objectively detected within a corpus, whereas the concept of interruptions demands a degree of interpretation—specifically, recognizing when one participant is infringing on the speaking rights of another. It's important to note that interruptions and overlaps are not interchangeable terms, despite the definitions provided earlier. Interruptions can also occur independently of overlap, for instance, when a speaker momentarily pauses (completing an IPU without yielding the turn) and the other participant begins to speak (Gravano and Hirschberg [2012]). This complexity

complicates the task of automatically identifying interruptions solely based on Voice Activity Detection (VAD) patterns. In fact, an interruption could occur without any overlap involved—essentially, taking the conversational turn following an IPU that isn't a Transitional Relevance Place (TRP). Gravano and Hirschberg (Gravano and Hirschberg [2012]), in their manual annotation of interruptions within a task-oriented dialogue corpus, observed that non-overlapping IPUs labelled as interruptions exhibited higher intensity, pitch levels, and speech rates at their onset.

### The key points of this Chapter:

#### *Multimodal human behaviour*

- *Nonverbal signals* such as body language, facial expressions, vocal tones, appearance, touch, personal space, and other physical indicators convey a substantial amount of information to those involved in communication, helping to emphasize the intended message.
- *Speech prosody cues* are integral components of language, they encompass the speaker's emotional state, attitude, certainty and the emphasis they wish to convey.
- *Facial expression* is the quintessential nonverbal conduit for conveying emotions and attitudes and can be encoded into Action Units through facial muscle movements.

#### *Human conversation*

- *Conversations* are far from being confined to the realm of language alone; they are dynamic tasks that emerge through the collaborative interplay of multiple modalities.
- *turn-taking* stands as a foundational organizational principle within human social interactions, it constitutes a central pillar in the realm of conversational analysis.
- *Backchannels* are brief and soft vocalizations like "mm hm", serve as communication cues from the listener, indicating ongoing attention and possibly conveying attitudes and uncertainties.
- *Interruption* reflects the interrupter's effort to seize the conversational floor before the ongoing speaker completes their utterance.

## Related Works

### Contents

---

3.1	Turn-taking cues . . . . .	27
3.2	Handling turn-taking . . . . .	29
3.2.1	Turn-taking in human-agent interaction . . . . .	29
3.2.2	End-of-turn detection and prediction . . . . .	30
3.2.3	Generating turn-taking cues . . . . .	32
3.3	Interruption cues . . . . .	32
3.4	Handling interruption in human-agent interaction . . . . .	34
3.4.1	User barge-in & overlap management . . . . .	34
3.4.2	Interruption prediction . . . . .	36
3.4.3	Interruption types . . . . .	37
3.5	Positioning . . . . .	38

---

This thesis focuses on the management of interruptions in interactions between human users and virtual agents. The overarching goal is to delve into the characterization of interruptions and empower virtual agents to initiate interruptions with appropriate behaviours and at suitable moments, ensuring effective communication. The preceding chapter introduced various terms pertinent to conversation analysis. In this chapter, we delve into the realm of existing interruption and turn-taking management approaches, encompassing both rule-based and data-driven methodologies. We explore the underlying principles of these approaches and shed light on their respective limitations.

## 3.1 Turn-taking cues

Despite the intricate integration of various modalities in contemporary human communication, it is essential to view the entire system as a collection of interconnected systems that have evolved over the course of approximately two and a half million years, coinciding with the emergence of humans as cognitively advanced, tool-using beings (Levinson and Holler [2014]).

The role of prosody in turn-taking has garnered significant interest and debate. Prosody encompasses the nonverbal elements of speech, including intonation, loudness, speaking rate, and timbre. It serves multiple crucial functions in conversations, such as highlighting prominence, disambiguating syntax, conveying attitude, uncertainty, and topic shifts (Ward [2019]). Ford and Thompson (Ford and Thompson [1996]) integrated intonation into their definition. Across various languages, research indicates that level intonation, occurring in the middle of the speaker's fundamental frequency range, near the end of an Inter-Pausal Unit (IPU), often acts as a turn-holding cue. This applies to languages such as English (Duncan [1972], Local et al. [1986], Gravano and Hirschberg [2011]), German (Selting [1996]), Japanese (Koiso et al. [1998]), and Swedish (Edlund and Heldner [2005]). In addition, studies on English and Japanese suggest that both rising and falling pitch can be found in turn-yielding contexts (Gravano and Hirschberg [2011], Local et al. [1986], Koiso et al. [1998]). However, research on Swedish reveals that while falling pitch serves as a turn-yielding cue, the rising pitch is not distinctly associated with either turn-holds or turn-shifts (Edlund and Heldner [2005], Hjalmarsson [2011]).

Exploring breathing in conversation, McFarland (McFarland [2001]) identified increased expiratory duration before speech onset during turn-shifts, possibly reflecting respiratory system preparation for speech production. Rochet-Capellan and Fuchs (Rochet-Capellan and Fuchs [2014]) also explored breathing as a coordination cue, finding no broad correlation between breathing and turn-taking rates or general breathing synchronization between participants. Torreira et al. (Torreira et al. [2016]) focused on inbreathes right before answering a question and noted their occurrence shortly after the question ended, suggesting a link between breathing and response planning. This suggests that breathing could be seen as a turn-initial cue, indicating that the next speaker has recognized the turn's end and is preparing a response. Ishii et al. (Ishii et al. [2014]) delved into breathing during multi-party interactions, finding that when holding the turn, a speaker inhales more rapidly and deeply than when yielding the turn. Additionally, speakers about to take the turn tend to take deeper breaths compared to non-speaking listeners.

In face-to-face interaction, eye gaze serves numerous vital communicative functions. Evolutionarily, humans have learned that others' eye gaze (and their attention) offers valuable information for coordinating activities (Tomasello et al. [2007]). Kendon (Kendon [1967]) conducted one of the earliest comprehensive analyses of eye gaze's role in turn-taking, observing video recordings of dyadic interactions. He noted a general pattern: the speaker initially averts gaze but shifts



### 3.1. TURN-TAKING CUES

---

it towards the listener at the turn's end. This observation aligns with findings from other studies (Goodwin [1981], Oertel et al. [2012], Jokinen et al. [2010]).

In multi-party interactions, gaze plays a pivotal role in selecting addressees and next speakers (Auer [2018], Jokinen et al. [2013], Ishii et al. [2016], Müller et al. [2021]). Gaze toward a participant serves as both a turn-yielding cue and a signal for selecting the next speaker. Hemamou et al. (Hemamou et al. [2019]) conducted research on significant non-verbal social cues in asynchronous job video interviews, specifically examining the connection between the recruiter's focus and specific portions of the candidate's responses. They found that increases in attention are more probable during turn-taking (at the start of the response) and turn-giving (at the conclusion of the response) which is similar to what is observed in in-person face-to-face interactions, and somehow aligned with the findings on gaze shifts.

Analyzing turn-taking cues, Duncan (Duncan [1972]) found specific gestures exerted a strong turn-holding effect. When speakers gestured, particularly with tense hand positions or movements away from the body, listeners rarely attempted to take the turn. The notion that completing hand gestures acts as a turn-yielding cue is supported by other studies (Zellers et al. [2016]). Holler et al. (Holler et al. [2018]) explored how bodily signals impact language processing in interaction, discovering that gestures accompanying questions led to quicker response times. This timing appeared to align with gesture terminations, suggesting that gestures help listeners anticipate turn endings. Sikveland and Ogden (Sikveland and Ogden [2012]) observed speakers temporarily pausing gestures during mid-turn clarifications or feedback from others before resuming their turn and gesture.

In the context of multi-party conversations, researchers have observed that the frequency of turn-taking is positively associated with cohesion among group members. High-cohesion groups tend to have more active and engaged participants who actively take turns in the conversation (Kantharaju et al. [2020]).

In the domain of communication, a sophisticated coordination of various articulators and modalities is required. Messages encompass both auditory and visual elements, spanning speech, non-speech vocalizations, and involving movements of the head, face, hands, arms, and torso, a multimodal combination of predictive features can lead to a good accuracy level for feedback timing prediction (Boudin et al. [2021]). Despite the complexity of this integration, it is noteworthy that multimodal messages tend to be processed faster than unimodal messages (Holler and Levinson [2019]). Across different modalities, turn-taking cues can be either redundant or complementary. Combining multiple cues can enhance recognition or prediction of partner intentions, potentially explaining the preference for face-to-face interaction. For conversational systems, where identifying these subtle cues poses challenges, employing various cues in combination could enhance system robustness.

# 3.2 Handling turn-taking

## 3.2.1 Turn-taking in human-agent interaction

One of the very first works in turn-taking models, as presented by Thórisson et al. (Thórisson [1998]), delves into the concept of real-time decision-making in human-computer interactions, particularly in face-to-face communication scenarios. Thórisson emphasizes the importance of real-time decision-making by computer agents involved in such interactions. In human communication, people make constant decisions about when to speak, when to listen, when to use non-verbal cues, and how to interpret the cues of others. To engage effectively in such interactions, agents should exhibit similar real-time decision-making capabilities. In addition to processing and understanding spoken language, computer agents must interpret non-verbal cues and decide when and how to respond appropriately. This involves a level of decision-making that goes beyond traditional natural language processing. Achieving seamless coordination between these modalities is crucial for effective human-agent interaction.

Cassell and colleagues (Cassell et al. [2007]) then introduced the concept of conversational coordination, which encompasses the harmonization of both verbal and non-verbal behaviours between conversational participants, extending this concept to interactions involving users and embodied conversational agents (ECAs).

In human-computer interaction (HCI), conversational coordination assumes a pivotal role in the establishment of rapport, a fundamental element of effective communication. Rapport-building with users significantly enhances the quality of interactions and overall user satisfaction. The integration of non-verbal cues, including gestures, facial expressions, and body language, holds considerable importance in facilitating communication and fostering rapport in HCI scenarios, particularly within the realm of embodied language processing. A comprehensive understanding of these dynamics is imperative for the development of highly effective and engaging computer agents, especially those designed to engage users through natural and intuitive interactions (Cassell et al. [2007]).

Cassell et al. mentioned turn-taking as an essential component of this coordination. Effective conversational coordination entails the seamless exchange of conversational turns between users and ECAs. Similar to Thórisson, Cassell et al. emphasize that ECAs need to exhibit an understanding of turn-taking dynamics, not just in terms of processing spoken language but also in interpreting non-verbal cues and making real-time decisions about when and how to respond. By incorporating effective turn-taking strategies into ECAs, rapport-building and the overall quality of human-computer interactions can be significantly improved.

In another line of research, turn-taking behaviour is influenced by the personalities and interaction objectives of those involved in the conversation. Maat et al. (Ter Maat and Heylen [2009], Ter Maat et al. [2010]) explored how the implementation of a simple communicative function, designed to manage interactions, could impact users' perceptions of an agent. They specifically investigated



how different turn-taking strategies, applied in human face-to-face conversations, could shape impressions of virtual agents in terms of personality (agreeableness), emotion, and social attitudes (e.g., friendliness).

Janowski et al. (Janowski and André [2019]) introduced a turn-taking model grounded in psychological theories that explore the interplay between an individual's personality, interpersonal stance, and diverse interaction goals. In this model, the agent's decision to speak or wait is determined by the expected utility of these actions in conveying the intended personality traits. They argued that variations in an agent's speaking style, arising from different Extraversion settings, can effectively convey the desired perceptions of its extraversion, agreeableness, and status. Personality is communicated through multimodalities and should be incorporated into embodied agents based on the specific context, adjusting their behaviour to align with user preferences (Kiderle et al. [2021]).

Fischer et al. (Fischer et al. [2021]) proposed effective in initiating interactions with people in public spaces by adapting the loudness of the robots' voice dynamically to the distance of the respective person approaching, thus indicating who it is talking to. It furthermore tracks people based on information on body orientation and adapts its gaze direction accordingly.

There are also several studies that explored the use of reinforcement learning for turn-taking strategies. Selfridge and Heeman (Selfridge and Heeman [2010]) proposed an approach in which turn-taking was treated as a negotiation process based on the perceived importance of the intended utterance. Jonsdottir et al. (Jonsdottir et al. [2008]) demonstrated how two artificial agents could develop turn-taking skills through interaction, learning to recognize each other's prosodic cues. Initially, they exhibited short pause durations and frequent overlaps, but over time, interruptions decreased, and their turn-taking patterns became more human-like. Khouzaimi et al. (Khouzaimi et al. [2015]) employed reinforcement learning to create a turn-taking management model in a simulated environment.

### 3.2.2 End-of-turn detection and prediction

Riest et al. (Riest et al. [2015]) delved into the mechanisms that underlie the human capacity for anticipation in turn-taking and investigated the diverse sources of information including prosody, syntax, context, and non-verbal cues, that contribute to this anticipation process. Listeners are indeed able to anticipate a turn-end and this strategy is predominantly used in turn-taking.

The examination of turn-taking in conversational systems primarily revolves around determining when the end of the user's turn is. This section delves into the existing approaches for this aspect. These approaches can be categorized into three types.

#### Silence-Based Models

Many systems employ an end silence duration threshold to identify the completion of a speech segment using Voice Activity Detection (VAD). After the system yields

the turn, it waits for user responses, allowing a certain duration of silence. If this silence exceeds the no-input-timeout threshold, the system resumes speaking, such as by reiterating the last question. When the user begins to speak, the end-silence timeout demarcates the conclusion of the turn. However, finding an ideal threshold to eliminate both issues entirely is often unattainable, leading to turn-taking challenges in systems following this simplistic model (Ward et al. [2005], Raux et al. [2006]).

### **IPU-Based Models.**

Operating under the assumption that the system should not initiate speech while the user is speaking, a common approach involves detecting the end of Inter-Pausal Units (IPUs) in the user's speech using VAD. This approach is somewhat similar to the silence-based model mentioned earlier but employs shorter silence thresholds, like 200ms. Once the end of an IPU is identified, the system uses turn-taking cues extracted from the user's speech to determine the presence of a Transition Relevance Point (TRP). If these TRPs are accurately recognized, the system can assume the turn with minimal gaps, while avoiding interrupting the user at non-TRPs. Bell et al. (Bell et al. [2001]) presented an early IPU-based model, that delves into the real-time management of fragmented utterances in dialogue systems. It tackles the challenge of handling spoken language that is often disfluent, fragmented, or interrupted. whereas Sato et al. (Sato et al. [2002]) introduced a more data-driven approach that also considered additional cues and employed a decision tree to classify pauses exceeding 750 ms and took the spotlight as a tool for determining turn-taking in spoken dialogue systems. Similar models were explored by Schlangen (Schlangen [2006]) to predict when a speaker will relinquish the floor and when a listener will take their turn, and Meena et al. (Meena et al. [2014]) for timing feedback responses in a map task dialogue system.

### **Continuous Models.**

In continuous models, the user's speech is processed continuously to identify suitable points to assume the turn or make projections. Incremental processing empowers the system to delve deeper into the user's utterance and make ongoing turn-taking decisions. Skantze and Schlangen (Schlangen and Skantze [2011]) introduced an early example of a fully incremental dialogue system. Skantze (Skantze [2017]) proposed a general continuous turn-taking model, trained self-supervisedly on human-human dialogue data. This model processed audio frame-by-frame (20 frames per second), employing an LSTM to predict speech activity for both speakers within a future 3-second window. Similar models were implemented by Ward et al. (Ward et al. [2018]) and Roddy et al. (Roddy et al. [2018a]), who delved into various speech features aiding predictions. Roddy et al. (Roddy et al. [2018b]) expanded the architecture by processing acoustic and linguistic features in separate LSTM subsystems with differing timescales. Recent works, such as Ruede et al. (Ruede et al. [2019]) and Hussain et al. (Hussain et al.

[2019]), have utilized LSTMs to predict backchannels and address backchannel-generation challenges using various features and reinforcement learning in contexts ranging from corpus analysis to human-robot interaction.

#### 3.2.3 Generating turn-taking cues

Up to this point, we have primarily discussed the interpretation of turn-taking cues from the user. Nevertheless, it's equally crucial to consider how to generate turn-taking cues effectively, ensuring that the user knows when it's their turn to speak and when it's not. If the system fails to manage this correctly, users may inadvertently begin speaking at the same time as the system. Several studies have explored methods for generating appropriate behaviours in animated agents to facilitate turn-taking (Pelachaud et al. [1996], Thórisson [1999], Cassell et al. [2000]).

In a study by Kunc et al. (Kunc et al. [2013]), researchers investigated the effectiveness of visual and vocal turn-yielding cues in a dialogue system that utilized an animated agent. Their findings indicated that visual cues were more successful in facilitating turn-taking than vocal cues.

Another study by Skantze et al. (Skantze et al. [2014]) delved into the impact of gaze, syntax, and filled pauses as turn-holding cues during pauses in a human-robot interaction scenario. In this context, the robot was guiding the user in drawing a route on a map placed between them. When the robot directed its gaze towards the map (rather than the user), used incomplete syntax, or incorporated filled pauses, the user was less likely to continue drawing or provide feedback compared to situations where the robot maintained eye contact with the user or used complete phrases.

Hjalmarsson and Oertel (Hjalmarsson and Oertel [2012]) examined the influence of gaze as a cue to invite backchannel responses. They conducted an experiment where participants were asked to give feedback while listening to a virtual agent tell a story. The agent typically looked away while speaking but made eye contact with the user at specific points. The results revealed that listeners were more inclined to provide backchannel responses when the agent established eye contact, although there was notable variability in their behaviour, suggesting the presence of additional significant factors.

### 3.3 Interruption cues

Numerous studies underscore the significance of prosodic features, particularly fundamental-frequency ( $f_0$ ) and intensity, as pivotal cues in conversation analysis (French and Local [1983], Kurtić et al. [2013], Truong [2013]). Further research by (Shriberg et al. [2001b], Gravano and Hirschberg [2012]) demonstrates that interrupters tend to amplify their energy and vocal tone when attempting to break into the ongoing discourse. (Hammarberg et al. [1980]) also provides similar evidence through observations of pitch and amplitude variations. The features of

### 3.3. INTERRUPTION CUES

---

interrupters, including speech rate, truncations, and repetitions, have been subject to analysis by conversational analysts. For instance, (Schegloff [2000]) identifies the use of variations in prosodic profiles and repetitions to indicate competitiveness among interrupters. This study further reveals that interrupting sentences often exhibit a faster speaking rate, shedding light on the role of speech rate in resolving speaker conflicts.

In the domain of interruption, researchers have explored the interplay between acoustic features and conversation dynamics. A study by (Gravano and Hirschberg [2012]) delves into the acoustic characteristics of telephonic conversations. They highlight significant differences in intensity, pitch level, speaking rate, and Inter-Pausal Unit (IPU) duration for interruptions compared to other turn transition instances. These findings underscore the intricate connection between prosody and the interruption phenomenon.

Gaze does not consistently serve the function of floor apportionment. Diverse studies have revealed varying outcomes regarding the gaze's role in interruption. (Cook and Lalljee [1972]) found more interruptions in the vision condition compared to the no-vision condition. (Jaffe et al. [2001]), conversely, observed fewer overlaps and shorter switching pauses in no-vision conditions. In contrast, (Argyle et al. [1968]) manipulated visibility levels and noted the highest interruption frequency in scenarios with limited visibility. Collectively, these investigations suggest that gaze might not play a central role in determining speaker turns during conversations.

Despite the limited number of interruptions observed, Kendon (Kendon [1967]) noted that speakers tend to maintain eye contact during problematic interruptions until one prevails. Brône et al. (Brône et al. [2017]) conducted investigations into dyadic and triadic conversations and found that individuals wishing to interrupt often averted their gaze before a problematic interruption and then typically initiated the interruption by making direct eye contact with the interrupted speaker.

Harrigan (Harrigan [1985]) examined verbal and non-verbal behaviours related to turn-taking and found that looking away was a prevailing strategy in problematic interruptions. Zima et al. (Zima et al. [2019]) discovered that during simultaneous speech, in 54% of cases with mutual gaze, speakers who first averted their gaze emerged as winners in the competition for a turn, and 80% of these speakers successfully completed their turn, whether it involved turn-holding (problematic interruption) or turn-yielding (unproblematic overlap). Furthermore, in 62% of interruption cases without mutual gaze, the speaker who gazed at the other speaker lost the competition for the turn, whereas, in 75% of cases where the speaker avoided another speaker's gaze, they won the competition.

Interruption frequency has been found to be linked to the relationships between interlocutors. Tannen et al. (Tannen [1994]) suggest that interruptions can serve as indicators of cohesion within a group. The occurrence of overlaps and interruptions during interactions has also shown a positive correlation with group cohesion. Additionally, instances of mutual gaze occurring during interruptions have been found to be positively correlated with the overall cohesion of the group (Kantharaju et al. [2020]).

(Beattie [1982]) emphasizes that interruptions are influenced by the conversation's contextual settings. For instance, in political interviews, interruptions are considerably prevalent as politicians interrupt almost twice as often as their interviewers.

Moreover, interruptions can provide insights into speech patterns and distinct speech styles linked to personality traits. For instance, (Rim [1977]) observed that high neurotic individuals experience more interruptions, while extroverts are more likely to interrupt and speak simultaneously compared to introverts.

Deliberately disregarding or being excessively attentive to these intentions can convey additional information about the context of the interaction, a participant's personality, or their attitude toward the other person. How virtual agents manage interruptions can also lead human observers to attribute similar characteristics which are linked to specific human personality traits and attitudes (Schiller et al. [2019]).

In the study conducted by Ravenet et al. (Ravenet et al. [2015]), the interruption behaviour of their agents was shaped by the interpersonal stance they held toward the speakers at that moment. The observers accurately identified the attitudes of these agents. Each agent had the capability to convey its attitude independently, regardless of the attitude of others, the arrangement of the group, or the gender of the user.

## 3.4 Handling interruption in human-agent interaction

ECAs should be aware of the ongoing dialogue, including the current speaker, the topic, and the user's intent. They should follow social rules of conversation, such as yielding the floor when appropriate, respecting turn-taking norms and adapting their behaviour based on the user's conversational style and preferences, including their tolerance for interruptions. Instead of waiting for a user's entire input to be completed before responding, ECAs should process and respond to input incrementally. This helps in handling interruptions more naturally (Cassell et al. [2000]).

### 3.4.1 User barge-in & overlap management

In the realm of human-agent interaction, dialogue systems can be implemented using either a simplex channel, where only one participant can speak at a time, or a duplex channel, allowing the system to listen to the user while it is speaking—that could result in overlapping speech. Enabling a duplex channel requires the system to incorporate echo cancellation, which eliminates its own voice from the audio picked up by the microphone. This configuration allows users to "barge in" during the system's speech, effectively interrupting the system. However, several complexities are associated with barge-in functionality.

One of the earliest studies on users' barge-in behaviour was conducted by Heins et al. (Heins et al. [1997]). Their research revealed that users often attempted to barge in without explicit notification about the possibility and did so frequently. These barge-in attempts were often observed to occur at specific points in the system's prompt, particularly at syntactic boundaries.

The concept of false barge-ins presents a common challenge when allowing barge-in functionality. False barge-ins are the detection errors possibly triggered by non-speech audio like external noise or coughing, or by users providing backchannels without intending to assume the speaking turn. If not effectively addressed, this can lead to confusion. Heins et al. (Heins et al. [1997]) proposed a potential solution: adjusting the threshold for detecting user speech based on the likelihood of interruptions. For instance, the threshold might be raised in scenarios where user interruptions are less likely and lowered when they are more probable, such as at syntactic boundaries. Furthermore, users might produce brief backchannels without an intention to take the speaking turn. Therefore, early prediction of whether incoming audio constitutes a complete utterance or not, preferably at the onset of Inter-Pausal Units (IPUs), is crucial. Neiberg and Truong (Neiberg and Truong [2011]) and Skantze (Skantze [2017]) demonstrated models trained on human-human data that make such predictions based on acoustic features.

When a user's utterance is verified as a genuine barge-in, it should be confirmed at the end of the IPU. If it turns out not to be a true barge-in, the system should ideally resume its speech. Strategies for handling such scenarios include utilizing a filled pause and then restarting from the last phrase boundary, as suggested by Ström and Seneff (Ström and Seneff [2000]). A comprehensive integration of these aspects is exemplified in Selfridge et al. (Selfridge et al. [2013]), where incremental speech recognition was employed to continuously determine whether to pause, continue, or resume the system's utterance. This model demonstrated improved task success and efficiency compared to more rudimentary approaches in a publicly available spoken dialogue system.

Furthermore, Heins et al. (Heins et al. [1997]) noted that users sometimes barge in before hearing crucial information they need. Hence, there may be scenarios where the system should not always permit barge-in, especially if it has important content to convey. Ström and Seneff (Ström and Seneff [2000]) explored how the system could communicate this to users when a barge-in attempt is detected. Adjusting the volume of the system's voice can indicate whether barge-in is allowed (lowering the volume) or not (raising the volume).

The REA agent represents one of the earliest virtual agent systems capable of handling interruptions from users (Cassell et al. [2001]). In the REA system, the agent yields the floor to the user as soon as they commence speaking or signal their intent to take a turn through gestures. In this system, users consistently succeed in interrupting the agent, regardless of the duration of speech overlap. A more sophisticated approach to managing verbal interruptions by users has been introduced in the virtual HWYD? ("How Was Your Day?") agent (Crook et al. [2012]). This system incorporates an intensity model designed to distinguish genuine interruptions from mere backchannel signals.



Furthermore, Gebhard et al. (Gebhard et al. [2019]) conducted a pioneering study using a real-time interactive agent system to investigate how various interruption handling times impact users' perceptions of the ECA. The findings from this study revealed that users evaluate the agent as less dominant, more friendly, and emotionally closer when the agent responds promptly to interruptions.

Cultural diversity also plays a significant role in shaping the dynamics of communication, Bennett et al. (Bennett et al. [2023]) developed a bilingual virtual avatar capable of autonomous speech during cooperative gameplay with a human participant in a social survival video game. The results showed significant differences between English and Korean speakers during the experiment. Korean speakers spoke less on average and had more negative speech sentiment, while the English speakers spoke more frequently and had more positive speech sentiment. The avatar was also more likely to interrupt the human's speech in English than in Korean, despite having the same design, the human user was more likely to interrupt the agent when it was less "chatty".

Several previous studies have delved into the intricate management of overlapping speech during interactions. One notably adopted strategy employed by certain commercial dialogue systems, as highlighted in Raux et al.'s work (Raux et al. [2006]), involves the agent responding to overlaps by simply disengaging temporarily. Later, more advanced attempts have been undertaken to model and handle overlapping speech behaviours, as evidenced by research conducted by Devault et al. (Devault et al. [2009]), Selfridge et al. (Selfridge and Heeman [2010]), and Zhao et al. (Zhao et al. [2015]).

These advanced approaches often incorporate incremental parsing techniques to gradually construct a partial understanding of the ongoing utterance. This incremental comprehension allows these models to identify opportune moments to assume control of the conversation (Skantze and Hjalmarsson [2010]). Incremental models have proven useful not only for managing overlapping speech but also for generating collaborative completions (Baumann and Schlangen [2011], Devault et al. [2009]) and providing timely feedback (Skantze and Schlangen [2009]) during a human speaker's turn.

#### 3.4.2 Interruption prediction

Cyra et al. (Cyra and Pitsch [2017]) conducted a study examining the practices of managing extended user utterances during human-agent interaction, particularly focusing on turn increments in the context of a speech-based assistive system. This investigation involved contrasting case analyses that explored the strategies employed by a human wizard to address long utterances. The findings highlighted the intricacies of human-human interaction, such as precise timing, which can be challenging to implement in technical systems.

The preliminary analysis identified two fundamental strategies for handling long utterances: interruption and wait-and-see. However, a more detailed examination from the participants' perspective revealed an additional dimension. It was observed that user acceptance of these strategies depended on whether their in-

put was acknowledged and the ongoing action was continued. Surprisingly, even interruptions, when strategically timed, were socially accepted by users, while the seemingly safer wait-and-see approach could lead to issues if the collaborative task was not continued by the system.

Several studies have focused on predicting interruption occurrences. Lee and Narayanan (Lee and Narayanan [2010]) utilized a hidden conditional random field (HCRF model) to forecast interruption instances in dyadic conversations. They identified cues for interruption prediction:

- For the Interrupter: mouth opening distance, eyebrow and head movement
- For the Interruptee: energy and pitch values of audio

These authors classified turn transitions into smooth transitions and interruptions. Their model anticipated upcoming turn exchange types based on interrupter and interrupted behaviour one second prior to the transition.

Chýlek et al. (Chýlek et al. [2018]) aimed to predict speaking turn switch timing. They categorized overlaps into three types: internal overlap (INT), overlap resulting in turn switches (OSW), and clean turn switches (CSW). INT corresponds to when speaker B starts speaking during speaker A's utterance but A continues. OSW occurs when A ends during overlap, and CSW involves no overlap. They evaluated various ML models and found that deep residual learning networks (ResNet-152) with acoustic features performed best.

#### 3.4.3 Interruption types

Research has also delved into distinguishing types of interruptions, particularly cooperative and competitive ones. The competitive interruption strategy becomes evident when the listener interrupts to assert dominance or control over the ongoing interaction. Conversely, a cooperative interruption typically contributes to the maintenance of the conversation (Goldberg [1990]).

According to (Yang [2001]), competitive interruptions are characterized by higher pitch and intensity levels, while collaborative interruptions tend to exhibit a relatively lower pitch level. (Lee et al. [2008]) propose a multimodal analysis technique for classifying interruption types, revealing that the absence of hand motions often indicates the occurrence of cooperative interruptions. Furthermore, the frequency of disfluencies in speech is notably higher in cases of competitive interruptions. Their classification approach combines hand motion with speech intensity to yield optimal results. Additionally, scholars such as Mondada (Mondada and Oloff [2011]) have explored the relationship between overlap types and accompanying gestures of continuity or abandonment.

Some studies focused on classifying interruption types, particularly cooperative and competitive interruptions. Khiat and colleagues (Truong [2013]) used SVM to classify interruptions with acoustic features, gaze behaviour, and head movement annotations. Chowdhury et al. (Chowdhury et al. [2015]) classified competition and cooperative interruptions using a Sequential Minimal Optimization (SMO)



model with prosody, voice quality, MFCC, energy, and spectral features. Egorow et al. (Egorow and Wendemuth [2019]) incorporated emotion dimensions with acoustic features for SVM-based interruption classification.

Cafaro et al. (Cafaro et al. [2016]) conducted a study to investigate the impact of different interruption types on the perception of engagement. Interruptions can take on a cooperative nature when the interrupter actively participates in the ongoing conversation by seeking clarification or expressing agreement. Conversely, interruptions can be considered disruptive when the interrupter exhibits behaviours like disagreement or changing the topic of conversation.

The findings of Cafaro et al. (Cafaro et al. [2016]) revealed that employing a cooperative interruption strategy, such as completing the speaker's sentence or asking a clarification question, was associated with a perception of increased engagement and greater involvement in the interaction. This suggests that cooperative interruptions, aimed at enhancing affiliation and fostering liking or friendliness, tend to contribute positively to the perception of engagement in a conversation.

## 3.5 Positioning

Previously, numerous studies have focused on making human-agent interactions more natural and seamless. Many of these studies delved into the realm of turn transitions, aiming to develop models that allow Embodied Conversational Agents (ECAs) to predict when a human user is about to finish their current turn and then take the conversational floor once the user has completed their turn.

To equip ECAs to handle interruptions initiated by human users, research has been conducted on detecting true interruptions, predicting human user interruptions, classifying interruption types, and studying interruption response times. These studies aimed to understand how various factors influence the user experience.

All of these research efforts started by observing and analyzing human conversational behaviours. They sought to simulate the natural patterns of human dialogue and apply them in ECA user experiments. The focus has largely been on ensuring that ECAs respond correctly to human behaviour to facilitate smooth interactions. However, these studies often overlooked the idea that ECAs could also initiate actions to guide and correct the interaction actively, engaging the user by, for instance, interrupting when necessary to take conversational turns.

Interruptions are common in human conversations; they can adjust the rhythm of dialogue. However, in current human-agent interactions, when agents interrupt human users, it is typically perceived as a system error that disrupts the interaction process. Thus, we asked, what if an agent could choose the right moment and method to interrupt? Would the interaction experience be different? Would such interruptions be acceptable to human users?

To our knowledge, there is currently no research addressing this specific question. Therefore, drawing on methods from previous studies on turn transitions and

### 3.5. POSITIONING

---

responses to interruptions, we began by analyzing human interactions. Building upon the previously mentioned turn-taking cues and interruption cues, we have opted to focus on nonverbal behaviour to dissect the actions of interlocutors. We categorized interruptions, analyzed behavioural patterns during interruptions in human interactions, and used real human interaction data to develop computational models that empower agents to initiate appropriate interruptions.

# Corpora

## Contents

---

4.1	Corpora: NoXi / AMI . . . . .	41
4.1.1	Description of Noxi Corpus . . . . .	41
4.1.2	Description of AMI corpus . . . . .	41
4.2	Features . . . . .	42
4.3	Data Cleaning . . . . .	46
4.3.1	Video Processing . . . . .	46
4.3.2	Audio Processing . . . . .	48
4.4	Conclusion . . . . .	48

---

In this Chapter, we present two corpora we have chosen to use.

The first one, the NoXi corpus Cafaro et al. (Cafaro et al. [2017]) is the foundation of our human interaction study. It is employed to train, test, and validate our models predicting interruption timing and generating facial gestures during the interruption period.

The second corpus built by Carletta et al. (Carletta [2007]), is used to train, test, and validate the model classifying interruption types. *AMI corpus* was initially created to develop meeting browsing technology. We extend it to include additional multimodal features related to *facial expression* and *low-level acoustic features*.

## 4.1 Corpora: NoXi / AMI

### 4.1.1 Description of Noxi Corpus

*NoXi Corpus* is a comprehensive collection of multimodal data, consisting of video and audio recordings capturing free dyadic interactions. Each interaction involves two participants who have been recorded separately, allowing easy separation of the audio sources. The setup of screen-mediated interactions facilitates clear separation of both audio and video flow that have been synchronized and transcribed. The video recordings capture almost the entire body of each participant, except for their feet, as depicted in Figure 4.1.

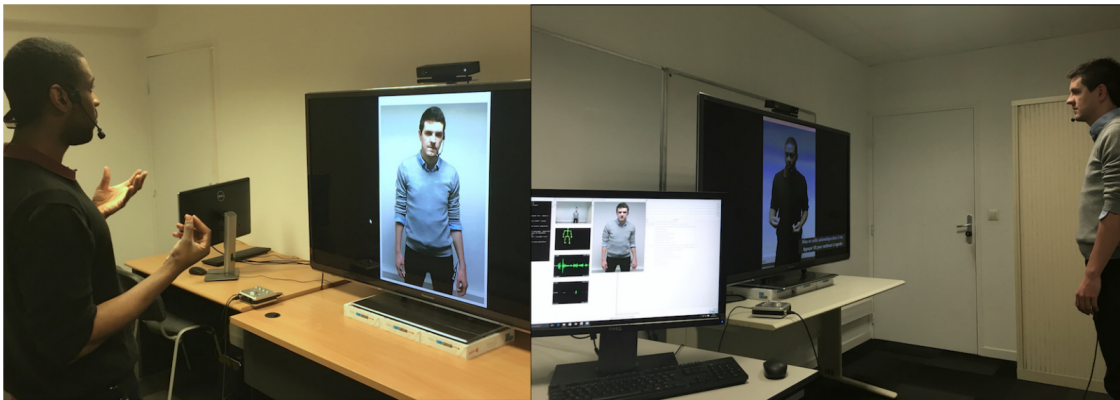


Figure 4.1 An example of NoXi dyadic conversation. Figure from (Cafaro et al. [2017])

Within the NoXi database, participants assume either the role of an "expert" or a "novice." The expert shares the knowledge with a novice on a specific subject, selected from over 45 given topics. Each interaction lasts approximately 20 minutes, resulting in a rich dataset of diverse and informative interactions.

The NoXi corpus encompasses data recorded in seven different languages. For our study, we specifically selected the French portion of the NoXi corpus, which includes 21 dyadic conversations, totalling approximately 7 hours of recorded interactions ( $21 * 20$  minutes). The videos are recorded with a frequency of 25 fps.

The NoXi corpus encompasses a wide range of modalities, including video, audio, and transcriptions. Through the meticulous examination of these modalities and their interplay, we can uncover patterns, subtleties, and multimodal cues that contribute to understanding how interruptions are perceived and managed within dyadic interactions.

### 4.1.2 Description of AMI corpus

The AMI Corpus (Carletta [2007]) is a multimodal database that presents a rich collection of 100 hours of free-flowing multi-party English meetings (as in Figure

## 4.2. FEATURES

---

4.2). Within the AMI Corpus, each meeting involves four participants who are engaged in discussions centred around specific topics. These interactions are carefully structured, lasting approximately 25 minutes each, recorded with a frequency of 15 fps.

The video recordings of the participants capture their upper body movements and expressions, facilitating the analysis of nonverbal communication cues. In this thesis, this corpus is only used to develop and evaluate the interruption classification model presented in Chapter 6.



Figure 4.2 AMI corpus: four-party English conversation

For our specific work, we focus on nine carefully selected meetings from the AMI Corpus, which amount to approximately four hours of conversation data. These particular meetings have been extensively annotated for essential aspects such as head function and focus of attention, the annotations are provided with original video and audio data.

## 4.2 Features

NoXi is designed to provide spontaneous and natural interactions. A lot of annotations and information are provided with the database through a web interface. Descriptors include low-level social signals (e.g. gestures, smiles), functional descriptors (e.g. turn-taking, dialogue acts) and interaction descriptors (e.g. engagement, interest, and fluidity).

AMI Meeting Corpus is created for the uses of a consortium to develop meeting browsing technology, it is useful for a wide range of research areas. Various annotations are provided with the meeting data, such as dialogue acts, disfluency, focus of attention, head and hand movement.

## 4.2. FEATURES

<i>Features</i>	<i>Collection Tool</i>	<i>Available Representations</i>
<b>Audio</b>	<b>OpenSmile</b>	Fundamental Frequency - $f_0$
		Loudness
		Voice probability
		Mel-frequency cepstral coefficients - <i>MFCC 0-12</i>
		Logarithmic harmonics-to-noise ratio - <i>logHNR</i>
		Jitter
		Logarithmic signal energy from pcm frames
<b>Action Units (AUs)</b>	<b>OpenFace</b>	Energy in spectral bands
		Shimmer
		AU1 - Inner Brow Raiser
		AU2 - Outer Brow Raiser
		AU4- Brow Lowerer
		AU5 - Upper lid raiser
		AU6 - Cheek Raiser
AU7 - Lid Tightener		
<b>Head Motion</b>	<b>OpenFace</b>	AU12 - Lip corner puller
		AU15 - Lip corner depressor
		Roll Euler angle - $R_x$
		Pitch Euler angle - $R_y$
		Yaw Euler angle - $R_z$
<b>Upper body Motion</b>	<b>AlphaPose</b>	Head position - $P_{x,y,z}$
		Gaze direction - $R_{x,y}$
		7 key points position of upper body

Table 4.1 Features utilized in our research.

While these annotations may enhance the accuracy of interruption classification models and decision models for timing, our ultimate goal is to enable real-time detection and generation of interruptions in human-agent conversations. Considering the challenges of real-time annotation, we have decided to forego the use of these annotations. Instead, we opt for the utilization of low-level features that can be readily extracted in real-time.

In this thesis, we focus on nonverbal signals such as facial action units, speech prosody and turn exchanges, to develop models capturing multimodal patterns exchanged during interaction and synthesizing human-like and expressive facial gestures.

As mentioned in Chapter 3, numerous studies have emphasized the significance of acoustic features in analyzing interruptions. Hence, we employed Opensmile to

## 4.2. FEATURES

---

extract pertinent acoustic features. Simultaneously, we also paid attention to other modalities during interruptions. Therefore, we utilized OpenFace to capture facial expressions and head movements of the conversational participants. Additionally, we used AlphaPose (Fang et al. [2022]) to extract body movements of the participants.

It's worth noting that certain facial action units are significantly influenced by speech content, which falls outside the scope of our study. Consequently, we excluded these particular action units from our analysis. In this section, we will exclusively enumerate the features utilized in our research, which are presented in Table 4.1 and further discussed in the following sections.

### Audio Features

The audio features we are considering in this corpus are prosodic and voice quality features.

More specifically, we consider at each time-step the fundamental frequency  $f_0$ , the loudness and the 13 Mel-frequency cepstral coefficients (MFCC 0-12).  $f_0$  variations capture pitch changes, which are essential for conveying intonation and melodic contour. The mel frequency cepstral coefficients (MFCCs) of a signal concisely describe the overall shape of a spectral envelope, they are often used to describe timbre.

The above acoustic features were extracted with OpenSmile (Eyben et al. [2013]) with a frequency of  $100fps$ . In this corpus,  $f_0$  values were restricted to the range of 50 to 550Hz, which is enough to enclose the vocal ranges of both male and female speakers. In fact, the vocal speech of a typical adult male has a  $f_0$  ranging from 85 to 180 Hz. That of a typical adult female ranges from 165 to 255 Hz (Baken and Orlikoff [2000], Titze [1994]).

### Action Units Features

Facial expressions are represented by "Action Units"(AUs), as defined in the Facial Action Coding Systems (FACS) manual that was developed by Ekman et al. [1971]. The AUs we study in the scope and context of this thesis are the ones related to *eyebrows* and *lip corners* movements. We do not consider the other facial muscles involved in articulatory movements, since we do not model them in the scope of this thesis. *Eyebrows* and *lip corners* motion are represented by the seven action units AU1, AU2, AU4, AU5, AU6, AU7, AU12 and AU15, listed and described in Table 4.1.

We extracted facial expression action units, head motion and gaze direction using the tool OpenFace (Figure 4.3, Baltrušaitis et al. [2016b]) with a frequency of  $25fps$  for NoXi corpus and  $15fps$  for AMI corpus.



## 4.2. FEATURES

Action Units are represented by values of intensity, which is a measure of how strong the activation of facial muscles is. In OpenFace, *AU* intensities are continuous values ranging from 0 - *lowest* intensity - to 5 - *highest* intensity. At each *AU* intensity is associated a “Success” score and a “Confidence” value. The “Success” score equals to 1 if OpenFace has detected the speaker’s face, 0 otherwise while the “Confidence” value, between 0 and 1, represents the confidence level of OpenFace.

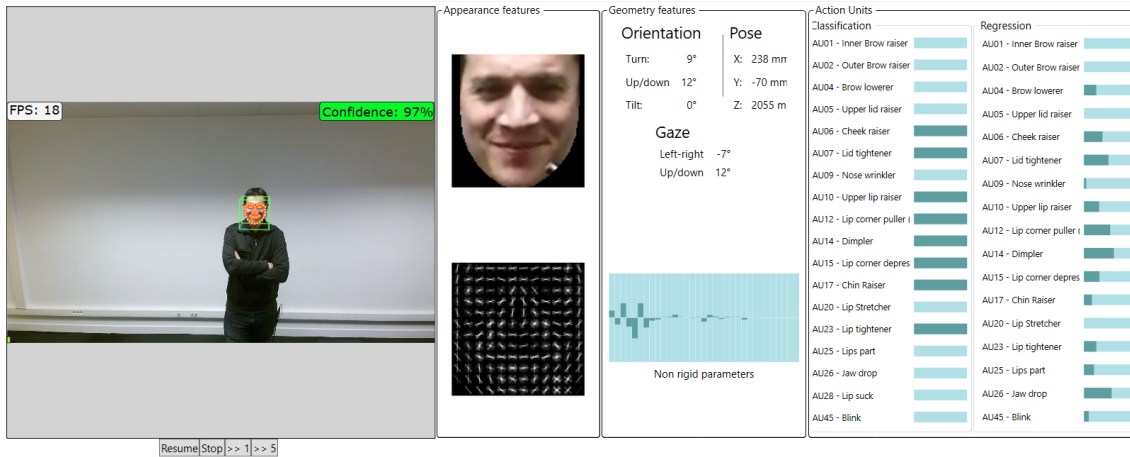


Figure 4.3 Open Face AU detection

### Head motion features & gaze direction

Head motion is represented by 3D head angles and 3D positions relative to the camera.

Head rotations have three degrees of freedom, represented by the Euler angles: *roll*, *pitch* and *yaw*. The angles are represented by  $R_x$ ,  $R_y$  and  $R_z$ , which are the rotations of the head with respect to  $x$ ,  $y$ , and  $z$  axes. Head rotation and position were also extracted using the tool OpenFace. Each frame of the video has a success score and confidence level.

Head positions are represented by  $P_x$ ,  $P_y$  and  $P_z$ , which are the positions of the head with respect to  $x$ ,  $y$ , and  $z$  axes (The origin is situated at the position of the camera.). To avoid the bias caused by the interactant’s initial position, we do not use absolute position but the head motion activity, defined at each time-step  $i$  using the following equation:

$$v_{Head}(i) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2} \quad (4.1)$$

Gaze direction is represented by  $R_X$  and  $R_Y$ , which are the relative horizontal and vertical angles regarding the camera.

### Hand motion features

To be invariant to the body position in the image, instead of using the absolute right and left-hand positions provided by Alphapose Fang et al. [2022] with the



frequency of videos, we centre the position by taking the middle of the two shoulders as the origin of the coordinate system (0,0). Moreover, for each video, we pick one frame when the interactant is facing the camera and note the distance between the two shoulders (*scale*). Then, the coordinate system is changed using a normalisation of  $x$  and  $y$  such as  $scale = 1$ . This is used to calculate the *scaled joint position*.

After scaling the joint position, the right and left-hand activity are computed. They can be interpreted as the amount of motion of each hand and are estimated using:

$$v_{Hand}(i) = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \quad (4.2)$$

where  $x_i$  and  $y_i$  are the coordinates of the hand (right or left) at time-step  $i$ .

Due to the arrangement of participants in the AMI corpus, where they sit around a table and lack frontal camera angles to provide reliable hand motion data, we opted not to extract hand movements of the participants in the AMI corpus.

## 4.3 Data Cleaning

The main purpose is to provide clean, structured, and aligned multimodal features of NoXi and AMI corpus for the following studies. As we started extracting the multimodal features from the raw videos and audios, we identified several situations where the extracted features were either missing or very noisy. We list the major undesired situations:

- During the recording, interlocutors may turn their heads to the side so that OpenFace cannot detect the entire face.
- The interlocutors in the conversations are recorded separately but sometimes the voice of other interlocutors or background noises are also recorded in the audio, which have to be filtered out.

We worked out a series of solutions to overcome these issues, which are explained in the following sections.

### 4.3.1 Video Processing

After extracting the facial expressions and head motions, the intensity values extracted by OpenFace are noisy and may be missing for some frames. We applied additional data smoothing and fitting techniques to further eliminate noise and fill in the gaps by determining the missing values.

**Linear interpolation.** We identified some cases where OpenFace did not detect well the speaker’s face, and for these cases, *Success* score was equal to 0 and no intensity values were extracted. For this reason, we further applied linear interpolation, to fill in the gaps where OpenFace failed to detect the interlocutors’ faces, as shown in Figure 4.5.

### 4.3. DATA CLEANING

---

**Median Filtering.** *Median filtering* is a smoothing technique that is frequently used to remove noise from an image or a signal. For each extracted AU intensity, body and head motion feature, we applied a *median filter* to remove noises. Figure 4.4 depicts the effect of applying a median filter with window sizes equal to 3, 5 and 7 to  $pose_{Rx}$  (head rotation on the x-axis) for the purpose of filtering out noises while preserving and highlighting edges. After testing with different window sizes, we applied the median filter with a *window size* equals to 7, since it eliminates noise and maintains the edges.

**Z-score normalization.** *Z-score normalization* scales the values of a feature to have a mean of 0 and a standard deviation of 1 by subtracting the mean of the feature from each value and then dividing it by the standard deviation. To eliminate redundant data and minimize personal biases, we apply a z-score normalization for head and body motion.

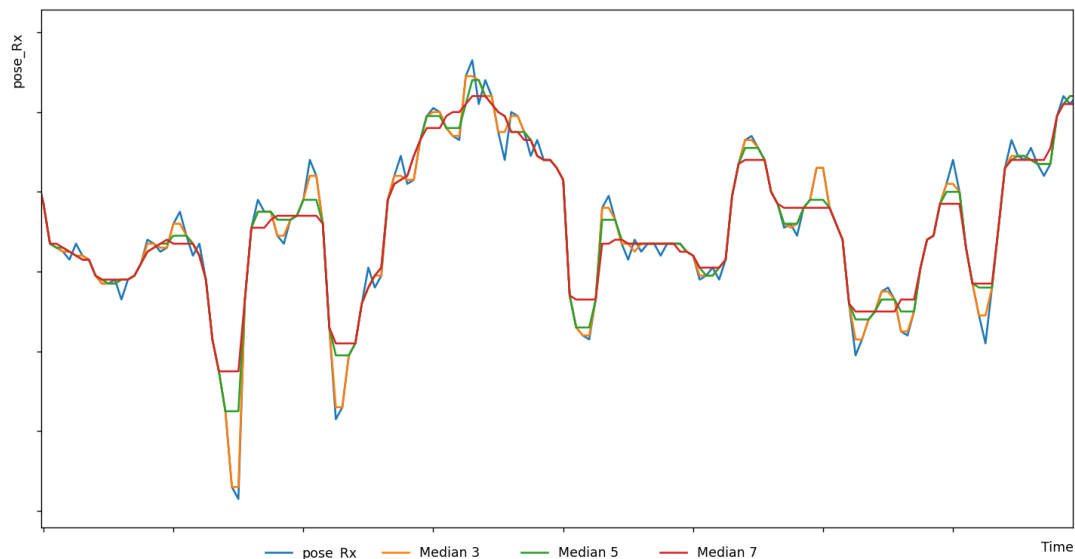


Figure 4.4 This figure is a plot of different median filtering window sizes applied to  $AU1$  signal.

## 4.4. CONCLUSION

---

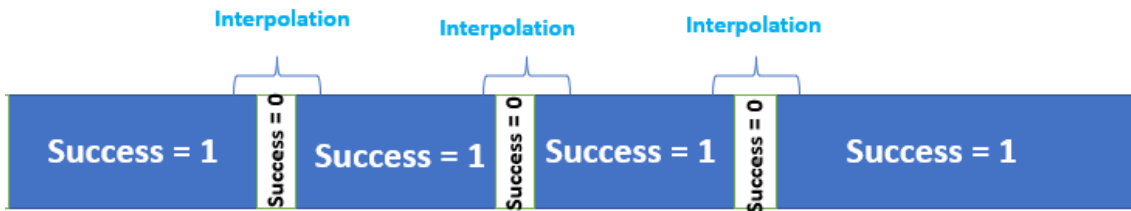


Figure 4.5 Linear Interpolation is applied on the frames where OpenFace’s *success* score is equal to 0.

### 4.3.2 Audio Processing

To avoid the impact of noises, we first filtered out these noises with Audacity, given a sample of the noise from other interlocutors, Audacity captures automatically the characteristics of the noise and reduces it over the full audio episode. This solves most of the problem, especially when two or more interlocutors speak at the same time. Acoustic features from OpenSmile are also z-score normalized.

At the same time, facial expressions, head motions, and hand motions are extracted at a frame rate of the videos (25 frames per second for NoXi, 15 frames per second for AMI), while acoustic features are extracted at a higher frame rate of 100 frames per second. In order to ensure compatibility and synchronization among all these features, we performed undersampling on the acoustic features. This involved averaging the acoustic feature values every 4 frames, aligning them with the slower frame rate of facial expressions and motions. By harmonizing the frame rates through undersampling, we create a consistent temporal framework that enables a unified analysis of these diverse features in relation to one another. This adjustment allows us to effectively examine the interplay between visual and acoustic cues, facilitating a comprehensive understanding of the interaction dynamics.

## 4.4 Conclusion

In this Chapter, we presented *NoXi Corpus* and *AMI Corpus*. *NoXi Corpus* presents a large amount of data that includes speech audio features, facial expressions, and head motions. *AMI Corpus* includes multimodal features and high-level annotations regarding head movements. We also presented the process of data cleaning on extracted features.

##### The key points of this Chapter:

###### *Corpus*

- *NoXi Corpus* is designed to provide spontaneous interactions with emphasis on adaptive behaviours and unexpected situations. We choose to utilize the French part of *NoXi Corpus* as the foundation of our study.
- *AMI Corpus* is a multimodal database that presents a rich collection of 100 hours of free-flowing multi-party English meetings, we focus on nine conversations that have been extensively annotated for essential aspects such as head function and focus of attention.
- *NoXi Corpus* and *AMI Corpus* consists of a large amount of *multimodal features*, which are: facial Action Units, Head Positions and Rotations, Gaze Direction, Voice Prosody and Body motions.

###### *Data cleaning*

- We filter the noises of the interlocutors' audios and extract low-level acoustic features with OpenSmile.
- We extract visual features with Openface and Alphapose, then apply a data preprocessing to clean the data and prepare it for the following studies.
- We applied undersampling to the acoustic features in order to ensure synchronization with visual features.

# Annotation schema & multimodal analysis

## Contents

---

5.1	Related Works . . . . .	51
5.2	Annotation Schema . . . . .	53
5.2.1	Schema . . . . .	53
5.2.2	Process . . . . .	56
5.2.3	Annotation accuracy . . . . .	57
5.3	Statistical results . . . . .	58
5.4	Multimodal Analysis . . . . .	60
5.5	Conclusion . . . . .	63

---

This thesis delves into the realm of interruptions within the context of human-agent interaction. With the intention of generating authentic exchanges between humans and Embodied Conversational Agents (ECAs), a critical aspect involves equipping these virtual agents with the capability to manage interruptions effectively, which is to initiate interruptions as necessary. To achieve this, we start with a study on human-human interaction that involves annotating, analyzing, and delineating the characteristics of interruptions to distinguish them from other forms of exchanges.

When an ECA interacts with a human, it must adapt to the dynamics of the conversation, especially in terms of interruptions. The ECA should be capable of measuring its human interlocutor’s intentions and reactions through multimodal signals. This includes recognizing when the human is actively listening, possibly indicated by backchannel signals, or when the human is attempting to seize the speaking turn, potentially involving interruption.

We delve into the study of natural speaking turn exchanges with the dyadic corpus known as NoXi (Cafaro et al. [2017]) as presented in 4. Our approach involves formulating an annotation schema that encompasses various types of interruptions. Furthermore, we conduct a comprehensive analysis of multimodal features, placing particular emphasis on prosodic attributes such as fundamental frequency ( $f_0$ ) and loudness. Additionally, we delve into facial expressions utilizing Action Units (AUs), as well as body movements encompassing head and hand gestures. Our objective is to unravel the intricate patterns of nonverbal behaviours that emerge during different types of turn switches thereby clarifying the distinctive characteristics that the ECA should be trained to identify and respond to, which correspond to the research question Q1.

## 5.1 Related Works

Interruption is a special phenomenon within the framework of turn-taking principles, which serves as both a natural occurrence and a compelling topic within the realm of conversational structure analysis. The work of Allwood and colleagues has contributed significantly to this discourse by incorporating interruption into their coding schema, categorizing it based on its functional role (Allwood [2001]). However, their categorization of interruption solely covers instances of speech overlap, thus excluding interruptions initiated within a pause.

Previous studies have engaged diverse conversation coding structures in their investigations (Nakazato [2000], Truan and Romary [2021], Ten Bosch et al. [2004], Christodoulides and Avanzi [2015], Jokinen et al. [2013], Enomoto et al. [2020], Heeman et al. [2006]). However, these studies have not delved into the finer distinctions among various interruption types, leaving a significant gap in understanding the nuances of this complex conversational phenomenon.

Schegloff and Sacks (Schegloff and Sacks [1973]) have laid out a comprehensive framework for categorizing simultaneous speech occurrences within conversations, encompassing three distinct types: interruption, overlap, and parenthetical comments, often exemplified by backchannels. Backchannels, while representing feedback messages, are not intended to seize the speaking turn from the current speaker.

Overlaps transpire when the listener accurately anticipates the imminent conclusion of the ongoing speaker's utterance. Here, there exists a willingness on the part of the speaker to renounce the speaking turn, fostering a harmonious transition. In sharp contrast, interruptions involve a more abrupt transfer of the speaking floor. In these instances, the listener forcefully takes control of the conversation, often against the wishes of the current speaker (Schegloff and Sacks [1973]).

To prevent any ambiguity between the concepts of overlap and interruption, Sacks and his colleagues (Sacks et al. [1978]) delineate overlap as the listener preemptively predicting the culmination of the ongoing speaker's dialogue, result-

## 5.1. RELATED WORKS

ing in an overlap between the last word or syllable of the current speaker and the first word of the listener's subsequent speech segment.

In contrast, interruption is characterized as a disruption to the current speaker's turn, a departure from the anticipatory but generally harmonious nature of overlap (Moerman and Sacks [2010]).

Presented here are two prominent methods for classifying interruptions. The first method, proposed by Beattie (Beattie [1981]), hinges on the assessment of simultaneous speech and the willingness to cede the conversational floor. This classification method is illustrated in Figure 5.1.

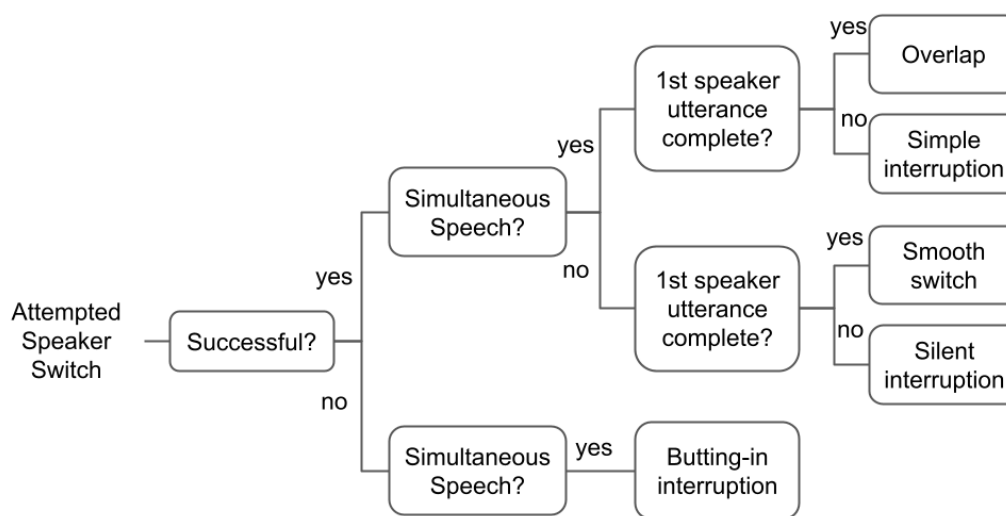


Figure 5.1 Classification of interruption and smooth speaker exchange (Beattie [1981]).

In this taxonomy, there are three types of interruptions.

- Butting-in interruption, in which there is overlap but the listener fails to grab the turn and the speaker continues to speak.
- Simple interruption, in which there is also overlap, but the listener succeeds in getting turns against the speaker's wishes.
- Silent interruption, without overlap, the listener takes turns, opposing the speaker's wishes, during a short pause.

In the alternate taxonomy presented by Li and Campbell (Li [2001]), interruptions are categorized into two overarching strategies based on their alignment with speech content: competitive and cooperative interruptions.

The competitive interruption strategy manifests when the listener interjects to seize control of the ongoing interaction. This type of interruption is often disruptive to the natural flow of the conversation between the interlocutors and may introduce an element of conflict into the discourse:

- **Disagreement:** The listener disagrees with what the current speaker is saying and expresses his or her own opinion.
- **Floor taking:** The switch does not change the current topic and usually expands on the current speaker's topic.
- **Topic change:** The listener changes the current topic of conversation.
- **Tangentialization:** The listener sums up the message from the current speaker to prevent listening to more unwanted information.

On the opposite, a cooperative interruption usually helps to maintain the conversation and can be:

- **Agreement:** The listener shows agreement, compliance, understanding or support to the speaker.
- **Assistance:** The listener interrupts to provide the current speaker with a word, a phrase or an idea to complete the utterance.
- **Clarification:** The listener asks the current speaker to clarify or explain the information about which the listener is not clear.

These two classification methodologies encompass a wide range of interruption scenarios commonly encountered in conversations. However, it's important to acknowledge that there are exceptions that fall outside their scope. For instance, the first taxonomy doesn't account for backchannels, which are brief feedback messages. Additionally, there are instances of interruptions that are abandoned swiftly, making it challenging to definitively categorize them using the second taxonomy. To address these limitations and create a more comprehensive framework, we introduce a novel structure that integrates and refines elements from both of these existing methods. This merged approach aims to provide a more nuanced and adaptable system for classifying interruptions across various conversational contexts.

## 5.2 Annotation Schema

In this section, we propose a new annotation schema and explain how we made the annotation. Then, we present annotation accuracy and some statistical results on annotated data.

### 5.2.1 Schema

The annotation schema we propose in Figure 5.2 includes all turn changes and comprises three classification levels.



## 5.2. ANNOTATION SCHEMA

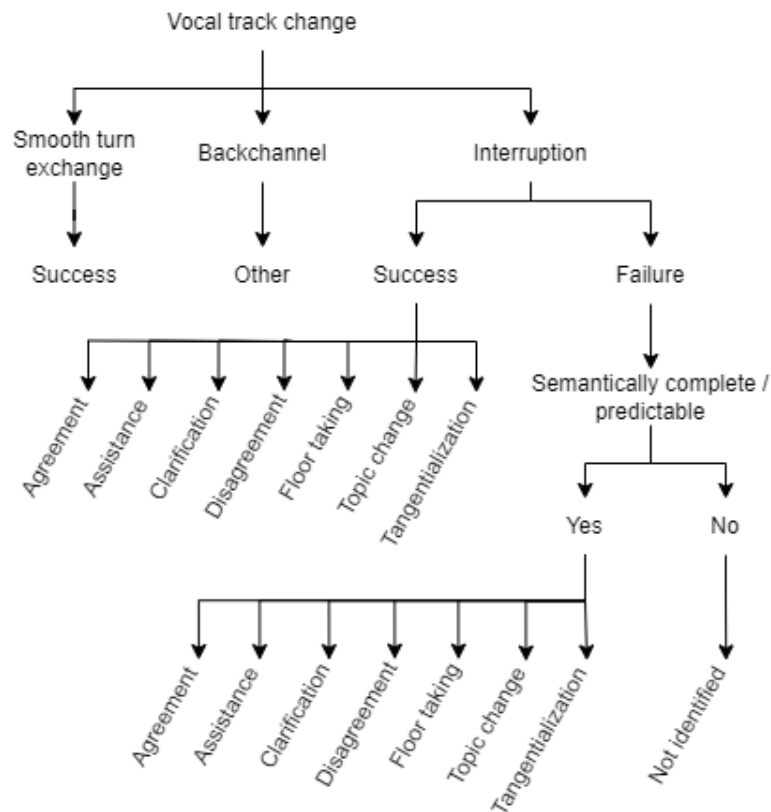


Figure 5.2 Interruption annotation schema

To begin with, our classification process involves categorizing each change in voice activity into three distinct types: *interruption*, *backchannel*, or *smooth turn-exchange*. Interruptions pertain to instances where the listener initiates a turn, either during a moment of silence or with an overlap, while the current speaker is still in the middle of their turn. On the other hand, a smooth turn-exchange occurs when the listener takes their turn as the current speaker concludes or is in the process of yielding their turn. Backchannels, which are concise messages indicating the listener’s attention or expressing agreement or disagreement with the speaker’s content (Allwood et al. [1992]), can also occur during a turn.

Importantly, it’s worth noting that differentiating between smooth turn exchanges and interruptions solely based on gaps and overlaps can be misleading. For instance, a smooth turn-exchange might involve an overlap if the listener starts speaking with overlapping words or syllables while the current speaker begins to yield the floor but has not yet completed their utterance. Conversely, an interruption might involve a gap in speech, such as when the speaker is momentarily stuck and searching for a word, and the listener offers a suggestion. Therefore, identifying interruptions requires a thorough consideration of the interaction between the participants’ speech patterns.

Backchannels serve as brief messages aimed at conveying the listener’s attention or expressing agreement or disagreement with the ongoing discourse (All-

wood et al. [1992]). These messages can vary in length, extending beyond a single word to encompass complete sentences. What sets backchannels apart from interruptions is their intent: they are not intended to seize the speaking turn or elicit a response from the speaker. Moreover, they do not disrupt the flow of speech in the same manner as interruptions.

Subsequently, we proceed to annotate the successful completion of turn exchanges. In this context, a smooth turn exchange denotes a seamless and successful transition of speaking turns (*success*). Interruptions, on the other hand, can manifest as either successful (*success*) or unsuccessful (*failure*) instances, depending on whether they effectively acquire the speaking turn. It's important to acknowledge that backchannels, while not aimed at seizing the speaking turn, are also included in this process. However, since they do not lead to turn exchange, their annotation falls under the category of *Other* in terms of accomplishment.

A successful interruption corresponds to the situation described below ('[ ' and ']' represent the start and the end of a simultaneous speech):

- The interrupter succeeds in grabbing the turn and the current speaker stops talking even though s/he has not finished his/her utterance.

*Example: Novice:*...basically not phy- [ sical but I ... *Expert:* [ I agree with you for..

- The interrupter talks over the speaker (e.g. to ask a clarification question). The speaker keeps the speaking turn but considers what the interrupter has said (e.g. by answering the interrupter's question).

*Example: Expert:* ...mushrooms and you [ have to be care- ] ful. yeah, especially the optics... *Novice:* [ Mushrooms?]

A failed interruption occurs when:

- The interrupter terminates the interruption before completing the utterance and let the current speaker continue his/her turn.

*Example: Expert:* ...your point of view, I unders- [ tand but finally ] maybe it's easy ... *Novice:* [ Ah no no you...]

- The interrupter begins to speak and tries to get the current speaker's attention, but the current speaker does not respond to the interruption after the interrupter has completed his/her utterance and continues speaking as planned.

*Example: Novice:* ...I didn't pay even one euro for Hearthstone, and I uh I still [ have my meta decks up to ] now, I can... *Expert:* [ Ah me neither I didn't pay for it. ]

In the last phase of our annotation process, we proceed to categorize each interruption based on its underlying speech content. This categorization extends to both successful and unsuccessful interruptions, aiming to capture the diversity of

## 5.2. ANNOTATION SCHEMA

interruption types. The distinct interruption types we have considered are elucidated in Figure 5.2.

In the case of failed interruptions, their type is determined when the speech duration is sufficiently long to discern the content and ascertain the intended interruption category. Conversely, when the speech duration is insufficient to grasp the intended interruption type, it is annotated as *Not identified*.

### 5.2.2 Process

The annotation process relies on the automated detection of voice activity. As previously discussed, the NoXi corpus encompasses a total of 21 dyadic conversations, with individual recordings of each participant. To facilitate the simultaneous display of video footage from both participants within each conversation, and to ensure a coherent synchronization between the visual elements and the corresponding voice activities, we employ the Nova tool (Baur et al. [2020]), depicted in Figure 5.3.

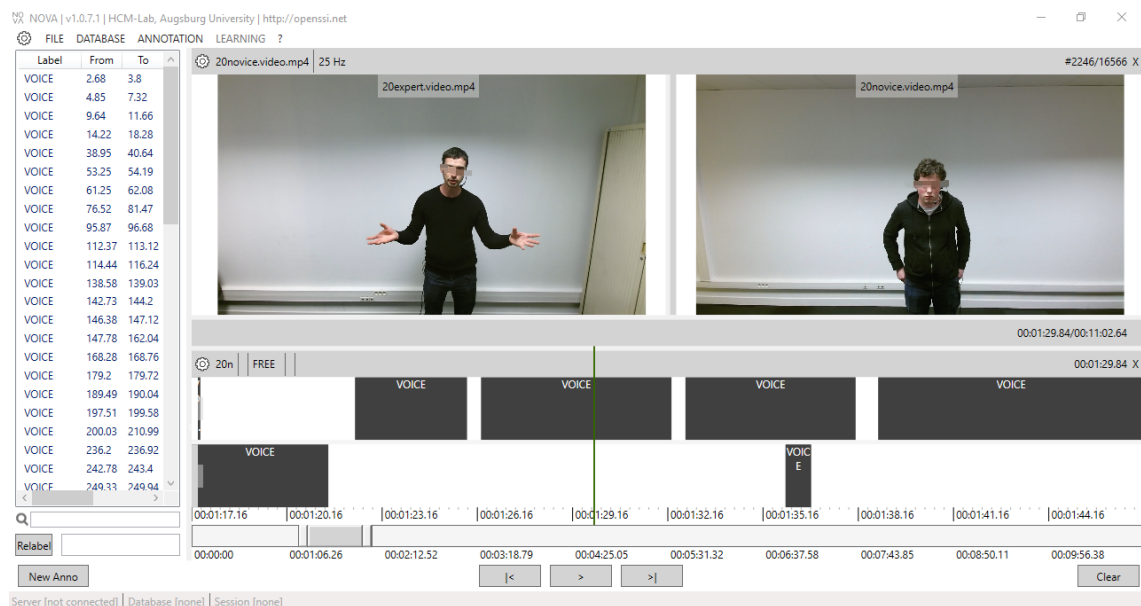


Figure 5.3 Nova annotation interface

We implement voice activity detection throughout the entire video duration. With each instance of voice activity detected from the current listener, we meticulously capture the onset time and subsequently annotate the transition of speaking turns based on the schema outlined in Figure 5.2.

Subsequently, we employ a script to automatically identify the termination point of the initial Inter-Pausal Unit (IPU) that emerges after this onset time. Inter-Pausal Units, which find extensive application in conversation analysis, refer to speech units from an individual speaker that lack pauses (Levitan and Hirschberg [2011]). As clarified by (Demol et al. [2007]), we define a pause as a period of silence surpassing 200 milliseconds, and we employ these pauses to segment voice

## 5.2. ANNOTATION SCHEMA

---

Type	Agreement	Assistance	Clarification	Disagreement	Floor taking	Topic change	Tangentialization	Not identified
Count	348	68	89	44	230	68	6	76
Percentage	37.46%	7.32%	9.58%	4.74%	24.75%	7.32%	0.65%	8.18%

Table 5.1 Probability distribution according to the 8 types of interruption.

Type	Agreement	Assistance	Clarification	Disagreement	Floor taking	Topic change	Tangentialization
Count	310	64	84	42	188	66	5
Percentage	40.84%	8.43%	11.07%	5.53%	24.77%	8.7%	0.66%

Table 5.2 Probability distribution of different interruption types for successful interruption.

activity into distinct IPU. This segmentation strategy facilitates the organization and analysis of spoken content, enhancing the accuracy of our annotation process.

### 5.2.3 Annotation accuracy

To ensure the precision of the annotation process, all videos underwent dual annotation sessions conducted by the same annotator. Following the methodology advocated by Chollet et al. (Chollet et al. [2019]), a span of one month was maintained between the two rounds of annotation. This temporal gap was strategically implemented to ensure that the annotator’s memory of the video content waned during the second annotation session. Subsequently, all annotations that exhibited disparities in terms of start points or annotation labels between the two rounds were extracted. These segments were then subjected to a third round of blind annotation, without any reference to the previous two annotations.

Upon completion of the first two annotation rounds, the concordance rate between them was computed. Remarkably, the overall self-agreement of the annotator reached an impressive 89.5% across all voice activity transitions characterized by the same onset point in the first two rounds. This self-consistency was maintained in terms of switch type, accomplishment, and interruption type in both annotation rounds. Specifically, the self-agreement rate for backchannel annotations was recorded at 93.5% for voice activity transitions marked as backchannels at least once in the first two rounds. A comparable self-agreement rate of 72.4% was observed for interruptions, while a substantially higher 95.3% self-agreement was achieved for smooth turn exchanges.

Following the third round of annotation, the global self-agreement rate of the annotator further elevated to 92.6%. A self-consistency rate of 84.07% emerged for interruption annotations, while smooth turn exchanges exhibited a self-agreement rate of 92%. Impressively, a remarkably high self-agreement rate of 98.8% was achieved for backchannel annotations, highlighting the reliability of the annotation process.

### 5.3. STATISTICAL RESULTS

Type	Agreement	Assistance	Clarification	Disagreement	Floor taking	Topic change	Tangentialization	Not identified
Count	38	4	5	2	42	2	1	76
Percentage	22.35%	2.35%	2.94%	1.18%	24.7%	1.18%	0.59%	44.71%

Table 5.3 Probability distribution of different interruption types for Failed interruption.

## 5.3 Statistical results

Eventually, a total of 3983 annotated records were amassed, representing voice activity changes in the French portion of the NoXi dataset. Among these records, 1403 instances corresponded to smooth turn exchanges, 1651 were classified as backchannels, and 929 were categorized as interruptions. This distribution indicates that voice activity changes were attributed to smooth turn exchanges 35% of the time, backchannels constituted 42%, and interruptions accounted for 23%.

Among the aggregate interruptions, there were 759 instances (81.7%) of successful interruptions and 170 instances (18.3%) of failed interruptions.

The distribution across the eight interruption types is outlined in Table 5.1. Within this spectrum, cooperative interruptions held the majority, totalling 505 occurrences (54.36% of all interruptions), while competitive interruptions numbered 348 (37.46% of all interruptions).

Upon further subdivision into eight sub-categories, interruptions of the *agreement* type emerged as the most frequent, predominantly within the cooperative interruption subset. Conversely, *floor taking* interruptions were the most prevalent in the category of competitive interruptions.

In combination with the *accomplishment* category (as demonstrated in Tables 5.2 and 5.3), *agreement* and *floor taking* remained the two dominant types in successful interruptions. However, for failed interruptions, a significant portion couldn't be classified, leading to the dominance of the *Not identified* category, which constituted 44.71% of failed interruptions.

An investigation into the duration of the first Inter-Pausal Unit (IPU) following a voice activity exchange revealed intriguing insights. Smooth turn exchanges exhibited the lengthiest first IPU, with an average duration of 4.31 seconds. In contrast, interruptions were accompanied by a shorter initial IPU, averaging at 2.91 seconds. Further dissection based on success revealed that successful interruptions boasted longer first IPUs (mean = 3.33 seconds) than failed ones (mean = 1.04 seconds). The classification of the interruption type also exerted an influence on the length of the first IPU. Notably, competitive interruptions yielded a lengthier first IPU (mean = 4.01 seconds) compared to cooperative interruptions (mean = 2.46 seconds). This discrepancy was particularly pronounced in the realm of successful competitive interruptions, where the mean first IPU duration was 1.89 seconds greater than that of successful cooperative interruptions (mean = 2.58 seconds).

We observe that overlaps occur frequently in interruptions (88%) and backchannels (71%), but rarely in smooth turn exchanges (29%), where the turn-taker tends to wait until the speaker completes the turn. Figure 5.4 illustrates the variation

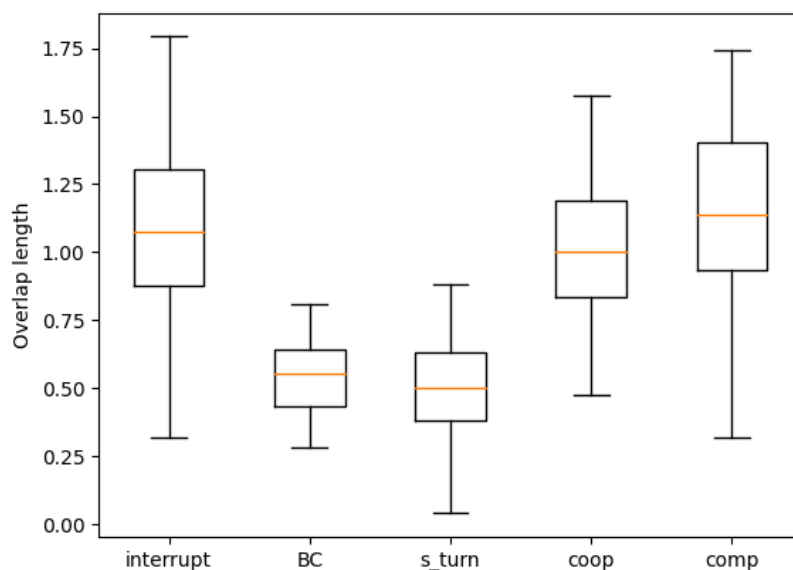


Figure 5.4 Overlap length in second for different types of exchanges. Labels: Interruption (interrupt), backchannel (BC), smooth turn exchange (s\_turn), cooperative interruption (coop), competitive interruption (comp).

in overlap duration across exchange types. Interruptions have the longest overlaps, with a mean of  $1.15s$ , indicating a high degree of competition for the floor. Smooth turn exchanges have much shorter overlaps than interruptions, with a mean of  $0.62s$ , suggesting a low level of conflict and a high level of coordination. Backchannels have the shortest overlap duration, as they are mostly single words or syllables that signal agreement or attention.

We also conduct an in-depth analysis of the temporal aspects of exchange initiation relative to the commencement of the speaker's last Inter-Pausal Unit (IPU), as depicted in Figure 5.5. Our findings reveal that smooth turn exchanges tend to occur earlier after the initiation of the IPU, with an average delay of 2.95 seconds, in contrast to interruptions and backchannels. This observation suggests that speakers provide cues for the conclusion of their turn. Interruptions and backchannels, on the other hand, exhibit a later timing, with average delays of 4.38 seconds and 4.93 seconds, respectively. This trend implies that either the speaker has no intention to relinquish the floor or the turn-taker requires more time to make a decision. Cooperative interruptions exhibit a slightly lengthier delay than backchannels, averaging at 4.99 seconds. This is noteworthy since cooperative interruptions are often utilized to express sustained interest or comprehension.

Furthermore, we delve into the impact of conversational roles (expert and novice) on interaction dynamics. Our analysis uncovers that novices contribute to approximately 29.4% of the conversation duration on average, while experts occupy approximately 69.8% of the conversation duration on average. This dis-

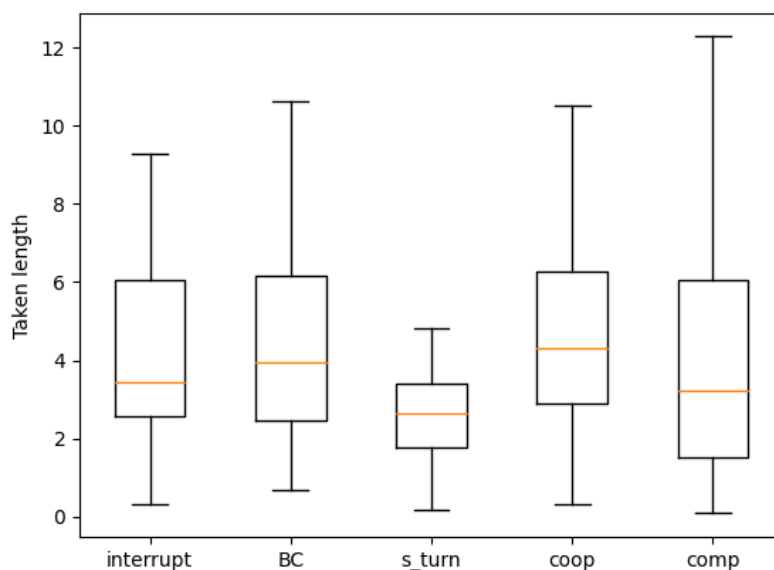


Figure 5.5 Relative distance in seconds between the exchange onset point and the start of the speaker’s last IPU for different types of exchanges. Labels: Interruption (interrupt), backchannel (BC), smooth turn exchange (s\_turn), cooperative interruption (coop), competitive interruption (comp)

crepancy mirrors the inherent asymmetry in knowledge and authority between these roles. Intriguingly, we also observe that a significant majority of interruptions (60.6%) and backchannels (76.5%) are initiated by novices. This insight might suggest that novices are more actively engaged in the discourse, potentially driven by the motivation to challenge assumptions or seek clarification from the expert.

In addition, we gauge the length of the first IPU following the exchange. Notably, the expert’s initial IPU is notably longer than that of the novice for both interruptions (3.98s vs 2.86s) and smooth turn exchanges (5.49s vs 3.65s). This discrepancy could signify that the expert possesses more substantial information to convey or exhibits greater confidence in holding the floor compared to the novice.

## 5.4 Multimodal Analysis

Conducting multimodal analysis, we want to see the difference between the three principal exchange types, and if there exists a pattern of nonverbal behaviour during the exchanges.

We observed extracted features from the French part of NoXi corpus, here we selected several major features for the analysis: pitch (f0) and loudness from the



## 5.4. MULTIMODAL ANALYSIS

---

speech signals, AU01(inner brow raiser), AU02(outer brow raiser), AU04(brow lower), AU06(cheek raiser) and AU12(lip corner puller) for facial expressions.

From the extracted AU features (notably AU01, AU02, AU04, and AU12) and other low-level signal (acoustic and visual) features.

Let us define the time steps illustrated in Figure 5.6:

- $t_1$ : start of the last speaker's IPU
- $t_2$ : end of the last speaker's IPU
- $t_3$ : start of the first exchange initiator's IPU
- $t_4$ : end of the first exchange initiator's IPU

We analyzed the features on the intervals  $[t_1, t_2]$ ,  $[t_3, t_2]$ ,  $[t_3, t_4]$  and  $[t_1, t_4]$ .

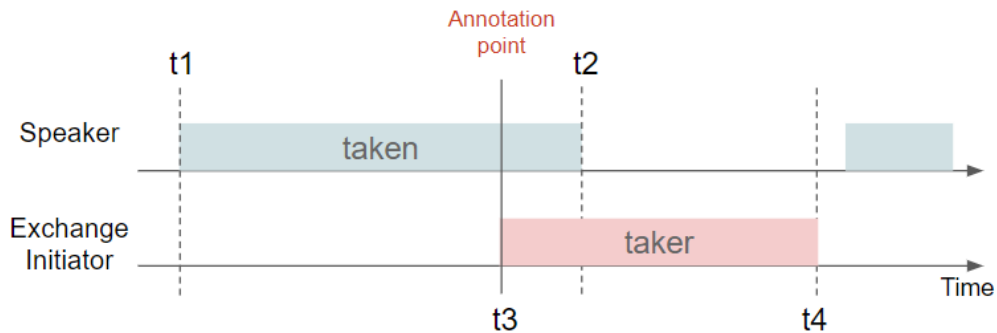


Figure 5.6 Explanation of data segmentation structure (taken: the last IPU of the speaker before the exchange, taker: the first IPU of the exchange initiator after the exchange onset point).

To facilitate a comparative analysis of the acoustic and visual characteristics between the speaker and the exchange initiator across different types of conversational exchanges, we initially calculate the average value of each feature within the designated time intervals. For the speaker, the time interval from  $t_1$  to  $t_2$  is utilized, encompassing their speech duration. For the exchange initiator, the time interval from  $t_3$  to  $t_4$  is employed, signifying the point of transition into taking over the turn. This approach allows us to scrutinize the disparities in prosody and facial expressions between the two roles and the potential variations based on the exchange type.

The distribution of average values for each feature is visually represented in Figure 5.7. To assess significant differences between the speaker and the exchange initiator, as well as across various exchange types, we employ a t-test ( $p < 0.05$ ). Our analysis reveals several noteworthy observations:

When initiating an interruption, the exchange initiator employs a higher pitch compared to the speaker, yet utilizes a lower loudness. This pattern suggests that they elevate their voice frequency as a signal of intent to interrupt, while deliberately moderating their vocal intensity to avoid coming across as overly aggressive.



## 5.4. MULTIMODAL ANALYSIS

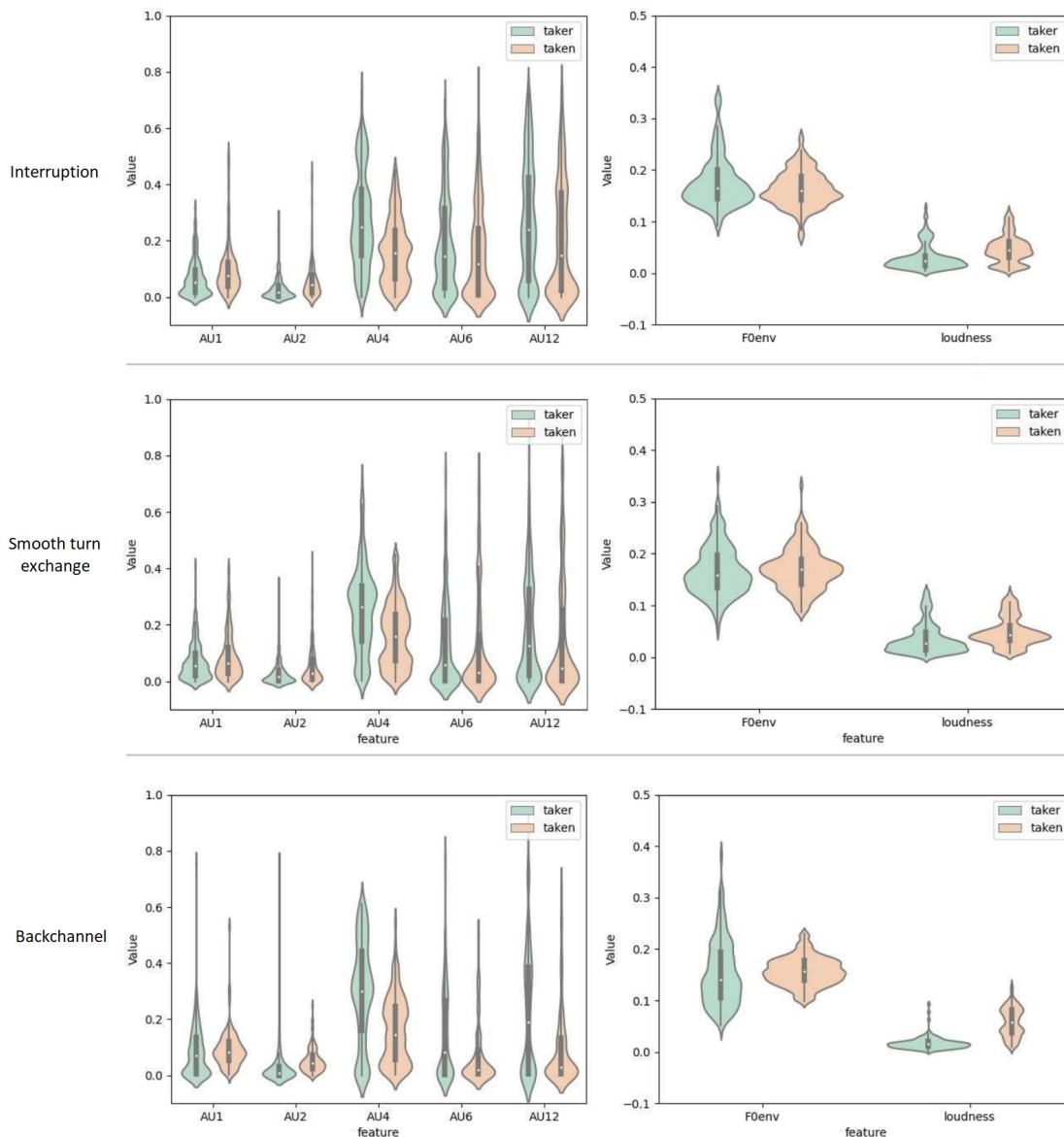


Figure 5.7 Average value of selected features during the corresponding intervals (taken:  $[t1, t2]$ , taker:  $[t3, t4]$ ).

Conversely, when initiating a smooth turn exchange or a backchannel, the exchange initiator employs a lower pitch and decreased loudness compared to the speaker. This shift in acoustic features indicates an intentional reduction in voice frequency and intensity, signalling agreement, acknowledgement, and a smooth transition to their speaking turn.

Additionally, the exchange initiator employs a higher pitch when initiating an interruption as opposed to initiating a smooth turn exchange or a backchannel. This implies a modulation of voice frequency in alignment with exchange type, using a higher pitch for more assertive exchanges and a lower pitch for supportive

ones. Similarly, the exchange initiator employs lower loudness when initiating a backchannel compared to initiating a smooth turn exchange or an interruption.

Turning to visual features, we analyze the facial action units (AUs) of both the speaker and the exchange initiator, reflecting movements of different facial muscles. We observe that the exchange initiator's AU01 (inner brow raiser) and AU02 (outer brow raiser) exhibit lower values than those of the speaker. Conversely, the exchange initiator's AU04 (brow lower), AU06 (cheek raiser), and AU12 (lip corner puller) exhibit higher values. Particularly, during interruption initiation, their AU06 and AU12 tend to be more active than during smooth turn exchanges or backchannels, indicating more frequent smiling when interrupting compared to smooth turn exchange or backchannel instances. We also note that the speakers' AU06 and AU12 exhibit reduced activity during the backchannel segment compared to the other two types.

## 5.5 Conclusion

In this chapter, we introduced a new interruption annotation schema we used to annotate the NoXi corpus. We conducted statistical analysis on the occurrence of the different interruption types. We also studied the length of the first IPU after the interruption and found some significant differences.

From our analysis, in the French part of the NoXi database, interruptions occur frequently in conversations. Most interruptions are successfully completed and are *cooperative* interruptions. Failed interruptions are often very short, which does not allow us to determine their type and justifies the introduction of the type *Not identified*. Agreement interruptions take up the most part of cooperative interruptions while floor taking are predominant for competitive ones.

NoXi gathers interaction of an expert giving information to a novice on a topic that interests both of them. This particular context of interaction explains why there are more cooperative interruptions than competitive ones.

We analyzed the multimodal signals that humans use to initiate and respond to exchanges such as smooth turn exchanges, interruptions, and backchannels.

Interruptions have longer overlaps than smooth turns and backchannels. We also noticed that the exchange initiator adjusts the voice pitch depending on the type of exchange, using a higher pitch for interruptions and a lower pitch for smooth turn exchanges and backchannels.

Examining the facial expressions that humans use during and after the exchanges we found that some of them are more active than others, depending on the types of exchanges. These patterns indicate that the visual features of the interlocutors are influenced by the type of the exchange.

### The key points of this Chapter:

*This Chapter addresses research question Q1.*

- How do humans perceive and categorize different interruption situations during their interactions?

#### *Annotation*

- We propose a new annotation schema for manual exchange/switch annotation in the conversation.
- The annotation results of *NoXi Corpus* indicate that interruptions are quite common during interaction and most of the interruptions successfully took the speaking floor.

#### *Analysis*

- Interruptions have longer overlaps than smooth turns and backchannels.
- Smooth turn exchanges are relatively closer to the beginning of the last speaker IPU than interruptions and backchannels.
- We analyzed the multimodal signals that humans use to initiate and respond to exchanges, such as smooth turns, interruptions, and backchannels.
- Different types of exchanges do show different patterns, but there are also common points in strategies.

#### *Publications related to Chapter:*

- Liu Yang, Catherine Achard, Catherine Pelachaud. Annotating Interruption in Dyadic Human Interaction. Thirteenth Language Resources and Evaluation Conference, LREC, 2022.
- Liu Yang, Catherine Achard, Catherine Pelachaud. Multimodal Analysis of Interruptions. International Conference on Human-Computer Interaction. HCII, 2022.
- Jiyeon WOO, Liu Yang, Catherine Achard, Catherine Pelachaud. Are we in sync during turn switch? 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE (SIVA workshop), 2023.

# MIC: Multimodal interruption classification in humans' interaction

## Contents

---

6.1	Related works . . . . .	66
6.2	Features . . . . .	67
6.3	The proposed model . . . . .	71
6.4	Result & discussion . . . . .	71
6.4.1	Results . . . . .	72
6.4.2	Comparative study . . . . .	73
6.4.3	Interruption Window length . . . . .	75
6.4.4	Adaptation of the temporal length during real-time application . . . . .	76
6.5	Conclusion . . . . .	77

---

Annotating and analysing the interruption during human-human interaction, we found different types of interruption, mainly divided into two categories: cooperative and competitive. Human uses different kinds of interruption to complete different goals, either to be supportive or to grab attention. For a Socially Interactive Agent to be capable of handling user interruptions in dyadic interaction, it should be able to detect interruption, recognize its type (cooperative/competitive), and then plan its behaviour to respond appropriately. This Chapter corresponds to the research questions of Q2 and Q3.

As a first step towards this goal, we developed a multimodal classification model using acoustic features, facial expression, head movement, and gaze direction from both, the interrupter and the interruptee. The classification model learns from the sequential information to automatically identify interruption types. We

also present studies we conducted to measure the shortest delay needed for our classification model to identify interruption types with high classification accuracy.

### 6.1 Related works

Previously, Goldberg and colleagues (Goldberg [1990]) discovered that the location of an interruption could offer insights into differentiating interruption types. In addition to the starting point, the duration of overlaps assumes a crucial role in distinguishing between cooperative and competitive interruptions. The findings in (Kurtic et al. [2010], Jefferson [2004]), as well as the analysis presented in Chapter 5, suggest that competitive overlaps tend to be longer than cooperative ones.

Moreover, a growing body of research has accumulated evidence indicating that prosodic features exhibit variations between cooperative and competitive interruptions. Yang *et al.* (Yang [2001]) argued that competitive interruptions manifest higher pitch and intensity compared to their cooperative counterparts. Shriberg (Shriberg et al. [2001a]) and Hammarberg *et al.* (Hammarberg et al. [1980]) discovered that individuals elevate their voice energy and pitch when attempting to interrupt the ongoing speaker. Schegloff *et al.* (Schegloff [2000]) contended that speakers employ prosodic variations and repetition to signify a strong desire to claim the conversational turn.

Numerous studies have proposed models for the automatic classification of interruption types using multimodal features. Lee *et al.* (Lee et al. [2008]) analyzed hand motion activity, speech intensity, and disfluency for classifying competitive/cooperative interruptions in spoken dyadic conversations. Hand motion activity and speech intensity were identified as reliable features for distinguishing interruption types. However, using a single modality resulted in significantly lower classification performance, with an accuracy of 71.2%.

Truong *et al.* (Truong [2013]) developed an SVM model to classify overlaps. They used low-level signals such as acoustic features ( $f_0$ , intensity, and voice quality), alongside high-level annotations like gaze direction and head movement communicative functions. With a sequence length of 0.6 seconds after the overlap onset point, the SVM model achieved good performance with an Equal Error Rate (EER) of 32.1%. They also noted that incorporating gaze information slightly improved accuracy, whereas adding acoustic information from the interruptee did not.

Chowdhury et al. (Chowdhury et al. [2015]) introduced a Sequential Minimal Optimization model for competitive/cooperative overlap classification, using features like prosody, voice quality, MFCC, energy, and spectral features. With an optimal subset of selected features, the model achieved an F1-score of 0.69. Subsequently, Chowdhury et al. (Chowdhury et al. [2019], Chowdhury and Riccardi [2017]) employed both acoustic (prosodic, spectral, voice quality, MFCC, and energy) and lexical features to classify cooperative/competitive overlaps. Among

various models tested, the best performance (F1-score of 0.70) was attained with a feed-forward neural network (FFNN) using both acoustic and lexical features.

Incorporating emotional dimensions (*control* and *valence*), Egorow *et al.* (Egorow and Wendemuth [2019]) employed acoustic features to classify overlaps in a telephone-based human conversation corpus. The SVM model achieved the best performance with an F1-score of 0.74.

While prior research has demonstrated that features such as hand activity, body motion, gaze, and head gestures enhance classification accuracy, acoustic features have emerged as critical elements for classifying interruptions and overlaps. Nonetheless, facial expressions have been overlooked in previous studies. Based on the experimental results from (Motley [1993]), conversational facial expressions are linked with the dialogue context and operate as nonverbal interjections, potentially offering supplementary information for interruption classification models.

Furthermore, the significance of sequential information during interruptions and overlaps has been neglected in prior works, which mainly rely on statistical projections as input for classification models. Yet, the analysis findings from (Truong [2013]) illustrate the temporal curve differences in acoustic features, which were utilized to classify overlap types.

Hence, we propose a novel method to classify interruptions by encompassing acoustic profiles, head activity, gaze behaviour, and facial expressions. Our approach also accounts for sequential information. Given our objective of implementing an online model that swiftly classifies interruption types during human-agent interaction, we investigate how classification performance varies with the length of the time window after the interruption point.

## 6.2 Features

In this study, we utilize two distinct corpora: the AMI corpus (Carletta [2007]) and the NoXi corpus (Cafaro *et al.* [2017]). Following the annotation schema expounded in (YANG *et al.* [2022]), as elucidated in Chapter 4, a total of 508 interruptions were annotated in the AMI corpus, comprising 230 cooperative interruptions and 278 competitive interruptions.

Our novel approach for classifying interruptions harnesses multimodal signals. To gauge the minimum delay requisite for accurate classification, we initially define a time window centred around the onset point of each interruption, subsequently extracting multimodal features within this delineated timeframe.

Our feature set encompasses acoustic attributes as well as visual cues such as facial expressions (eyebrow movement), head motion activity, and gaze direction. Each modality encompasses two distinct types of features: *local* features denote values extracted from each frame within the chosen time window, while *global* features represent statistical summaries of each *local* feature over the predefined time span.

## 6.2. FEATURES

---

Additionally, for the AMI corpus, we incorporate two supplementary features which are provided with the database: the annotated communicative functions of head movements (concord, discord, deixis, emphasis, negative, turn, and other) and the presence of mutual gaze (when the interlocutors look at each other). These two annotations are not provided in the NoXi corpus.

In order to establish the optimal interruption window, we meticulously analyze the duration of interruption overlaps within the annotated interruptions from both corpora. The outcome of our statistical analysis is depicted in Figure 6.1. Notably, the average duration of cooperative interruptions amounts to 1.11 seconds, whereas competitive interruptions exhibit an average duration of 1.49 seconds.

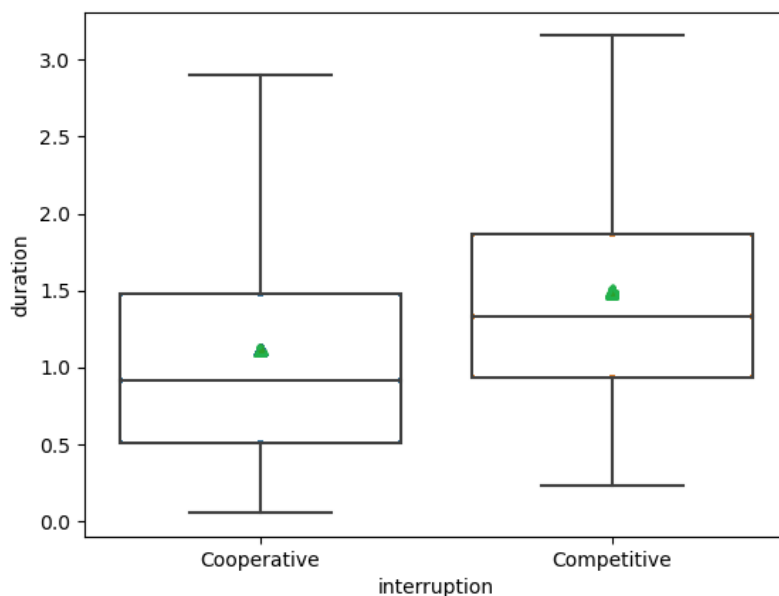


Figure 6.1 Interruption overlap duration in second (AMI corpus).

For each instance of interruption, as illustrated in Figure 6.1, it's noticeable that the majority of interruption overlaps tend to last longer than 0.6 seconds. We designate the initiation point of the overlap as  $t_0$ . Given the objective of classifying the interruption promptly and responding accordingly before the overlap concludes, it's imperative to select a window size that's less than 0.6 seconds. In this context, Truong and colleagues (Truong [2013]) recommend an interruption window length of 0.6 seconds after the interruption onset point. We opt to adopt this same window size for our study and focus on the temporal segment spanning from  $t_0 - 0.6$  seconds to  $t_0 + 0.6$  seconds, which we refer to as "interruption windows" hereafter. This choice aligns with our primary goal, which is to equip a virtual agent with the capability to promptly manage interruptions and react to them in a timely manner, preferably before the overlap concludes. Notably, during the initial phase leading up to the interruption at  $t_0$ , spanning from  $t_0 - 0.6$  seconds to  $t_0$ , only the individual being interrupted (interruptee) is speaking. In contrast,



## 6.2. FEATURES

the subsequent phase encompasses both interlocutors speaking, at least to some extent.

The classification of interruptions hinges on the multimodal features extracted from both the interrupter and the interruptee across the interruption window. Depending on the specific methodology employed for the classification model (as detailed in Section 6.4.2), we either compute *global* features across the entirety of the interruption window or extract *local* features at each time step within the window (25 *fps* for NoXi, 15 *fps* for AMI as presented in Chapter 4).

Since the NoXi corpus and the AMI corpus have different frame rates, for ease of comprehension, we will use the AMI corpus as an example to explain the features used. We applied the same processing to the NoXi corpus, with the only difference being that the features for the NoXi corpus correspond to a frame rate of 25 *fps* instead of 15 *fps*.

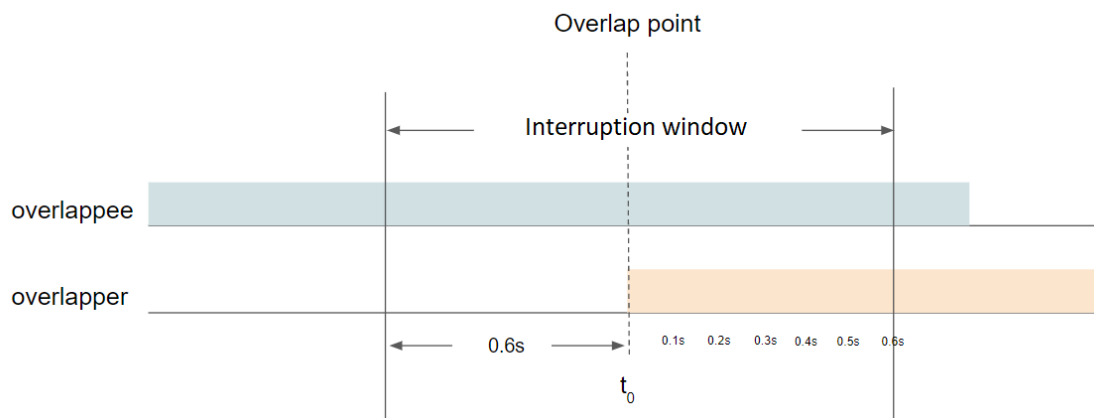


Figure 6.2 Segmentation of features.

**Local acoustic features:** The set of 33 acoustic features we have taken into consideration comprises various parameters such as pitch (Fundamental frequency  $f_0$ ,  $f_0$ -envelope), loudness, voice-probability, jitter, shimmer, logarithmic harmonics-to-noise ratio (logHNR), Mel-frequency cepstral coefficients (MFCC 0-12), Logarithmic signal energy from pcm frames, Energy in spectral bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), roll-off points (25%, 50%, 70%, 90%), centroid, flux, max-position, and min-position, as proposed by (Chowdhury et al. [2019]).

These acoustic features are extracted at a frequency of 100 frames per second and then resampled to 15 *fps* to match the frequency of the visual features (eyebrow movement, head motion activity, and gaze direction). This resampling is carried out to ensure synchronization between the acoustic and visual features. Consequently, we generate a series of temporal values for each acoustic feature for both the interrupter and the overlapped, subsequently storing these values in a vector. This vector is referred to as the "local acoustic features vector" pertaining to a given time window around an interruption event. The dimensions of the local acoustic features vector denoted as  $A_l$ , are determined by the product of 33 features per interlocutor, multiplied by 18 frames, resulting in a size of  $(33 * 2) * 18$ .

**Global acoustic features:** For each feature mentioned previously, we extract their statistical projection over the interruption windows as proposed by (Chowdhury et al. [2019]). They are composed of values such as min and max position, range, linear and quadratic regression coefficients and approximation errors, variance, standard deviation, skewness, peaks, mean peak distance and mean peak.

These statistical projections are normalized by z-scores, using the mean and standard deviation of all windows. They composed the global acoustic features vector  $A_g$  of size 792 ( $33*2*12$ , the value of the above 12 statistical projections for the 33 features for each interlocutor).

**Local eyebrow features:** In this research, we focus on the three action units AU01, AU02 and AU04 representing eyebrow movements. Eyebrow features are extracted with a frequency of 15 frames per second, composing a feature vector  $E_l$  of size  $(3 * 2) * 18$  (3 AUs for both interlocutors over 1.2 sec (18 frames)).

**Global eyebrow features:** Statistical projections of each eyebrow feature are estimated as done for the acoustic features. They composed a vector  $E_g$  of size 48 (The value of the 8 statistical projections: min and max position, range, linear and quadratic regression coefficients and approximation errors, variance and standard deviation, for the 3 features for each interlocutor).

**Local head activity features:** For head movement features, we use the head position (in x-y-z axis) at the frequency of 15 frames per second for both corpora. For the AMI corpus, we also use the head functions annotations: *concord*, *discord*, *deixis*, *emphasis*, *negative*, *turn*, and all *other* communicative head gestures. The final vector  $H_l$  is of size  $(10*2)*18$  (head activity + head function one-hot encoding for both interlocutors over 1.2 sec (18 frames)).

**Global head movement features:** The statistical projections of head activity are calculated on the interruption window. For the head function annotation of AMI corpus mentioned above, we use a one-hot encoding by setting the value 1 if the event is present at least at the one-time step of the interruption window. The Global head movement vector is of size 34 (the value of the 8 statistical projections: min and max position, range, linear and quadratic regression coefficients and approximation errors, variance, standard deviation, for 1 feature + the vector's one-hot encoding for each interlocutor).

**Local gaze features:** We also consider the gaze direction (in x-y axes). For AMI corpus, we also use the *focus of attention* annotations of which the binary value is computed to indicate the presence of mutual attention (when the interrupter and interruptee are looking at each other) or its absence (when the interrupter or interruptee is looking somewhere else). The final vector  $G_l$  is of size  $(4 * 2) * 18$  (4 gaze features for both interlocutors over 1.2 sec (18 frames)).

**Global gaze features:** Statistical projections of gaze direction are estimated as done for acoustic features. For the *focus of attention* annotation we use a binary value for mutual attention as presented in the local gaze features. The Global gaze vector is of size 34 (the value of the 8 statistical projections: min and max position, range, linear and quadratic regression coefficients and approximation errors, variance, standard deviation, for 2 gaze direction features + the vector one-hot encoding for each interlocutor).

## 6.3 The proposed model

In contrast to prior methodologies (Chowdhury et al. [2019], Chowdhury and Riccardi [2017], Truong [2013]), our approach deviates by refraining from manually extracting global features. Instead, we delegate this task to a neural network, this shift not only enhances the model’s adaptability but also aligns seamlessly with our primary objective of real-time application, as demonstrated in the subsequent evaluation section.

Diverging from conventional feed-forward neural networks, the long short-term memory (LSTM) architecture boasts connections that facilitate the processing of not only singular data points but also entire sequences of data. Each LSTM unit comprises pivotal elements, including a memory cell, an input gate, an output gate, and a forget gate. Collectively, these gates regulate the inflow and outflow of information within the memory cell.

The LSTM architecture we have adopted is visually illustrated in Figure 6.3. At each time step  $t$ , the input  $x_t$  is formed through the amalgamation of all local features. We extract the hidden state from the final time step, which is subsequently channelled into a dense layer responsible for generating the ultimate classification. Despite its apparent simplicity, this model has yielded the most promising outcomes. Interestingly, introducing more intricate architectures (such as stacked LSTM) has demonstrated a propensity for significant overfitting, thus undermining overall performance.

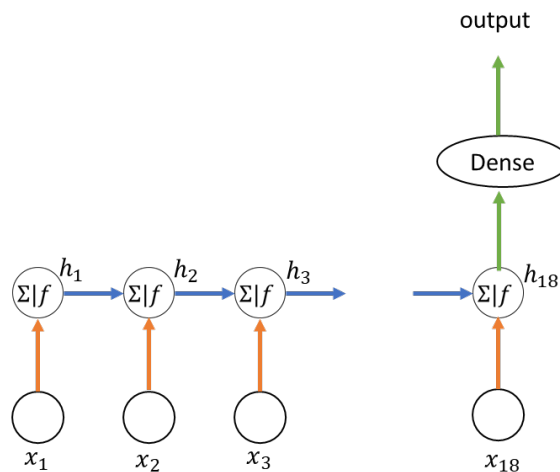


Figure 6.3 The long-short-term memory (LSTM) architecture.

## 6.4 Result & discussion

In this section, we present the experimental classification results we obtained using different modalities and compare them to the state of the art. Then we also introduce the study we conducted on the time window length to be considered.

## 6.4. RESULT & DISCUSSION

	Audio		Facial, head, gaze		All modalities	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN with modalities of Chowdhury et al. [2019]	0.74	0.73	-	-	-	-
FFNN with our modalities	0.74	0.73	0.69	0.67	0.79	0.78
SVM with modalities of Truong [2013]	0.69	0.66	0.65	0.61	0.72	0.72
SVM with our modalities	0.72	0.72	0.68	0.67	0.77	0.76
LSTM with our modalities	0.75	0.73	0.69	0.68	<b>0.81</b>	<b>0.80</b>

Table 6.1 Accuracy and F1 measure for FFNN, SVM and LSTM model with different combinations of modalities for the AMI corpus.

	All modalities		Audio, Facial, Head		Audio, Facial, Gaze		Audio, Head, Gaze	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN	0.79	0.78	0.75	0.75	0.77	0.77	0.77	0.76
SVM	0.77	0.76	0.74	0.73	0.75	0.74	0.75	0.75
LSTM	0.81	0.80	0.76	0.76	0.79	0.78	0.79	0.79

Table 6.2 Ablation study of FFNN, SVM and LSTM model for the AMI corpus with our modalities.

### 6.4.1 Results

Training Procedure: The methodology was implemented and evaluated using TensorFlow. The dataset was divided into training (70%), validation (10%), and test (20%) subsets. The input to the LSTM layer, with a dropout rate of 0.2, is constructed by concatenating all the aforementioned local features over the 1.2-second interruption window (18 frames), resulting in a vector of size  $100 \times 18$ . The latent vector, with a dimension of 10, is then passed through a dense layer with a dimension of 8. The final output layer, with a dimension of 1, employs a sigmoid activation function.

Throughout our experiments, the model was trained using mini-batches consisting of 64 interruption instances. An Adam optimizer with a fixed learning rate of  $1e-5$  was employed for the training process.

The results are presented in Table 6.1 for the AMI corpus. Analogous to previous studies, acoustic features alone demonstrate commendable performance. However, the inclusion of facial expressions, head movement, and gaze substantially enhances performance, achieving an accuracy of 81% and an F1-score of 0.80.

A comparison among facial expressions, gaze, and head activity reveals the predominant role of acoustic features in classification. Further insights are gleaned from ablation studies as showcased in Table 6.2. These experiments focus on evaluating the impact of each modality on classification accuracy by systematically excluding features from individual modalities.

While ablation experiments were also conducted for each specific feature, the isolated influence of a single feature was found to be relatively minor, and therefore, those results are not detailed here. Interestingly, among the three models, the accuracy experiences a significant decline when gaze-related features are removed, surpassing the reduction observed when the other two modalities are omitted.

We compare our result with random accuracy, which is calculated with the equation (Tharwat [2020]):

$$\begin{aligned} \text{accuracy} &= P(\text{class} = 0) * P(\text{prediction} = 0) + P(\text{class} = 1) * P(\text{prediction} = 1) \\ &= (348/(505 + 348))^2 + (505/(505 + 348))^2 \\ &= 0.5162 \approx 0.52 \end{aligned}$$

Where, in our case, class 0 represents competitive interruption and 1 represents cooperative, therefore  $P(\text{class} = 0) = P(\text{prediction} = 0) = 348/(505+348) = 0.41$ , and  $P(\text{class} = 1) = P(\text{prediction} = 1) = 505/(505 + 348) = 0.59$ .

The outcomes of our models are displayed in Table 6.3 for the NoXi Corpus. The most favourable results (accuracy = 0.69) are also achieved through the utilization of multimodal features, although they only exhibit a marginal improvement over random accuracy (0.52). Upon a comparison between facial expression, gaze, head activity, and acoustic features, it becomes evident that acoustic characteristics play a pivotal role in the classification process. In a manner similar to our AMI corpus study, we conducted ablation studies detailed in Table 6.4. In these experiments, each modality was isolated, and its related features were systematically omitted. However, unlike the AMI corpus, no significant differences were observed among the three modalities.

An important consideration in interpreting these results lies in the differences between the AMI and NoXi databases. Apart from the language variation (French for NoXi and English for AMI), the conversational contexts differ significantly. The NoXi corpus comprises dyadic screen-mediated conversations, while the AMI corpus consists of four-party face-to-face interactions. Even without considering multi-party interruptions, dyadic interruptions are more prevalent in AMI compared to NoXi. The AMI context involves four interlocutors engaged in brainstorming for product design, which could lead to more competitive interruptions arising from attempts to assert individual ideas. The higher number of participants in the AMI conversations could also contribute to the increased occurrence and frequency of interruptions.

On the other hand, the NoXi database involves participants adopting roles as experts or novices discussing various topics. Experts tend to dominate the conversation, and most interruptions arise from novices seeking information or expressing opinions, typically less competitive in nature. These differing scenarios give rise to interactions with varying degrees of dynamism and interruptions characterized by varying levels of competitiveness.

It is plausible that the discrepancies in settings and scenarios contribute to the divergent results between the two corpora. Furthermore, the absence of certain high-level annotation features in the NoXi corpus might also contribute to these variations.

## 6.4.2 Comparative study

We conducted a comprehensive comparison of our method with previous approaches that have been developed for classifying interruptions in conversations.

## 6.4. RESULT & DISCUSSION

	Audio		Facial, head, gaze		All modalities	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN with modalities of Chowdhury et al. [2019]	0.63	0.61	-	-	-	-
FFNN with our modalities	0.63	0.61	0.61	0.60	0.66	0.64
SVM with modalities of Truong [2013]	0.57	0.52	-	-	-	-
SVM with our modalities	0.61	0.57	0.59	0.57	0.62	0.60
LSTM with our modalities	0.65	0.62	0.62	0.60	<b>0.69</b>	<b>0.65</b>

Table 6.3 Accuracy and F1 measure for FFNN, SVM and LSTM model with different combinations of modalities for the NoXi corpus.

	All modalities		Audio, Facial, Head		Audio, Facial, Gaze		Audio, Head, Gaze	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
FFNN	0.66	0.64	0.65	0.65	0.65	0.64	0.64	0.63
SVM	0.62	0.60	0.61	0.60	0.60	0.58	0.61	0.61
LSTM	0.69	0.65	0.68	0.66	0.66	0.65	0.68	0.67

Table 6.4 Ablation study of FFNN, SVM and LSTM model for the NoXi corpus with our modalities.

Specifically, we evaluated our approach against Truong’s SVM model and Chowdhury’s FFNN model, which were designed to classify interruptions using different sets of features. Here’s an in-depth exploration of our comparison:

Truong’s SVM model was trained using both the original features presented in (Truong [2013]) and the features extracted by our method. Chowdhury’s FFNN model, on the other hand, was trained using the acoustic features described in (Chowdhury et al. [2019]), whereas we employed our global multimodal features, which were automatically generated from the local features, as detailed in Section 6.2. It’s worth noting that both Truong’s and Chowdhury’s methods rely on global feature vectors that are manually crafted, whereas our approach leverages a neural network to learn and extract the most pertinent features.

Our comparative analysis, presented in Table 6.1 for the AMI database and Table 6.3 for the NoXi database, demonstrates that, across all models, acoustic features are more informative for the classification task when compared to facial expressions, head activities, and gaze direction. In fact, the achieved accuracy using only facial expression, head activity, and gaze direction is relatively modest for all models.

Interestingly, the multimodal features we introduced have the potential to enhance classification performance for all models. The improvements are notable, with a 5% boost for the FFNN model, a 6% enhancement for the LSTM model, and a 3% increase for the SVM model. This indicates that facial expressions and head activities indeed convey valuable information for interruption classification.

Ultimately, our proposed multimodal LSTM classification model outperforms the other models, underlining the significance of sequential information gleaned from time series data. The model’s ability to extract relevant features from this temporal information enables real-time recognition of interruptions, as highlighted in Section 6.4.4. An important aspect is that our model doesn’t require information

about the interruption's endpoint, making it well-suited for real-time application scenarios.

### 6.4.3 Interruption Window length

To thoroughly examine the impact of interruption window length on classification accuracy, we adopt a systematic approach. Our aim is to rapidly identify the interruption type in order to apply it in real-time scenarios. To achieve this, we keep the window length fixed prior to the interruption's onset point at 0.6 seconds, and then systematically vary the window length after the onset point in intervals of 0.2 seconds, ranging from 0 to 1 second.

The motivation behind this study is to strike a balance between accuracy and response time, considering that an ideal virtual agent should determine the interruption type as accurately as possible while responding promptly. The results, depicted in Figure 6.4, indicate that the accuracy increases as the interruption window length expands when employing our LSTM model. However, it's worth noting that certain cooperative interruptions have shorter overlap durations, as illustrated in Figure 6.1. Opting for a longer interruption window could delay the virtual agent's response.

To address this trade-off between classification accuracy and response time, a prudent choice seems to be a fixed interruption window length of 0.6 seconds following the onset point of interruption. This choice optimally balances the accuracy of classification and the agent's prompt response time.

Another idea, possibly using temporal series as proposed with the LSTM model, is to adapt this temporal length during real-time application as proposed just below.

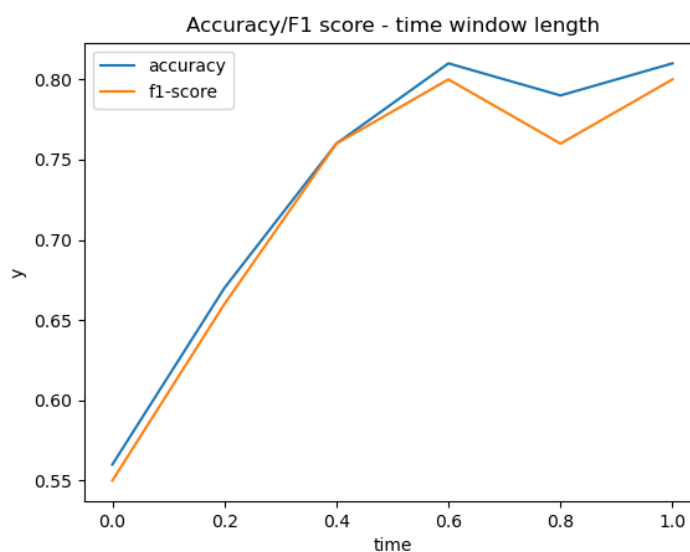


Figure 6.4 Accuracy & Macro F1-score with different interruption window lengths.



### 6.4.4 Adaptation of the temporal length during real-time application

In the context of real-time human-agent interaction, the swift classification of interruptions is crucial. As soon as an interruption is detected, marked by the onset of overlap, we initiate the inference process using multimodal signals within the temporal window  $t_0 - 0.6s, t_0s$ . By applying the LSTM model to this interval, we can initialize the latent vector within the architecture, as illustrated in Figure 6.3. A significant advantage of LSTM is its ability to update this latent vector frame by frame, starting from  $t_0$ , and predict cooperative or competitive interruption probabilities at each time step. Although the result from the sigmoid layer is considered as probabilities, this part of the model allows us to set a threshold for classification, determining whether to execute the classification at each frame.

To experimentally validate this concept, we systematically study the feasibility of classifying interruptions with AMI corpus. We begin classification at  $t_0$  and incrementally add incoming data, frame by frame until reaching 1.2 seconds (18 frames) post  $t_0$ . Denoting  $y$  as the output of the sigmoid at the studied frame, we perform classification at the frame if  $\max(y, 1 - y)$  surpasses the set threshold. However, if this condition is not met, we proceed to the next frame. Naturally, classification is executed regardless of the threshold if the overlap concludes or when the maximum interruption window length of 1.2 seconds post  $t_0$  is attained. Utilizing a lower threshold results in a shorter reaction time, yet possibly at the expense of accuracy.

Figure 6.5 illustrates the percentage of interruptions classified against the number of frames after  $t_0$  for various threshold values. As anticipated, a threshold of 0.5 classifies all interruptions at  $t_0$ . Elevating the threshold extends the classification time, leading to a longer reaction time. Notably, overly high thresholds might hinder classification, as achieving such a classification score becomes challenging. For instance, only around 50% of interruptions are classified after  $t_0 + 14$  frames (roughly 1 second) when the threshold is set to 0.9. Conversely, with a threshold of 0.6, over 95% of interruptions are classified within 0.4 seconds (6 frames), while the remaining 5% necessitate additional time. Irrespective of the chosen threshold, the LSTM framework empowers us to adapt the reaction time according to the complexity of classification. This flexibility ensures the virtual agent's responsiveness while maintaining accuracy. However, proceeding quickly with poor classification results is not acceptable as well. We thus present in Table 6.5 the mean accuracy and the mean reaction time according to the threshold values. As expected, the higher the threshold, the better the accuracy and the longer the response time. The model is requested to classify the interruption as soon as possible with acceptable accuracy. We find from Figure 6.5 and Table 6.5 that a threshold of 0.8 seems to be a good choice.

## 6.5. CONCLUSION

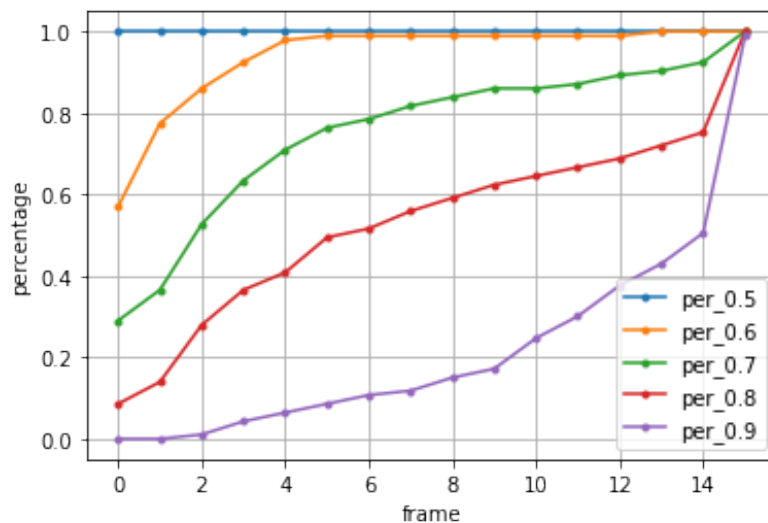


Figure 6.5 Percentage of classified interruptions according to the number of frames after the beginning of the overlap, for different thresholds.

Threshold	Mean accuracy	Mean reaction time length (frame and second)
0.5	0.66	0 frame (0s)
0.6	0.72	1 frame (0.07s)
0.7	0.76	4 frames (0.27s)
0.8	0.83	8 frames (0.53s)
0.9	0.87	12 frames (0.8s)

Table 6.5 Mean accuracy & mean reaction time with different thresholds

## 6.5 Conclusion

In this chapter, we proposed to classify cooperative and competitive interruptions in conversation with an LSTM model and evaluated different existing models. We experimented with different combinations of modalities and achieved our best performance using the acoustic profiles, facial expression, head movement and gaze features from both interrupter and interruptee. The experiments indicate that our LSTM model is able to learn accurate information from sequential series and improves classification performance. Tested with different interruption window lengths, the designed LSTM model is able to classify the interruption 0.6s after its start point with an accuracy of 81%.

Another advantage of LSTM is to make the classification more or less faster, according to the difficulty of the interruption. Using a classification threshold fixed to 0.8 allows us to classify interruption within 0.53s on average, with an accuracy of 83%. 40% of interruptions will be classified in less than 0.26s while the 30% of the most difficult examples will take more than 0.8s to be classified.

### The key points of this Chapter:

*This Chapter addresses research question Q2 and Q3.*

- How can we identify different types of interruption during the interaction? What's the most effective modality combination?
- How long would it take to get enough usable information for the interruption classification? Is it possible for ECAs to identify the interruption types in real time?

*MIC: Multimodal interruption classification.*

- Different from previous studies, we take the sequential information of multimodal features into account using an LSTM model.
- We train and evaluate the proposed model on AMI and NoXi corpus, proposing different features and annotations.
- In order to adapt our model in the real-time application, we manually settle a threshold on the classification score.
- Using a classification threshold fixed to 0.8 allows us to classify interruption within 0.53s on average, with an accuracy of 83%.

*Publications related to Chapter:*

- Liu Yang, Catherine Achard, Catherine Pelachaud. Multimodal classification of interruptions in human interaction. Proceedings of the 2022 International Conference on Multimodal Interaction. ICMI, 2022.

# One-PredIT: Prediction of Interruption Timing in dyadic interaction using one-class classification model

## Contents

---

7.1	Related works . . . . .	80
7.2	Approach . . . . .	81
7.3	Comparative study . . . . .	82
7.4	Subjective Evaluation . . . . .	82
7.4.1	Stimuli . . . . .	83
7.4.2	Comparison variables . . . . .	86
7.4.3	Questions . . . . .	87
7.4.4	Comparison Results . . . . .	87
7.5	Discussion . . . . .	92
7.6	Conclusion . . . . .	93

---

We aim for Embodied Conversational Agents (ECAs) to possess the capability to interrupt human users when necessary during human-agent interactions. To ensure that the interruptions made by ECAs are reasonable and not perceived as system errors, we need to consider two significant aspects: when to interrupt and how to interrupt. In this chapter, we will first address the first question, "when to interrupt." Choosing an appropriate point in time for interruptions is of paramount importance. This chapter corresponds to the research questions posed in Q4, Q5, and Q6.

We propose a novel approach to find possible interruption initiation timing in dyadic interactions using multi-modal features only from the speaker since this

model is to be applied to a virtual agent, of which the behaviour may be different from the real human. Our approach is based on a one-class classification model that has been trained on a corpus of dyadic interactions. We evaluate the model's accuracy through a perceptual study that compares model-predicted interruptions with ground truth data and random interruption timings.

We also evaluate the influence of interruption types (cooperative and competitive) on the perception of interruptions. Interesting results arise from this study, specifically on the timing where an interruption can be produced by the virtual agent. This important result may help future researchers equip agents with the ability to interrupt their human interlocutor.

## 7.1 Related works

Interruptions are common, but in most cases speaking turn exchanges smoothly during a conversation, smooth turn exchange is found predictable due to various cues that indicate the end of a turn: Ruth E. Corps et al. (Corps et al. [2019]) proposed a model that predicts turn-ends by using the semantic content and timing of the preceding speech. Sacks et al. (Sacks et al. [1978]) and Levinson et al. (Levinson and Torreira [2015]) provided insights into the systematic organization of turn-taking and its implications for processing models of language. Additionally, Garrod and Pickering (Garrod and Pickering [2015]) have investigated the use of content and timing to predict turn transitions. Raux and Eskenazi (Raux and Eskenazi [2009]) proposed a finite-state turn-taking model for spoken dialogue systems, which takes into account factors such as speech rate and gaze behaviour to predict when a speaker will finish their turn. Skantze (Skantze [2021]) proposed a continuous model of turn-taking using LSTM recurrent neural networks, which takes into account contextual information.

To manage turn-taking during human-agent interaction, there are also researches focusing on interaction strategies for affective conversational agents, which can respond appropriately to user interruptions and manage the flow of conversation. Crook et al. (Crook et al. [2010], Smith et al. [2011]) developed a model for handling user's interruptions in an embodied conversational agent, which takes into account the user's intent and the system goals. They proposed a set of interaction strategies for an affective conversational agent, including strategies for handling user interruptions and recovering from communication breakdowns. Chylek et al. (Chylek et al. [2018]) proposed to use low-level acoustic features to predict interruptions and overlaps with a deep residual learning network. Their method allows for predicting interruption timings using the speaker's acoustic features.

Current studies highlight the importance of effective turn-taking and interruption management in human-agent interactions and focus more on handling the interruptions initiated by the human user, while it's also important that the agent should take the initiative to interrupt the human user's floor and adjust the conversation flow. To improve the prediction performance, we propose a one-class classification method using multimodal features such as acoustic features, head

movement and facial expression. We also conducted a perceptual study to evaluate the acceptability of generated interruptions, including their timing and type.

## 7.2 Approach

We use NoXi corpus for this study. During a conversation, interruptions can occur at various timings. The first step in building an interruption prediction model is to create a database. We use the interruption annotation described in Chapter 5. However, even if an interruption did not occur at a given moment during a real interaction, this does not mean that it could not have happened. Nevertheless, interruptions may occur at any moment. Some moments are perhaps less appropriate. We refer to these moments as negative samples.

Obtaining the ground truth of positive samples (occurrence of interruptions) is thus possible but obtaining negative samples (where interruptions should not occur) is more challenging: how to know that at a given moment it is impossible to interrupt?

To avoid this issue, Chylek et al. (Chylek et al. [2018]) assumed that the current speaker was purposefully not interrupted before a real interruption ( $t - 0.7s$ ), making the same number of negative examples as the positive ones. Having both positive and negative samples allows the use of discriminative methods such as SVM or neural networks.

Another approach we proposed in this article, to overcome the limitation of missing negative samples is to use a one-class classification model that does not need negative samples and can learn to detect interruptions based on existing positive samples alone. In this Chapter, we compared our approach with the method proposed by Chylek et al. (Chylek et al. [2018]) using multi-modal features extracted on 1s length temporal window.

We leveraged acoustic features extracted from openSmile (Eyben et al. [2010]): fundamental frequency, loudness, and 12 mel-frequency cepstral coefficients (MFCC). We also use facial expressions, gaze and head movements extracted by OpenFace (Baltrušaitis et al. [2016a]), including Action Units (AU) 01, 02, 04, 05, 12, and 15, gaze direction, as well as head activity and rotation as described in Chapter 4.

For all multimodal features, we calculated their average values on the corresponding temporal window length (0.7s for the approach of Chylek et al. (Chylek et al. [2018]), 1s for our approach ) and used this feature vector as input to both models. First, we followed the works of Chylek et al. (Chylek et al. [2018]) and used a deep residual learning network (ResNet-152) to classify samples. Positive samples correspond to the moment of real interruptions manually annotated and negative samples correspond to the timing of positive ones minus 0.7s. Moreover, as in (Chylek et al. [2018]), data are augmented by offsetting each moment by 1 to 3 samples. We use the presented multimodal features as we obtain better results than using only audio features as done in (Chylek et al. [2018]).

The second method we proposed, which does not need to create negative samples, is a one-class SVM with specific hyperparameters  $\gamma=0.1$  and  $\nu=0.3$ .

The output of the one-class SVM is a score representing the similarity of the input feature vector to the targeted class, which in our case is interruptions. The higher the output score, the higher the probability that an interruption will occur. By manually setting a threshold on the output, based on the frequency of interruptions on the validation data, the model outputs whether an interruption would happen.

To compare our one-class SVM model with the method presented by Chýlek et al. (Chýlek et al. [2018]), we followed a similar approach to use the annotated interruption onset moments as positive samples, and defined the moment  $-0.7s$  as negative samples, with an offset of 3 frames. For both methods, the model is trained on 19 conversations (validation set: 2 half videos from the 19 ones) and tested on the 2 remaining ones.

### 7.3 Comparative study

To evaluate the efficacy of the proposed approach, we conducted a comparative study with the method proposed by (Chýlek et al. [2018]). The results are presented in Table 7.1. It is important to note that we do not exactly obtain the results presented in (Chýlek et al. [2018]) as we use another database where participants spoke another language and discussed other topics in a different interaction setting.

The comparison showed that the inclusion of facial expressions and head motions enhances the prediction accuracy of interruptions compared to the use of only acoustic features. Our proposed one-class SVM model performs slightly better than the neural network model, even so, the accuracy is still low.

To further evaluate our proposed model, we conducted a perceptual study, which is presented in the next section.

	Accuracy	F1-score
Deep residual learning network (Chýlek et al. [2018] acoustic only)	0.56	0.56
Deep residual learning network (Chýlek et al. [2018] all modalities)	0.59	0.58
One-class SVM (ours)	0.61	0.61

Table 7.1 Accuracy & F1-score for Deep residual learning network and One-class SVM models with different modality combinations.

### 7.4 Subjective Evaluation

In this study, we want to evaluate the timing prediction model using perceptual study, by comparing ground truth, predicted interruptions and randomly selected ones. We consider 4 independent variables: interruption timing (ground truth,

## 7.4. SUBJECTIVE EVALUATION

---

predicted model, randomly chosen), interrupter speech (ground truth, scripted), interrupter audio voice (natural human audio or synthesised voice), interruption type (agreement, clarification, disagreement) and we added one more variable to be tested: interruption turn (ground truth turn or false-positive turn). The value of these variables is explained below. Thus we obtained 8 conditions we referred to as group 1...8:

- group1: ground truth timing – ground truth interrupter speech - natural human voice.
- group2: ground truth timing – scripted interrupter speech - natural human voice.
- group3: predicted timing - scripted interrupter speech - natural human voice.
- group4: ground truth timing - scripted interrupter speech - synthesised voice
- group5: predicted timing - scripted interrupter speech - synthesised voice
- group6: random timing - scripted interrupter speech - synthesised voice.
- group7: predicted timing - scripted interrupter speech - synthesised voice - ground truth interruption turn.
- group8: predicted timing - scripted interrupter speech - synthesised voice - false-positive turn.

Each group was evaluated by 30 participants using a questionnaire composed of 11 questions. In total, we had 270 participants with an acceptance rate >95%, including 115 males and 142 females, with 13 participants not specifying their gender. The majority of participants were between 18 and 30 years old (149 participants) or between 31 and 50 years old (105 participants), with 13 participants being between 51 and 70 years old. All the participants were fluent in speaking French.

### 7.4.1 Stimuli

To assess the predicted interruption timings, we compared them to the ground truth and randomly selected interruption timings. To accomplish this, we created stimuli (<https://youtu.be/WXvoUBEfADc>) for the different conditions described in the following sections.





Figure 7.1 Screenshot of generated video for an interruption. Interrupter on the left side ("Yep, yep."), and interruptee on the right side ("So after coming back from school he has some time to play and ...").

### Visual

We utilized a static image featuring two stylized young individuals to accompany the interrupter and interruptee audios. As shown in Figure 7.1, the person on the left represents the interrupter role, and the person on the right represents the interruptee. Subtitles were displayed concurrently with the audio for both the interrupter and interruptee beneath their corresponding roles.

### Interruption sentences

For the ground truth interruptions, we had access to the audio and the speech of the interrupters. However, this was not the case when the timing of interruption was chosen randomly or predicted by our model. Thus, from the database, we chose five interruptions where the interrupter used rather common sentences that are possible to be used in general cases. These five interruptions were categorized into three types: agreement, disagreement, and clarification, as shown in Table 7.2.

## 7.4. SUBJECTIVE EVALUATION

Agreement	1. 'Oui, oui oui' ('yep yep'). 2. 'Ouais c'est ça ouais' ('Yeah that's it').
Disagreement	3. 'Oh non, pas du tout' ('Oh no, not at all'). 4. 'Ah je suis pas d'accord du tout' ('Oh I don't agree at all').
Clarification	5. 'C'est à dire?' ('What's that?').

Table 7.2 Scripted interrupter speech sentences, selected from interruptions in our corpus.

### Audio voice settings

To examine the impact of the interrupter's voice, we used either natural human voice interrupter audios which were cut from the videos of NoXi database or synthesised voice interrupter audios with the selected sentences for different groups:

- groups 1, 2, 3: natural human voice
- groups 4, 5, 6, 7, 8: synthesised voice

In all conditions, we used the original audios from the database for interruptees, once interrupted, the interruptee audios were cut off right after completing the current word, as in Figure 7.2.

#### **Ground Truth:**

Interruptee: .... à une espèce de version qui est encore en cours de développement  
 Interrupter: C'est quoi ton point de vue sur les DLC?

#### **Predicted:**

Interruptee: .... à une espèce de version qui est encore  
 Interrupter: C'est à dire?

#### **Random:**

Interruptee: .... à une espèce de version  
 Interrupter: C'est à dire?

Figure 7.2 Interruption simulation: for the same speaking turn of interruptee, predicted interruption and random interruption may occur at different timing as ground truth. Once interrupted, the interruptee audio was cut off after finishing the current word.

### 7.4.2 Comparison variables

We present the independent variables that we considered for the evaluation.

#### **Interrupter speech**

We compared the predicted interruption timing with the ground truth timing (groups 2 and 3). Group 1 was added to measure the effect of generic sentences on the perception of interruption. Groups 1, 2 and 3 all used natural human voice to avoid extra impact from the synthesised voice. We selected 20 turns from the conversations of the test data set where there is an interruption (ground truth interruption) and where our model predicted one (predicted interruption) in the same turn. All the selected interruption samples were of the three interruption types: agreement, clarification and disagreement. Groups 2 and 3 used scripted interrupter speech sentences. For each interruption sample of groups 2 and 3, we randomly chose one sentence from the selected scripted speech that was of the same interruption type as the ground truth.

#### **Interruption timing**

We compared the predicted interruption timing with ground truth timing and random timing (groups 4, 5 and 6). We used scripted interrupter speech and synthesised voice for the three groups. We used the same interruption samples as selected above for groups 1, 2 and 3. The interrupter speeches were also the same as for groups 1, 2 and 3, but all with synthesised voices.

#### **Interrupter audio voice**

We compared the impact of the synthesised voice and the natural human voice on interruption perception, we conducted two comparisons to measure for both ground truth interruption timings (group 2 vs. group 4) and predicted interruption timing (group 3 vs. group 5), this can give the insight on how the interrupter audio voice would impact the perception of interruption timing.

#### **Interruption turn & interruption type**

As previously mentioned, the absence of an interruption at a specific moment does not imply that an interruption could not have occurred. In other words, interruptions may occur during speaking turns other than those with annotated ground truth interruptions. We thus compared the predicted interruptions for speaking turns that contained a ground truth interruption (i.e., ground truth turns, group 7) and those that did not but were predicted to have one (i.e., false positives turns, group 8). For Group 7, we selected 10 turns from the 20 which were used for groups 1 to 6, while for false positive interruptions in group 8, we selected 10 turns from the conversations where we could only find the predicted interruption and no ground truth interruption.

## 7.4. SUBJECTIVE EVALUATION

---

To further investigate how different types of interruptions are perceived, we generated videos for all three interruption types using selected interrupter speech sentences. For each interruption sample in groups 7 and 8, we made one video for each interruption type, thus, for each interruption sample, there were 3 videos with different types of interrupter speech. A total of 30 videos (10 turns \* 3 types) were created for each group.

### 7.4.3 Questions

For each video, participants answered 11 questions (see Table 7.3) related to the timing of the interruption, the type of interruption, and their perception of the interrupter, using a 5-point Likert scale ranging from "strongly disagree" to "strongly agree". Participants were instructed to select the option that best reflected their opinion based on their first impression.

Do you think the interruption is	1. well placed? 2. acceptable? 3. coherent?
Do you think the interrupter is	4. competitive? 5. cooperative? 6. dominant? 7. friendly?
Do you think the interrupter	8. is trying to control the conversation? 9. intend to take the floor? 10. should let his interlocuter finish what he was about to say? 11. shouldn't have interrupted?

Table 7.3 Evaluation questions. Separated into three question sets. The 11 questions were asked for all the evaluated interruptions, randomly ordered.

### 7.4.4 Comparison Results

In this section, we present the comparison of different groups regarding the independent variables as mentioned above, an overview of the 11 questions for the 8 groups is presented in Figure 7.3. We conducted a post-hoc pairwise comparison using Tukey's honestly significant difference (HSD) test, the specific results are displayed in Table 7.4. In the case of mean differences, a negative value indicates that the first group is rated lower than the second group for the corresponding question, reversely for a positive value, and the farther the mean difference value from 0, the stronger the evidence.

## 7.4. SUBJECTIVE EVALUATION

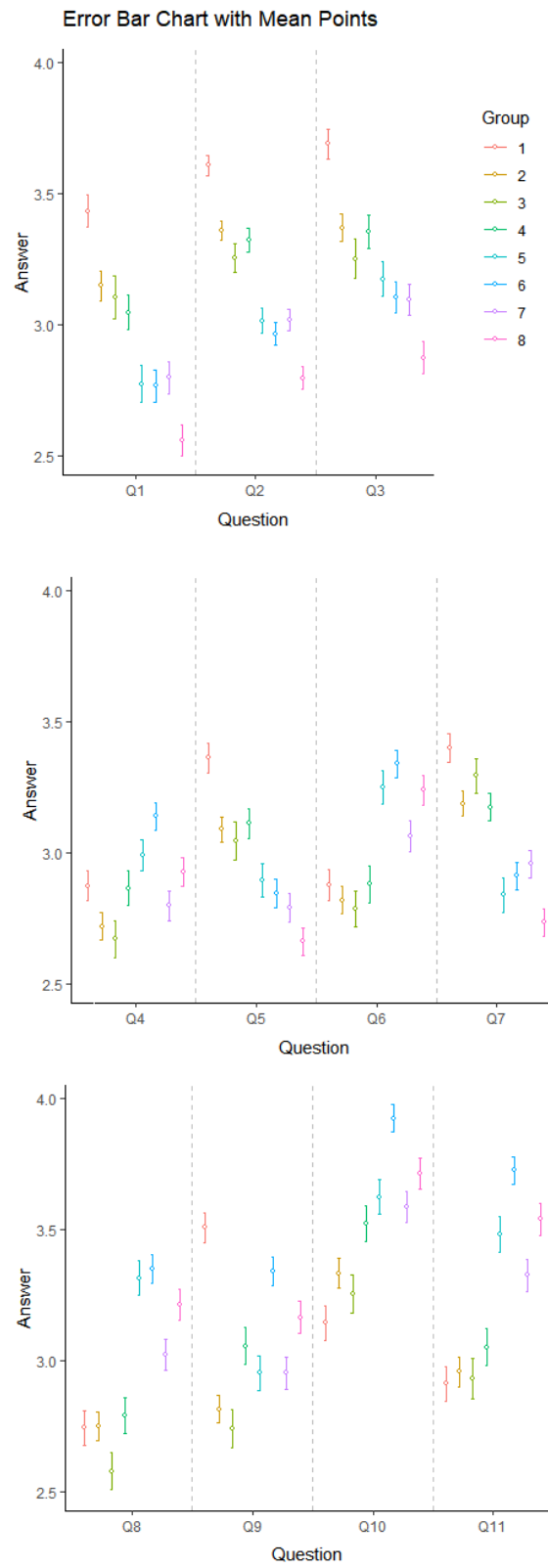


Figure 7.3 Error bar plot of 8 groups by question.

## 7.4. SUBJECTIVE EVALUATION

Groups	Comparison	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
G1-G2	Original sentence vs. scripted interrupter speech	<b>0.29</b>	0.17	<b>0.32</b>	0.15	<b>0.27</b>	0.06	0.21	-0.01	<b>0.69</b>	-0.19	-0.05
G1-G3		<b>0.33</b>	0.27	<b>0.44</b>	0.20	<b>0.32</b>	0.09	0.11	0.17	<b>0.77</b>	-0.11	-0.02
G2-G3		0.04	0.09	0.12	0.05	0.05	0.03	-0.11	0.17	0.08	0.08	0.02
G4-G5	Ground truth timing vs. predicted timing	0.27	<b>0.44</b>	0.18	-0.13	0.22	<b>-0.37</b>	<b>0.34</b>	<b>-0.52</b>	0.10	-0.10	<b>-0.43</b>
G4-G6		<b>0.28</b>	<b>0.47</b>	0.25	<b>-0.27</b>	<b>0.27</b>	<b>-0.46</b>	<b>0.26</b>	<b>-0.56</b>	<b>-0.29</b>	<b>-0.40</b>	<b>-0.68</b>
G5-G6	vs. random timing	0.01	0.03	0.07	-0.15	0.05	-0.09	-0.07	-0.04	<b>-0.39</b>	<b>-0.30</b>	<b>-0.24</b>
G2-G4	Natural human voice vs. synthesised voice	0.10	0.06	0.01	-0.15	-0.02	-0.06	0.01	-0.04	<b>-0.24</b>	-0.19	-0.09
G3-G5		<b>0.33</b>	<b>0.40</b>	0.08	<b>-0.32</b>	0.15	<b>-0.46</b>	<b>0.46</b>	<b>-0.74</b>	-0.21	<b>-0.37</b>	<b>-0.55</b>
G7-G8	Ground truth vs. false positive	0.24	0.22	0.22	-0.13	0.13	-0.18	<b>0.22</b>	-0.19	-0.21	-0.13	-0.21

Table 7.4 Comparison of different groups for 11 questions are presented in the table. Mean differences are reported, with a positive value indicating that the first group scored higher than the second group. Significant differences ( $p < 0.05$ ) between the first and second groups are highlighted in green colour.

First groups	Second groups	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
Agreement	Clarification	0.24	0.09	0.18	-0.16	0.10	0.00	0.15	-0.13	0.01	-0.10	-0.03
Agreement	Disagreement	<b>0.38</b>	<b>0.35</b>	<b>0.28</b>	<b>-0.50</b>	<b>0.45</b>	<b>-0.39</b>	<b>0.45</b>	<b>-0.58</b>	<b>-0.72</b>	-0.13	-0.16
Clarification	Disagreement	0.14	<b>0.26</b>	0.10	<b>-0.34</b>	<b>0.34</b>	<b>-0.39</b>	<b>0.30</b>	<b>-0.44</b>	<b>-0.72</b>	-0.02	-0.13

Table 7.5 Comparison of different interruption types for 11 questions are presented in the table (groups 7 and 8). Mean differences are reported, with a positive value indicating that the first group scored higher than the second group. Significant differences between the first and second groups are highlighted in green colour.

### Interrupter speech

We conducted the comparison of the stimuli of groups 1, 2, and 3 to see how interrupter speech impacts the perception of the interruption, which corresponds to the first three comparison results in Table 7.4. The stimuli of group 1 used the original interrupter speech and audio, that of group 2 used scripted interrupter speech with a natural human voice but maintained the ground truth interruption timing, and the stimuli of group 3 used predicted interruption timing with scripted interrupter speech and a natural human voice. The interruptions of all three groups were perceived as acceptable (Q2), with an average score higher than the neutral level of 3 (see Figure 7.3). However, the interruptions of group 1 were evaluated as significantly more coherent (Q3) than those of groups 2 and 3, which used scripted interrupter speech even though the interruption timings were the same for groups 1 and 2.

Regarding the perception of interrupters, Q4 ~ Q7 in Table 7.4 revealed that interrupters in group 1 were perceived as more cooperative (Q5) than those in groups 2 and 3 which received similar scores. There were no remarkable differences in competitiveness (Q4), dominance (Q6), and friendliness (Q7) between the three groups. All three groups were perceived as friendly and not particularly competitive/dominant. Furthermore, the interrupters in group 1 were perceived as more likely to grab the turn (Q9) than the interrupters in groups 2 and 3 with an average score of above 3, and the other two groups below the neutral level. All three groups were rated as the interrupter should let the speaker finish talking

(Q10), but none of them tried to control the conversation (Q8), aligning with the result for the perception of the interrupter's dominance.

In summary, the difference between the stimuli of group 1 and of groups 2 and 3 lies in the interrupter speech content. Group 1's stimuli used the original ground truth interrupter speech and audio, which was coherent with the conversation, while those of groups 2 and 3 used scripted interrupter speech which were "common" sentences (e.g., 'yep yep') but which were not related to the current conversation's content. This highlights the importance of the interruption speech content, and that may explain why stimuli in groups 2 and 1, which did not differ in interruption timing, were perceived with significant differences.

### **Interruption timing**

We compared the stimuli of groups 4, 5, and 6, which all used synthesised voice with scripted interrupter speech to study how the different interruption timings are perceived. Group 4 used the ground truth interruption timing, group 5 used predicted interruption timing, and group 6 used randomly chosen interruption timing. These results correspond to the 4th to 6th comparisons in Table 7.4.

From Table 7.4, group 4's interruptions were perceived as more acceptable (Q2) than groups 5 and 6, while groups 5 and 6 were rated similarly. Group 4's interruptions were also found to be better placed (Q1) than those of group 6, but no significant difference was found between the comparisons of group 4 vs. group 5, and group 5 vs. group 6. All three groups' interruptions were perceived as coherent (Q3), with no significant difference in score and all rated above the neutral level.

In terms of perception of the interrupters (Q4 to Q7 in Table 7.4), the interrupters of group 4 were perceived as more cooperative (Q5) and less competitive (Q4) than those of group 6, but no significant difference was found between the comparisons of group 4 vs. group 5, and group 5 vs. group 6. Group 4's interrupters were also perceived as more friendly (Q7) and less dominant (Q6) than the other two groups, while the stimuli of groups 5 and 6 showed no remarkable difference.

Table 7.4 also shows that the interrupters of groups 5 and 6 were perceived as more likely to control the conversation (Q8) compared to those of group 4, with no significant difference between those of groups 5 and 6. The interrupters also received a higher score for the term "should not have interrupted" (Q11) than those of group 4, although all three groups were rated above the neutral level. Compared to the interrupters of groups 4 and 5, those of group 6 were perceived as more likely to grab the floor (Q9) and should let the speaker finish the turn (Q10), where no significant difference was found between the interrupters of groups 4 and 5.

In summary, the distinguishing factor among groups 4, 5, and 6 lies in the timing of the interruptions. All three groups utilized the same interrupter speech and audio. Under these conditions, Group 4 (ground truth) was perceived as more

acceptable and friendly compared to Groups 5 (predicted) and 6 (random). This highlights the importance of selecting the appropriate interruption timing.

### **Interruption turn & interruption type**

We examined the results obtained from the stimuli of groups 7 and 8, which all utilized synthesised voice with scripted interrupter speech. The stimuli of group 7 were based on predicted interruptions on speaking turns that are also interrupted in the ground truth (but with different timing of occurrence), while those of group 8 were based on false positive predicted interruptions. The purpose of this test was to investigate how false-positive turn interruptions were perceived and whether they differed from ground-truth turn interruptions.

The results presented in Table 7.4 show that the only significant difference between the stimuli of groups 7 and 8 is that the interrupters of group 7 were perceived as more friendly (Q7) than those of group 8.

We further analyzed the differences between different interruption types in groups 7 and 8: agreement, clarification, and disagreement. The results are shown in Table 7.5. Cooperative types of interruptions (agreement vs. clarification) were perceived similarly in all aspects. Agreement and disagreement interruptions were significantly different in all aspects. Compared to disagreement interruptions, the agreement type was perceived as better placed (Q1), more acceptable (Q2) and coherent (Q3), with interrupters perceived as more cooperative (Q5) and friendly (Q7), and less competitive (Q4) and dominant (Q6). Clarification and disagreement interruptions had no significant differences in terms of placement (Q1) and coherence (Q3).

Regarding the perception of the interrupter, interrupters of agreement and clarification interruptions were more likely to control the conversation (Q8) and grab the turn (Q9) compared to those of disagreement interruptions, but all three types were rated similarly in terms of 'should not have interrupted' (Q11) and 'should let the speaker finish the turn' (Q10).

### **Interrupter audio voice**

To figure out the influence in perception when natural human voice or synthesised voice was used for different interruption timing, we compared the results between group 2 and group 4, where both groups are of ground truth interruption timing, group 2 used a natural human voice and group 4 a synthesised one. We also compared group 3 and group 5, both are of predicted timing, group 3 with natural human voice and group 5 with synthesised one. The stimuli of groups 7 and 8 that have ground truth interruption timing were evaluated with almost no significant difference while those with predicted timings were perceived with statistical differences.

Comparing the two groups of predicted interruptions, the natural human voice group interruptions were perceived as more acceptable (Q2) than the synthesised voice group, and the interrupters of the natural human voice group were per-



ceived as less competitive (Q4) or dominant (Q6) than the synthesised voice ones, who were perceived as more like interrupting unnecessarily (Q10, Q11). Participants might become more sensitive to interruption timing when there is no longer natural intonation.

## 7.5 Discussion

Overall, our experiments provide insights regarding how different factors, including speech content, interruption timing and its type, affect the perception of interruptions and interrupters. These findings have implications for how interruptions are perceived in various contexts.

The interruptions of group 1 (ground truth stimuli) compared to those of group 2 (similar to group 1 stimuli but with scripted common sentences such as "yep yep") were evaluated as better placed and more coherent. The interrupters in group 1 stimuli were considered more cooperative, indicating that the content of an interruption is an important factor in how it is received by the listener. Thus, the ground truth interruptions with coherent speech content were perceived as the most effective.

The difference between the perception of stimuli of groups 4, 5, and 6 lies in the interruption timing, all using scripted interrupter speech with a synthesised voice. Ground truth timings were perceived as more acceptable, friendly, and cooperative. In contrast, predicted interruption timings and randomly chosen timings were perceived as more dominant, interrupting unnecessarily, and trying to control the conversation. Predicted timings showed differences with randomly chosen timings in that the interrupters of the random group were perceived as more likely to grab the floor and should let the speaker finish. The placement of an interruption (be ground truth timing (group 4), predicted timing (group 5) or random timing (group 6)) has an influence on participants' perception. But this influence was not major.

The predicted interruptions were found to be rated with scores similar to the group with ground truth timing when using scripted interrupter speech with the natural human voice (group 3 vs group 2). The only variable that differs between these comparisons is the interrupter's audio voice. Previous studies on conversation interruption indicated that interrupters may use a special pattern to gain attention and grab the floor, such as, for example, higher pitch and loudness ([French and Local \[2018\]](#), [Hilton \[2018\]](#)). However, the synthesised voice is rather flat, and our stimuli did not model the intonation contours as in a natural human voice. The lack of natural-sound interrupter's intonation seems to have an impact on the perception of the interruption and the interrupter. For stimuli with scripted sentences and synthesised voice, there is a decrease in coherence between both interlocutors' speech. This is reflected in the results of group 2 vs. group 4 and of group 3 vs. group 5.

Moreover, we can see that randomly chosen interruptions were rated rather similar to the ground truth interruption timings in several aspects and did not

show many differences compared to predicted ones. Moreover, when using synthesised voice and scripted sentences, interrupting as in the ground truth (group 7) or at other moments (group 8, false positive computed by our model) did not make a significant difference in perception. One possible reason is that interruptions may not have to occur at specific timings during a conversation. There seems to be quite a lot of flexibility in the occurrence timing of the interruption. However, this result is modulated by other factors (coherence of the interrupter speech sentences and voice quality).

When an interruption occurs, the interruption type would play a rather important role. In our study, interruption type affects significantly how interruption and interrupter are perceived. Interruption type was recognized by participants as playing a critical role and was perceived differently even with the same interruption timings. The perception of the three types of interruptions (agreement, clarification, disagreement) showed remarkable differences in the evaluation score.

## 7.6 Conclusion

In this Chapter, we presented a novel approach to predict interruptions during conversations through the use of a one-class classification model with multimodal features from the speaker. To evaluate the effectiveness of our model, we conducted an objective study and a perceptual experiment to gain insights into how interruptions are perceived under different conditions. Our model achieved an overall accuracy of approximately 0.61. This result, even though better than those obtained in the state of the art, suggests that there is still room for improvement. However, our perceptual experiment highlights that interruption timing may not be the prime factor; rather it is the quality of the voice (i.e. the use of natural voice) that seems more important.

### The key points of this Chapter:

*This Chapter addresses research question Q4, Q5 and Q6?*

- How should ECAs decide when to interrupt during real-time interaction?
- What are the most effective modalities to find possible interruption timing?
- How are the predicted ECA interruptions perceived? What are the major factors that impact the perception of interruptions?

*One-PredIT: One-class Prediction of Interruption Timing.*

- We use a one-class SVM model to avoid the problem of the negative samples definition.
- We trained the proposed model on the NoXi corpus and conducted a perceptual study to evaluate the predicted interruptions.
- In order to adapt our model in the real-time application, only the speaker's nonverbal behaviour before each annotated interruption point was used to predict the possibility of the very next moment.
- Using a one-class classification model, we manually settled a threshold based on the distribution of the interruption.
- Multiple variables concerning interruption were considered and evaluated (Interrupter speech, Interruption timing, Interrupter audio voice, Interruption turn and Interruption type), Interruption timing was found not the major factor impacting the interruption perception but the Interrupter audio voice and Interruption type.

# AI-BGM: Agent Interruption Behaviour Generation Model.

## Contents

---

8.1	Related works . . . . .	96
8.1.1	Nonverbal turn-taking behaviour for virtual agents . . . . .	96
8.1.2	Nonverbal behaviour generation . . . . .	97
8.1.3	Transfer learning . . . . .	98
8.2	Corpus & Features . . . . .	101
8.3	Interruption preparation duration . . . . .	101
8.4	Smooth turn exchange vs. interruption . . . . .	103
8.5	ASAP introduction . . . . .	104
8.6	Proposed model . . . . .	106
8.7	Objective evaluation . . . . .	107
8.8	Subjective evaluation . . . . .	109
8.8.1	Video preparation . . . . .	109
8.8.2	Questionnaire . . . . .	111
8.8.3	Subjective evaluation results . . . . .	112
8.9	Discussion . . . . .	113
8.10	Conclusion . . . . .	114

---

Building upon the findings from the previous chapter’s experiments, it was revealed that in the context of human-agent interaction, the timing of when a virtual agent interrupts a human is not the most crucial factor. Surprisingly, the quality

of the sound and the content of the interruption have a more significant impact on the perception of interruption. Given that our previous experiments only employed audio data, we are particularly intrigued by the influence of the agent’s nonverbal behaviour on the overall perception. Hence, we introduce a generative model designed to generate nonverbal behaviours, encompassing facial expressions and head rotations, exhibited by the interrupter during the interruption phase, enhancing the authenticity and naturalness of interruptions within human-agent interactions. It ensures that the interruptions are imbued with nonverbal cues to align with effective communication. This chapter corresponds to the research questions posed in Q7, Q8 and Q9.

In order to delve deeper into this aspect, we aim to investigate the effects of nonverbal behaviour generated by the agent during interruptions. To accomplish this, we will maintain control over sound quality, interruption content, and timing—utilizing the original dataset as a foundation. Employing the technique of transfer learning, we extend the capabilities of an already proficient model designed for generating general virtual agent interaction nonverbal behaviour. This model is fine-tuned using data coming specifically from the interruption periods, enabling it to generate nonverbal behaviour (facial expressions and head movements) tailored to instances when the agent acts as an interrupter.

Through a series of experiments, our objective is to closely observe the virtual agent’s actions during human interruption scenarios and evaluate whether these generated nonverbal behaviours are perceived as acceptable by human participants. Furthermore, we compare the quality and appropriateness of the generated nonverbal behaviour with the nonverbal behaviour exhibited by actual human interrupters in the original corpus.

By conducting this research, we address the gap in the understanding of virtual agent nonverbal behaviour generation during interruption scenarios. We seek to gain deeper insights into the impact of such behaviour on human users and provide substantial guidance for enhancing the social interaction capabilities of virtual agents.

## 8.1 Related works

### 8.1.1 Nonverbal turn-taking behaviour for virtual agents

When exploring the existing landscape of virtual agent systems in terms of interruption management, the majority of research has predominantly focused on treating the virtual agent as the party being interrupted. Within this context, studies delve into how virtual agents should respond and handle interruptions from human users, such as determining when to cease speaking and yield the conversational floor. Our focus then shifts towards a closely related aspect—virtual agent turn-taking management.

In theory, nonverbal behaviour plays a crucial role in turn-taking, encompassing elements like gestures and eye gaze. Zhou et al. (Zhou et al. [2018]) demon-

strated the use of gaze shifts to signal turn-taking and body posture shifts to indicate changes in topic. Similarly, Kontogiorgos et al. (Kontogiorgos et al. [2019]) utilized gaze cues to coordinate turn-taking. In their work, prior to an utterance, the agent employed a gaze shift to the subject to establish attention. During the utterance, the deictic gaze was directed towards a referent object, indicating the agent's retention of the conversational floor. At the conclusion of the utterance, a gaze shifts back to the participant established at the end of the turn. Andrist, Leite et al. (Andrist et al. [2013]) conducted a study where an agent managed a group of children. This agent employed multimodal cues, including gaze, gesture, and proxemics, to facilitate the exchange of speaking turns. The study demonstrated that an agent incorporating all three cues resulted in a more equitable distribution of speaking turns within the group, reducing variance. When compared to an agent utilizing only vocal cues or a subset of nonverbal cues, the comprehensive use of all three cues proved most effective in managing the conversation without diminishing enjoyment.

However, the aforementioned turn-taking nonverbal behaviours are typically rule-based, and designed according to predefined protocols. In real conversations, turn-taking and interruptions are flexible and variable, demanding that virtual agents adapt to diverse situations. Anticipating every possible scenario and designing pre-established responses is often unfeasible. Consequently, we have developed a model capable of generating interruption-specific behaviours in real-time. This dynamic approach aims to equip virtual agents with the adaptability required to handle a range of unforeseen circumstances in both turn-taking and interruption scenarios.

### 8.1.2 Nonverbal behaviour generation

The task of generating nonverbal behaviours can be likened to predicting forthcoming sequences of non-linguistic actions, suggesting an intriguing avenue to explore existing sequence prediction techniques for potential applicability to nonverbal behaviours. The generation of multimodal behaviours in Socially Interactive Agents (SIAs) necessitates the accurate modelling of the temporal dynamics underlying exchanged social cues such as facial expression and head motion to enhance the authenticity and naturalness of the interaction. Facial gestures are consciously or unconsciously used to accentuate words or mark speech pauses. Many facial expressions and head nods are tied to the speech's syntactic and prosodic structure.

Past research endeavours that delve into intrapersonal temporality have put forth models capable of generating facial expressions and communicative gestures in coordination with speech. These studies harness the power of Deep Learning (DL) techniques, including Feed-Forward Neural Networks (FFNs), Bidirectional Long Short-Term Memory (BLSTM) networks, Conditional Variational Autoencoders (CVAEs), Generative Adversarial Networks (GANs), and Transformers (Alexanderson et al. [2020], Bhattacharya et al. [2021], Ding et al. [2015], Fares et al. [2022], Ferstl et al. [2019], Greenwood et al. [2017], Hasegawa

et al. [2018], Karras et al. [2017], Sadoughi and Busso [2018], Yuan and Kitani [2020]).

Feng et al. (Feng et al. [2017]) mentioned that attention is directed towards the relationship between a human user and an SIA. They employ a Feed-Forward Neural Network (FFN) model to generate the agent’s facial gestures based on previously predicted facial gestures from both the agent and the human. This approach solely employs visual features like facial landmarks and doesn’t leverage the multimodal information available in the interaction. Grafsgaard et al. (Grafsgaard et al. [2018]) adopt a Long Short-Term Memory (LSTM) model to encode multimodal signals (facial expression, body motion, and speech). Their model is used to predict the facial expression and motion of a partner by incorporating speech from both partners and their respective facial expressions and motion features. Dermouche et al. (Dermouche and Pelachaud [2019]) also delve into the interpersonal relationship, framing it as an interactive loop for agent behaviour generation. They introduce the Interactive Loop LSTM (IL-LSTM), which considers the upper-face behaviours of both the agent and the user to model the agent’s nonverbal behaviours. Similar to (Feng et al. [2017]), IL-LSTM’s limitation lies in taking only unimodal input features (facial gestures), leading to jerky movements due to its sliding window prediction.

Motion generation entails the use of generative models like Generative Adversarial Networks (GANs) (Goodfellow et al. [2014]) and normalizing flow-based models to produce more diverse and realistic motions. An extended version of the MoGlow system (Henter et al. [2020]) is employed by Jonell et al. (Jonell et al. [2020]) to predict the agent’s facial expression based on audio input from both partners and the human’s facial expression. Tuyen et al. (Tuyen and Celiktutan [2022]) forecast upper body motions (face, body, and hand landmarks) through a context-aware model comprising three components: a context encoder, a generator, and a discriminator.

Focusing on modelling reciprocal adaptation for SIA behaviour generation, Woo et al. (Woo et al. [2023]) present the Augmented Self-Attention Pruning (ASAP) neural network model. ASAP seamlessly integrates a recurrent neural network, the attention mechanism from transformers, and a pruning technique to facilitate learning of reciprocal adaptation through multimodal social signals. This novel approach encapsulates the dynamic interaction between SIAs and humans, paving the way for more sophisticated and contextually aware nonverbal behaviour generation.

### 8.1.3 Transfer learning

Despite interruptions being a common occurrence in everyday conversations, their overall frequency remains relatively limited. Consequently, we face a challenge in obtaining a sufficient amount of data to train an entirely new action generation model from scratch. To address this issue, we have opted to leverage a pre-trained and validated model that has demonstrated its capability to generate general virtual agent interaction nonverbal behaviour. Building upon this foundation, we

intend to utilize existing interruption data to retrain this model, thereby adapting it specifically for interruption behaviour generation. Given the scarcity of interruption-specific data, this approach capitalizes on the knowledge and general behaviour-generation abilities already embedded in the pre-trained model. By incorporating interruption data into the training process, we aim to fine-tune the model's responses to align with the dynamics of interruption scenarios. This strategy not only maximizes the use of available resources but also allows us to harness the nuanced knowledge captured by the initial general behaviour model.

In the realm of artificial intelligence and machine learning, the concept of transfer learning has emerged as a powerful strategy that enables models to leverage knowledge gained from one task to enhance performance on another (Weiss et al. [2016]). Instead of building isolated models for each specific task, transfer learning facilitates the sharing of insights across related domains, paving the way for more efficient and effective learning processes. By enabling models to transfer and adapt knowledge, transfer learning not only accelerates the training process but also enhances the generalization capabilities of AI systems, leading to improved performance and resource utilization (Weiss et al. [2016]).

At its core, transfer learning involves training a model on a source task with the objective of transferring the acquired knowledge to a related target task. The intuition behind transfer learning is grounded in the observation that real-world tasks often share common underlying patterns and structures. By leveraging the information extracted from one task, the model can gain a head start when confronted with a new, yet related, problem. This is particularly valuable when labelled data for the target task is limited or expensive to obtain.

Transfer learning encompasses various methods, each catering to different scenarios and learning objectives.

### Homogeneous Transfer Learning

Homogeneous Transfer learning approaches are developed and proposed to handle situations where the domains are of the same feature space. In Homogeneous Transfer learning, domains have only a slight difference in marginal distributions. These approaches adapt the domains by correcting the sample selection bias or covariate shift.

**Instance-based.** It covers a simple scenario in which there is a large amount of labelled data in the source domain and a limited number in the target domain, domains differ only in marginal distributions (Chattopadhyay et al. [2012]).

Instance-based transfer learning involves reassigning different weights to instances from the source domain in the loss function of the target domain. This approach allows the model to prioritize learning from instances that are more relevant to the target task.

**Parameter-based.** The parameter-based transfer learning approaches transfer the knowledge at the model/parameter level. Parameter-based transfer learning focuses on adapting the model's parameters from the source task or domain to the target task or domain. Fine-tuning and freezing certain layers of a pre-trained



model are common techniques used in parameter-based transfer learning (Torrey and Shavlik [2010]).

- **Soft weight sharing:** In this approach, model parameters are adapted gradually, allowing the model to adjust to the nuances of the target domain while retaining some learned knowledge from the source domain.
- **Hard weight sharing:** Here, certain layers of the pre-trained model are shared entirely, ensuring that the model capitalizes on the shared knowledge and structure present in the source domain.

**Feature-based transfer.** Feature-based transfer learning centres on extracting valuable features from the source domain and adapting them to the target domain (Zhuang et al. [2020]). Through this approach, domain-specific discrepancies are minimized, allowing the model to capitalize on shared patterns and information. This approach can further be divided into two subcategories, i.e., asymmetric and symmetric Feature-based Transfer Learning.

Asymmetric approaches transform the source features to match the target ones. In other words, we take the features from the source domain and fit them into the target feature space. There can be some information loss in this process due to the marginal difference in the feature distribution. Symmetric approaches find a common latent feature space and then transform both the source and the target features into this new feature representation.

**Relational-based transfer.** Relational-based transfer learning approaches mainly focus on learning the relations between the source and a target domain and using this knowledge to derive past knowledge and use it in the current context (Weiss et al. [2016]). Such approaches transfer the logical relationship or rules learned in the source domain to the target domain.

### Heterogeneous Transfer Learning

Homogeneous transfer learning involves deriving representations from a previous network to extract meaningful features from new samples for an inter-related task. However, these approaches forget to account for the difference in the feature spaces between the source and target domains. Heterogeneous transfer learning is pertinent when the source and target domains differ significantly in terms of data distributions and feature spaces. This category addresses the challenge of adapting knowledge from one domain to another that may have disparate characteristics (Weiss et al. [2016]).

In summary, transfer learning is a versatile methodology that encompasses a spectrum of techniques catering to different data scenarios. By enabling models to learn from prior experiences in related contexts, transfer learning accelerates the learning process, enhances predictive accuracy, and enables models to tackle real-world challenges effectively. We have explored several approaches in this work, discussed in the subsequent sections.

## 8.2 Corpus & Features

For the purpose of our study, we opted to utilize the French segment of the NoXi database. Within this methodology, visual features were extracted for each time-step, encompassing gaze direction data (around both the x and y axes), head rotation data (around the x, y, and z axes), and the activations of 6 upper face Action Units (AUs) – specifically, AU1, AU2, AU4, AU5, AU6, and AU7 – along with the smile expression (AU12).

Concurrently, we captured a range of acoustic features, encompassing the fundamental frequency, loudness, voicing probability, and a set of 13 Mel-frequency cepstral coefficients (MFCCs) (Logan et al. [2000]).

The extraction and cleaning of both audio and visual features have been detailed in Chapter 4.

## 8.3 Interruption preparation duration

Watching the original video data of interrupters and interruptees during interruption instances, a noticeable pattern emerged wherein the interrupter exhibited brief and distinct actions shortly before commencing speech. These actions encompassed changes in facial expressions, as well as subtle movements in head and body posture. These actions seemed to clearly convey the intent of the interrupter to speak. Notably, psycholinguistic research has demonstrated that producing even a single-word utterance takes at least 600 milliseconds, a time during which preparatory actions are typically observed (Indefrey and Levelt [2004]). In our pursuit of enhancing the human-agent interaction experience to align more closely with natural conversation habits, we incorporated the generation of these preparatory actions into consideration.

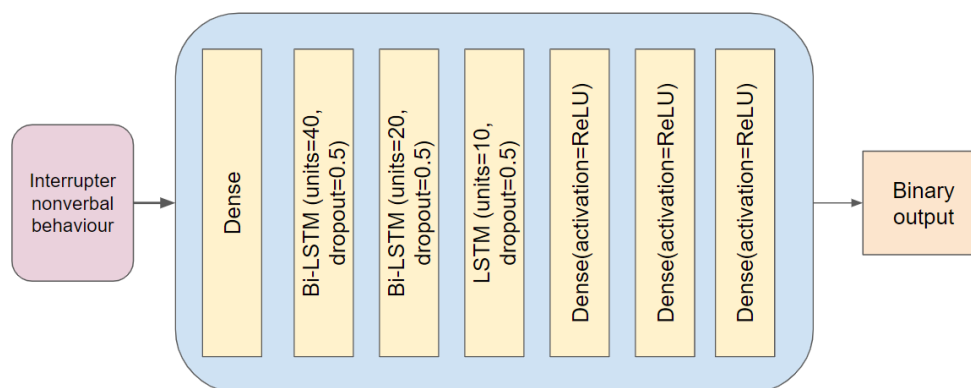


Figure 8.1 Classification model architecture.

Given the exceedingly brief nature of preparatory actions, an exact temporal duration is hard to be determined by analyzing data from various modalities.

### 8.3. INTERRUPTION PREPARATION DURATION

Thus, we resorted to training a classification model. This model used facial expressions (AU1, AU2, AU4, AU5, AU6, AU7, and AU12), gaze direction (around both the x and y axes) and head rotations (around the x, y, and z axes) of the interrupter as inputs to classify time segments prior to an interruption and those further away, serving as periods of attentive listening. The model's accuracy and classification capabilities served as indicators. When the model struggled to consistently differentiate between pre-interruption and listening periods, we could infer a rough estimate of the preparatory action duration.

Initiating the process, we extracted input data from the listening periods and excluded instances where the listener interjected backchannels. From these segments, we took the middle second, ensuring that it was far from the listener entering a listening state or preparing to speak at the end of a turn. After gathering all listening period data, we chose an equal number of segments (train, validation, and test) as in the interruption periods, totalling 929 segments.

For the periods before interruptions, we selected one-second segments at various distances from the interruption onset point (when the interrupter begins speaking), with offset duration ranging from 0s to 1.2s in increments of 0.2s (See Figure 8.2). This resulted in 929 segments for each time offset.

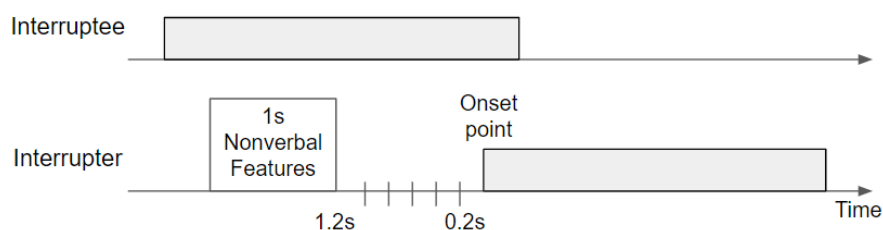


Figure 8.2 One-second segments with offset distance (each 0.2s from 0.2s to 1.2s).

Out of the 21 conversations, we reserved three conversations as separate test data, while the remaining 18 were divided into train and validation data in an 80% to 20% ratio. Subsequently, we trained the model to distinguish between listening periods and the different time offsets before interruptions (0s, 0.2s, 0.4s, 0.6s, 0.8s, 1.0s, and 1.2s). We then evaluated the model's accuracy on the test data.

We employed a multi-layer bi-directional LSTM model, as illustrated in Figure 8.1. This model structure was identified through training for classification between listening periods and interruption periods 0.2s before the interruption, as it demonstrated favourable classification performance. Given the consistent input-output requirements, we maintained the same model structure for training all seven models.

## 8.4. SMOOTH TURN EXCHANGE VS. INTERRUPTION

---

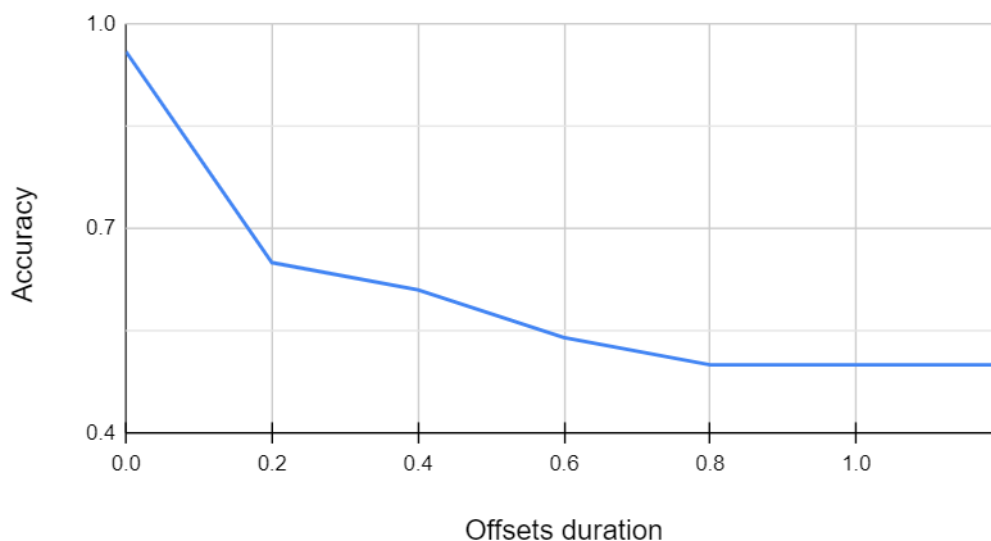


Figure 8.3 Classification model accuracy with different offset duration (0s, 0.2s, 0.4s, 0.6s, 0.8s, 1.0s, and 1.2s)

Figure 8.3 presents the test accuracy of the seven models corresponding to the different time offsets (0s, 0.2s, 0.4s, 0.6s, 0.8s, 1.0s, and 1.2s) before interruption. Up to 0.6s, four models sustained an accuracy of over 50%. Importantly, the 0.6s duration aligns with the language preparation duration mentioned in [Indefrey and Levelt \[2004\]](#). As a result, we selected the period of 0.6s prior to the interruption onset point as the baseline for generating interruption behaviour.

### 8.4 Smooth turn exchange vs. interruption

Recognizing the challenge posed by the limited data available for training the interruption behaviour generation model, our aim was to access a larger dataset. To address this issue, we turned our attention to examining data related to smooth turn exchanges. Drawing from insights in psycholinguistic research, as highlighted in [Indefrey and Cutler’s work \(Indefrey and Levelt \[2004\]\)](#), it is apparent that turn-taking in conversations often involves preparatory actions preceding the initiation of speech. This observation raises an intriguing possibility: could the preparatory actions associated with turn-taking be similar to those preceding interruptions? We sought to test whether the preparatory behaviours for smooth turn exchanges and interruptions exhibited similarity. If such similarity was established, it would open the possibility of using the preparatory behaviour data from smooth turn exchanges to train the interruption behaviour generation model.

Given the constraints of our data analysis methods and the challenge of effectively observing the interplay of various modalities, which might introduce biases in the results, we opted for a method analogous to our preparatory behaviour duration determination. We employed classification models to train and evaluate their accuracy and classification capabilities in assessing the similarity between the preparatory behaviours of smooth turn exchanges and interruptions.

For this investigation, we selected the data precisely one second before the onset point of both smooth turn exchanges and interruptions. We employed facial expressions (AU1, AU2, AU4, AU5, AU6, AU7, and AU12), gaze direction (around both the x and y axes) and head rotations (around the x, y, and z axes) as inputs. Similarly, we designated three conversations as test data and used the remaining 18 conversations for training and validation data, in an 80% to 20% distribution. Utilizing a model identical to that used for determining preparatory behaviour duration, the results demonstrated an accuracy of 83% in classifying the preparatory behaviours of both smooth turn exchanges and interruptions. This finding indicates the distinctiveness between the two, suggesting that the preparatory behaviours of smooth turn exchanges cannot be employed as training data for the interruption behaviour generation model. Consequently, in subsequent model training endeavours, we only utilized interruption data.

## 8.5 ASAP introduction

As previously mentioned, due to the limited dataset, we opted to utilize a pre-trained model that had been verified for its capability to generate general interaction behaviour. In this regard, we selected the Augmented Self-Attention Pruning (ASAP) model proposed by Woo et al (Woo et al. [2023]). Their objective was to model reciprocal adaptation by considering both intrapersonal and interpersonal temporality, as well as multimodality, along with the aspect of continuity, for the generation of an agent’s nonverbal behaviour.

The architecture of the ASAP model, depicted in Figure 8.4, presents a novel approach to model reciprocal adaptation. ASAP calculates the behaviour of the agent in real-time based on the behaviour of the human and its own behaviour. For each time frame ( $t + 1$ ), the model takes into account the previous 100 frames ( $t - 99 : t$ ) for both the human and the agent. ASAP incorporates three fundamental techniques: data augmentation, self-attention pruning, and autoregressive adaptive online prediction.

**Data Augmentation:** To effectively learn reciprocal adaptation, Woo et al. introduced a data augmentation technique that equally learns from both interlocutors. This method enables the model to capture the characteristics of both interacting partners. During each batch of the training phase, the model randomly assigns the interlocutor identity to be portrayed by the agent, allowing it to learn and predict behaviours for that particular interlocutor. This process is then alternated, with the agent assuming the identity of the other interlocutor, thereby fostering comprehensive learning.

**Self-Attention Pruning:** In order to better capture reciprocal adaptation and encapsulate the coherence, synchrony, and multimodality of interpersonal behaviours, a selection of relevant features is implemented through an attention mechanism. The self-attention process utilizes multi-head attention from Transformers (Vaswani et al. [2017]), involving all features (including eye movements, head rotations, smile (AU12), and 6 upper face AUs) across all interlocutors and

## 8.5. ASAP INTRODUCTION

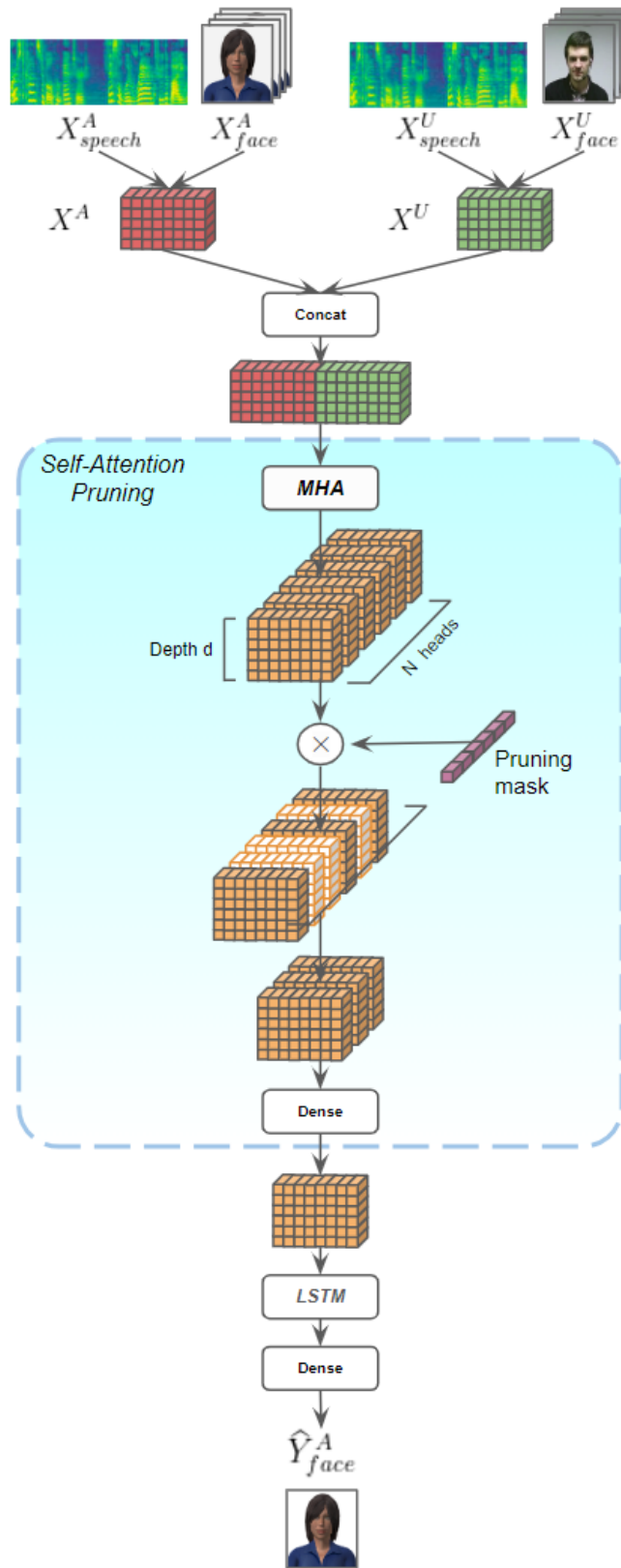


Figure 8.4 ASAP model architecture. Image from Woo et al. [2023].

modalities (visual and acoustic). This mechanism strategically captures pivotal information essential for modelling the occurrence of behaviours, mimicry, and synchronization in a comprehensive manner. However, within the multi-head attention (MHA), several heads often contain redundant information, leading to potential overfitting. To address this, they employed a pruning technique that eliminates redundant heads, allowing the model to focus solely on the unique and informative heads. This not only curbs overfitting but also enhances inference speed by discarding repetitive computations. The model acquires the ability to determine which heads are pertinent for each specific frame through a pruning mask.

**Autoregressive Adaptive Online Prediction:** Throughout the entire interaction, the ASAP model continuously updates its memory, akin to the approach in Yang et al. [2017]. Past information is retained within memory cells and utilized for making new predictions. Moreover, the model operates in an autoregressive manner, where the predicted values of previous time steps are fed back as input for predicting subsequent time steps. This approach allows the model to adapt and refine its predictions as the interaction unfolds, capturing the dynamic nature of human-agent exchanges.

## 8.6 Proposed model

Our primary focus lies in the nonverbal behaviour of the virtual agent around an interruption - how the interruption is initiated. Consequently, we concentrated our data selection solely on the time span from 0.6 seconds before the interruption onset point to the end of the first Inter-Personal Unit (IPU) of the interruption. We employed a training dataset consisting of 17 conversations to train the model, while an additional set of four conversations was reserved for testing purposes. This test dataset was utilized for both objective and subjective evaluations of the model's performance.

This selective data range was chosen deliberately to capture the critical moments leading up to and immediately following the interruption. By isolating this specific temporal window, we aim to examine the virtual agent's nonverbal behaviours during the preparatory phase and their subsequent adaptation as the interaction dynamics shift due to the interruption. This approach enables us to closely analyze the agent's nonverbal cues during an interruption.

To refine the generation of interruption behaviour more effectively, we tested several modifications to the training approach of the ASAP model, while maintaining the same input-output structure as the original model:

- **Fine-Tuning:** In this approach, we initiated the process by freezing all layers of the ASAP model except the final fully connected layer. We then trained the model with interruption data until the loss converged. Subsequently, we unfroze the previously frozen layers and continued training with interruption data for an additional 20 epochs. By adopting this strategy, we aimed to fine-



tune the model’s learned representations and adapt them specifically to the nuances of interruption behaviour.

- **Weighted Training:** We decided to fully retrain the ASAP model from its original architecture. For this, we employed data from all moments within the 17 conversations. However, unlike the previous approach, we introduced manually assigned weights to the data based on the proportions of interruption data within the entire dataset. This weighting allowed the model to focus more on learning interruption-related behaviours, aligning its attention with the specific context of interest.
- **Input Modification:** Similar to the weighted training method, we performed a complete retraining of the ASAP model using its original architecture. However, we introduced a binary signal to the input to represent whether an interruption was ongoing or not. Given this alteration in input data, we also made necessary adjustments to the ASAP model’s parameters to optimize its training performance while considering the new input structure.
- **Combined Input and Weighted Training:** This approach merges the concepts of the weighted training and input modification methods. While introducing the binary signal to indicate interruption status, we also applied the weighted training principle. By integrating both modifications, we aimed to capture the contextual significance of interruptions while ensuring the model’s attention remained aligned with the interruption-related behaviours.
- **Interrupt-Specific Training:** In this approach, we focused solely on retraining the ASAP model using interruption data exclusively. By isolating and emphasizing the interruption instances, the model learned to generate actions tailored specifically for interruption contexts. This training strategy aimed to enhance the model’s ability to generate interruption behaviours with a heightened degree of specificity.

Each of these modified training methods was designed to enhance the model’s capability to generate interruption behaviours. Through these adaptations, we sought to tailor the model’s learned behaviours to the unique dynamics and cues associated with interruptions, thereby fostering the generation of contextually relevant and adaptive nonverbal behaviours.

## 8.7 Objective evaluation

Evaluating the sequences of SIA’s nonverbal behaviours presents a complex and ambiguous challenge. Our responses to our interlocutors vary depending on factors such as personal traits, time of day, and mood. Nonverbal behaviours, including head and body motions or facial muscle movements captured by Action Units (AUs), are naturally expressed as temporal sequences with changing values over time.

A common approach to assessing the accuracy of generated behaviour sequences involves comparing their values against the ground truth sequence at each time step, provided they share the same length. This measure is frequently utilized as a loss function in the training of neural networks. Among the measures employed are Mean Squared Error (MSE) (Ding et al. [2015], Sadoughi and Busso [2018]), Root Mean Squared Error (RMSE) (Dermouche and Pelachaud [2019]), and Average Position Error (APE) (Ahuja et al. [2019], Ahuja and Morency [2019]). Furthermore, correlation analysis has been explored by various researchers (Ding et al. [2015], Grafsgaard et al. [2018], Sadoughi and Busso [2018]). Such metrics enable the establishment of loss functions for training neural networks. For point-to-point assessment, we adopt the Root Mean Square Error (RMSE).

The quality of nonverbal behaviours can also be evaluated through the verification of their probability distribution. Methods involving log-likelihood and density comparison (Aliakbarian et al. [2021], Jonell et al. [2020], Mao et al. [2021], Sadoughi and Busso [2018]) assess the disparities between predicted and actual sequences. Woo et al. (Woo et al. [2023]) introduced the Kolmogorov-Smirnov (KS) two-sample test (Massey Jr [1951]), which quantifies quality by gauging the density probability distinction between the generated and ground truth sequences for each output dimension. The KS test calculates the difference between the distributions of generated ( $g(x)$ ) and real ( $r(x)$ ) data. Applied across all features, the KS test yields an average score.

In evaluating these six models, including the ASAP model, we employed all interruptions from the four test conversations. To ensure the cohesiveness of the generated actions during interruptions with the preceding interaction, we utilized a frame-by-frame ground truth input, spanning from the beginning of the dialogue, to predict and generate the subsequent behaviours during interruptions. However, the generated output did not feed back into subsequent predictions until an interruption occurred. Starting from 0.6 seconds before the interruption onset point, the model entered an autoregressive mode, using each generated behaviour as input for the next prediction. This continued until the first Inter-Personal Unit (IPU) of the interruption concluded, at which point the model reverted to using ground truth inputs. Evaluation specifically focused on the autoregressively generated interruption behaviour (-0.6s to the end of the first IPU) and compared it against the ground truth interruption behaviour.

Tables 8.2 and 8.3 respectively display the evaluation results for RMSE and the KS test across the six models. Both individual output and average performance are presented. Among these models, the fine-tuning approach yielded the most favourable overall results, with the lowest average RMSE and KS test scores. Considering the outcomes of both evaluations, the "Combined Input and Weighted Training" model exhibited the weakest performance among the six models.

## 8.8 Subjective evaluation

We initially aimed to design a real-time human-agent interaction system to subjectively evaluate the performance of our generation model. However, due to the inherent unpredictability and open-ended nature of dialogues, generating real-time speech content for interruptions proved to be a significant challenge. As demonstrated in the previous chapter’s experimental results, the content during interruptions can have a substantial impact on perception. Additionally, with the application of the PredIT model, we lacked control over whether participants would be interrupted during the dialogues and the frequency of interruptions, introducing uncontrollable variables when assessing action generation models.

Taking these considerations into account, we opted for a different approach. We decided to conduct a third-party evaluation study where completed interruptions were presented through video recordings. In this study, participants were asked to assess only the generated nonverbal behaviours of the interrupters. This approach allowed us to maintain control over the stimuli and ensured a more controlled and consistent evaluation process.

Building upon the results of the objective evaluation, we have made the decision to proceed with a subjective evaluation using the fine-tuning model. This model achieved the best performance outcomes in the objective assessment. While the objective evaluation provided us with quantifiable metrics and insights into the various models’ performance, it is equally crucial to gain an understanding of how human observers perceive the generated nonverbal behaviours during interruption scenarios. This subjective evaluation allows us to delve into the qualitative aspects of the behaviours, taking into account factors that objective metrics might not fully capture, such as naturalness, appropriateness, and human-like responses.

### 8.8.1 Video preparation

For the subjective evaluation, we selected 15 samples from the interruptions within the four test conversations. Each of these chosen interruptions contained a first Inter-Personal Unit (IPU) that exceeded 4 seconds. The video segments were carefully selected to encompass a period preceding the interruption’s initiation, offering participants a contextual understanding and a relatively complete semantic overview. The videos concluded at the end of the first IPU of the interruption. To maintain visual continuity, we employed ground truth data before transitioning to the model-generated interruption behaviour at the 0.6-second mark before the interruption, Figure 8.5 explains the timeline of the videos.

## 8.8. SUBJECTIVE EVALUATION

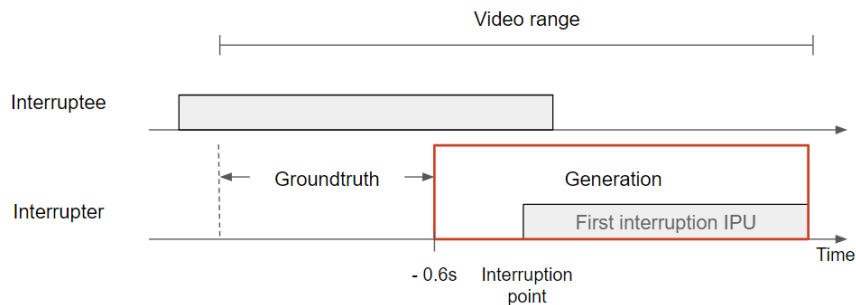


Figure 8.5 Video timeline.

To control the impact of interruption timing, audio quality, and speech content, we employed the original conversation’s interruption audio and timing. Furthermore, we utilized two virtual agent characters, one male and one female, to match the gender of the interrupter based on the conversation.

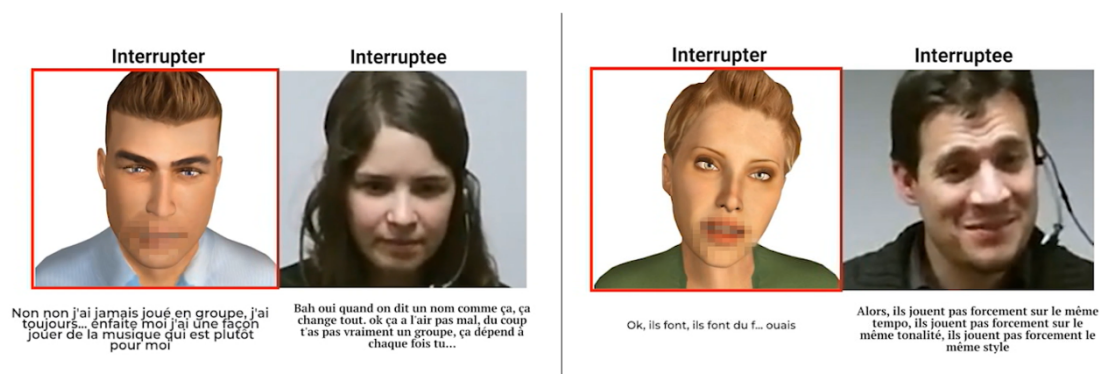


Figure 8.6 User perception test video clip examples of interactions between a male/female SIA and a human participant from NoXi database.

The Socially Interactive Agent (SIA) was animated using the open-source Greta SIA platform (Niewiadomski et al. [2009]). This was achieved by conveying visual features (predictions from computational models or ground truth) along with the audio from the ground truth. A video screenshot illustrating the setup is shown in Figure 8.6, where the SIA is displayed on the left side of the screen, and a human participant is shown on the right side.

Our model does not generate mouth movements for the agent. Since these movements are highly influenced by speech content, we opted to generate mouth animations based on the action units of the actual interrupter. However, the animations produced through action units simulation did not achieve a realistic effect. To prevent any potential visual distraction or interference caused by lip movements that might not align with the speech, we applied blurring to the mouth region during the study. This ensured that participants’ focus remained on the nonverbal behaviours being assessed without being influenced by unrealistic lip movements.

## 8.8. SUBJECTIVE EVALUATION

---

Throughout the video playback, real-time subtitles were displayed to accompany the visual content. Additionally, in order to alert participants to the initiation of interruption behaviour generation, a red border was shown on the side of the agent’s screen when the video reached 0.6 seconds before the interruption. This red border remained visible until the generated behaviour finished playing (end of the video). Participants were instructed to evaluate only the agent’s nonverbal behaviour when the red border was present. This visual cue ensured that participants focused their evaluation specifically on the interruption behaviour generation period.

### 8.8.2 Questionnaire

We generated 15 interruption videos each for the fine-tuning (best of the objective evaluation), original ASAP (as baseline), and ground truth interruption behaviour scenarios. After each video, we requested participants to evaluate the observed interruption behaviour based on their initial impressions and respond to the following questions, each rated on a 7-point Likert scale:

- Do you think the animation of the virtual character is natural?
- Do you think the animation of the virtual character is expressive?
- Do you think the interruption is competitive/cooperative?
- Do you think the virtual character is friendly?
- Do you think the virtual character is dominant?
- Do you think the virtual character is trying to energize the interaction?
- Do you think the behaviours of the virtual character and the human are in sync?

To conduct the evaluation, we divided it into three separate questionnaires, each containing 15 interruption videos representing one of the three scenarios (fine-tuning, original ASAP, and ground truth). The duration of the videos ranged from 6 to 21 seconds, and participants took an average of 15 minutes to complete each questionnaire. The evaluation was conducted using the Prolific platform, with each questionnaire being completed by 30 participants.

Comparison	natural	expressive	friendly	dominant	dynamize	in line	comp-coop
Fine tuning - ASAP	0.233	0.336	-0.047	0.011	0.111	0.031	0.067
Fine tuning - ground truth	1.107	1.102	0.340	-0.551	0.233	0.131	0.356
ASAP - ground truth	0.873	0.767	0.387	-0.562	0.122	0.1	0.289

Table 8.1 Comparison of different comparison groups for 7 questions are presented in the table. Mean differences are reported, with a positive value indicating that the first group scored higher than the second group. Significant differences between the first and second groups are highlighted in green colour ( $p < 0.05$ ).

## 8.8. SUBJECTIVE EVALUATION

	RotX	RotY	RotZ	AU1	AU2	AU4	AU5	AU6	AU7	AU12	GazeX	GazeY	Average
ASAP	1.84	1.74	1.57	0.12	0.07	0.25	0.05	0.33	0.25	0.32	1.48	2.22	0.85
Fine tuning	1.02	0.96	1.00	0.13	0.09	0.20	0.08	0.21	0.21	0.24	0.97	1.07	0.52
Weight	1.11	0.98	1.21	0.14	0.07	0.24	0.10	0.22	0.29	0.26	0.99	1.10	0.56
Input	1.34	1.32	1.19	0.22	0.17	0.26	0.05	0.22	0.21	0.25	1.32	1.40	0.66
Input & weight	1.35	1.44	1.17	0.22	0.26	0.25	0.05	0.21	0.20	0.26	1.43	1.39	0.69
Interrupt	1.09	0.95	0.98	0.11	0.07	0.22	0.05	0.21	0.20	0.26	0.98	1.19	0.52

Table 8.2 RMSE values of generated features for the six models.

	RotX	RotY	RotZ	AU1	AU2	AU4	AU5	AU6	AU7	AU12	GazeX	GazeY	Average
ASAP	0.20	0.21	0.14	0.46	0.63	0.19	0.54	0.19	0.10	0.09	0.11	0.25	0.26
Fine tuning	0.28	0.15	0.05	0.15	0.10	0.01	0.11	0.04	0.09	0.15	0.16	0.26	0.13
Weight	0.23	0.10	0.17	0.10	0.66	0.11	0.11	0.03	0.14	0.18	0.08	0.15	0.17
Input	0.20	0.10	0.10	0.20	0.24	0.11	0.63	0.06	0.02	0.23	0.12	0.13	0.18
Input & weight	0.30	0.27	0.14	0.20	0.38	0.07	0.66	0.02	0.05	0.09	0.22	0.22	0.22
Interrupt	0.26	0.15	0.09	0.74	0.66	0.03	0.66	0.05	0.09	0.16	0.15	0.24	0.27

Table 8.3 KS test values of generated features for the six models.

### 8.8.3 Subjective evaluation results

A one-way ANOVA analysis revealed significant differences among all animation conditions for all constructs: behaviour naturalness ( $F = 55$ ,  $p < 0.001$ ), behaviour expressivity ( $F = 53.1$ ,  $p < 0.001$ ), cooperative/competitive ( $F = 12.6$ ,  $p < 0.001$ ), friendliness ( $F = 10.4$ ,  $p < 0.001$ ), dominance ( $F = 46.5$ ,  $p < 0.001$ ), dynamization ( $F = 2.8$ ,  $p > 0.05$ ), and behaviour in-line ( $F = 0.9$ ,  $p > 0.05$ ).

To further analyze the differences, we conducted a post-hoc pairwise comparison using Tukey’s honestly significant difference (HSD) test. The specific results are displayed in Table 8.1 and Figure 8.7, which shows the differences between each pair of conditions for each question. According to the research findings, the fine-tuning model generated interruption behaviour with significantly higher scores in naturalness and expressivity compared to the original ASAP model and ground truth interruption behaviour. Moreover, the original ASAP model’s interruption behaviour also scored significantly higher than ground truth interruption behaviour.

In terms of friendliness, dominance, and cooperative/competitive interruptions, there were no significant differences between the fine-tuning and original ASAP models. However, compared to the ground truth, both models generated interruption behaviours that were perceived as more friendly, cooperative, and less dominant.

Regarding dynamization and behaviour in-line, the three conditions were rated with no significant difference.

## 8.9. DISCUSSION

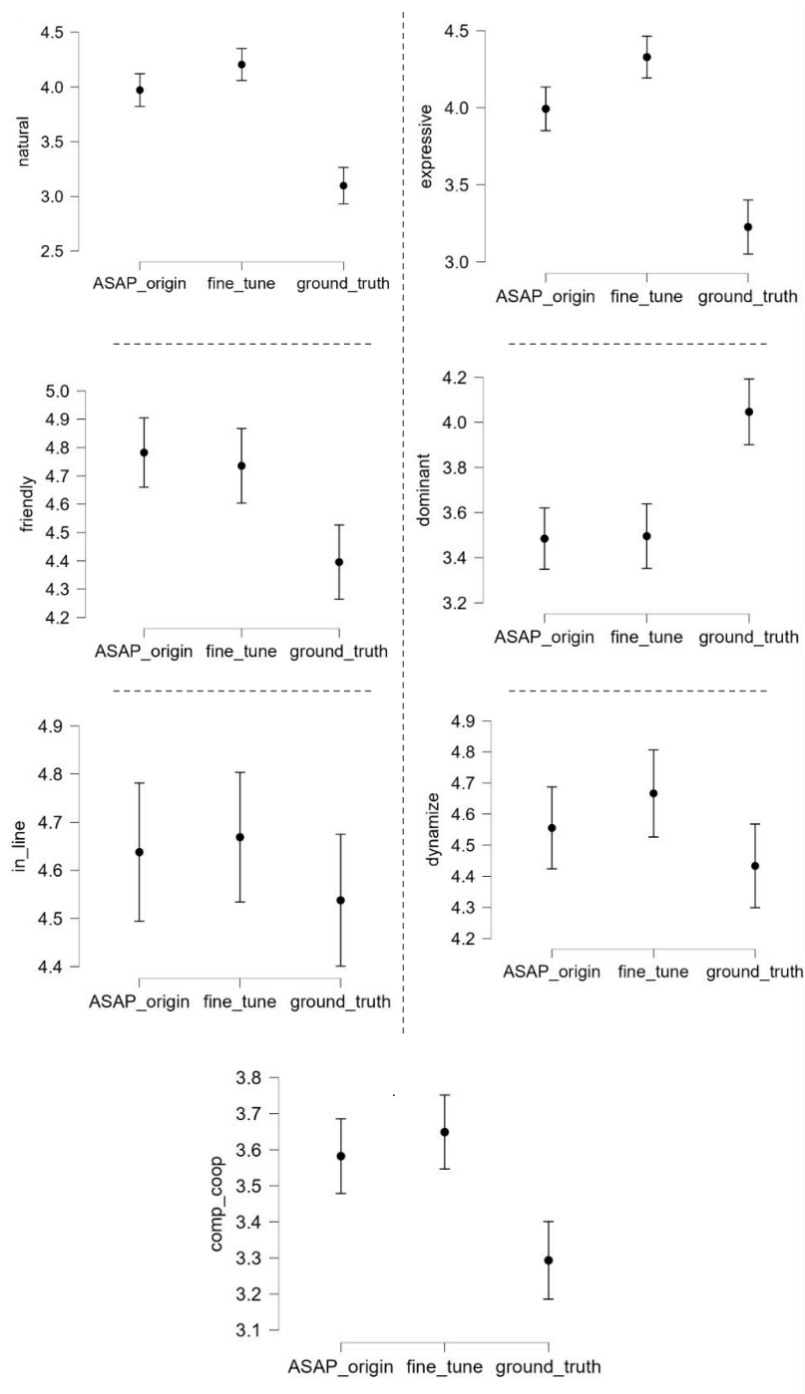


Figure 8.7 Error bar plot for seven questions rating results.

## 8.9 Discussion

Based on the evaluation results, we observe that the fine-tuned model performs better than the original ASAP model and ground truth. Compared to the model using a general interaction behaviour generation approach, the fine-tuned model



adjusted based on interruption characteristics demonstrates higher levels of naturalness and expressivity, aligning more with expectations. This finding also underscores the unique nature of interruption behaviour.

The original ASAP model, fine-tuned model, and ground truth all scored similarly in terms of dynamization and in-line behaviour, highlighting both models' capacity to adapt generation according to the partner's behaviour.

Unexpectedly, the ground truth's rating turned out to be the lowest among the three conditions, in contrast to the two generative models. Upon reviewing and comparing the videos from the three conditions, we noticed that the interruption behaviours generated by the two models exhibit greater head movements and facial expressions than those in the ground truth actions. Because the video content only includes the head of the agent and human partner, the conveyed information is limited to facial expressions and head turns, excluding gestures and body posture. Due to these limitations and the short duration of the videos, participants strive to extract information from the visible aspects to understand the events. In this context, participants might tend to assign higher scores to videos with more expressive and frequent facial expressions and actions, which is the case for the fine-tuned model.

Regarding the instances where the model results surpassed human-generated ground truth, this experiment's outcomes offer an intriguing insight. We should question whether a virtual agent must closely mimic human behaviour patterns. Our perception of humans and virtual agents differs; virtual agents lean more towards being tools. Expectations towards them vary depending on usage scenarios. Perhaps, instead of solely pursuing anthropomorphism, designing behaviour patterns that are better aligned with the virtual agent's unique traits and application contexts could yield better results. Kontogiorgos et al. (Kontogiorgos et al. [2019]) also suggest that complete anthropomorphism of virtual agents isn't always advantageous in communication.

## 8.10 Conclusion

In this chapter, we have undertaken the task of fine-tuning a pre-trained generative model using interruption data to specifically generate interruption behaviour. Through the implementation of a classification model, we effectively distinguished between the states of listening and preparing for an interruption. Our findings revealed that interrupters have an approximate preparation time of 0.6 seconds before initiating speech, during which their preparation behaviours differ from those associated with a smooth turn exchange.

We proceeded to subjectively evaluate the quality of interruption actions generated by our fine-tuned model. Our fine-tuned model achieved the highest scores, further reinforcing the unique nature of interruption behaviour. This outcome serves as a validation of our approach to fine-tune the model using interruption data, indicating its effectiveness in capturing the intricacies of interruption behaviour and generating high-quality outcomes.



## 8.10. CONCLUSION

---

This research contributes valuable insights into generating interruption behaviour and offers a promising avenue for improving the naturalness and authenticity of virtual agents' interactions, especially in scenarios involving dynamic turn-taking and interruption dynamics.

### The key points of this Chapter:

*This Chapter addresses research question Q7, Q8 and Q9*

#### *Interruption preparation behaviour*

- We generate the nonverbal behaviour of the interrupter with a preparation period of 0.6s.

#### *AI-BGM: Agent Interruption Behaviour Generation Model.*

- We fine-tuned a pre-trained model ASAP, which has been evaluated as capable of generating adaptive interaction behaviour.
- Compared to the interruption nonverbal behaviour generated by the original ASAP model and of ground truth, the behaviour generated by our model was evaluated as the most natural and expressive.

# Conclusion

## Contents

---

9.1 Summary . . . . .	116
9.2 Contributions . . . . .	117
9.3 Limitations and Future Work . . . . .	118

---

This chapter serves as the conclusion of this dissertation, encapsulating the key findings and contributions presented throughout the previous chapters. The conclusion is structured into three main sections: a summarized overview of the chapters, a discussion of the research’s contributions and implications, and a reflection on open issues and potential future research directions.

## 9.1 Summary

The primary objective of this dissertation was to bridge the gap between human conversation and human-agent interaction by comprehending interruption dynamics in human dialogues. Interruptions, commonly encountered in everyday conversations, play a pivotal role in shaping the flow of discourse. While virtual agents strive to avoid interrupting human users, a complete absence of interruptions in human-agent interactions might lead to interactions feeling rigid. On the other hand, poorly executed interruptions can result in user frustration and perceptions of system errors.

The overarching goal was to equip virtual agents with the capability to manage interruptions within interactions, especially in terms of appropriately interrupting human users without being perceived as committing a mistake, encompassing both the timing and nonverbal aspects of interruptions.

In Chapter 5, we laid the foundation by introducing an annotation schema specifically designed for manual interruption annotation. Analyzing various modes

of switching within this schema, we identified nonverbal features that can effectively characterize interruptions. This groundwork set the stage for subsequent research.

Building upon the analytical insights from Chapter 5, Chapter 6 delved into the development of a simple LSTM-based model. This model leveraged acoustic information, facial expressions, head motions, and hand movements of interrupters and interruptees to rapidly categorize interruptions into competitive and cooperative types.

In Chapter 7, a novel approach was unveiled, focusing on identifying suitable points for initiating interruptions based on the nonverbal behaviours of the current speaker. The acceptability of virtual agents interrupting human users was confirmed through subjective evaluations. However, it was noted that the pertinence of interruption content to the conversation context and the naturalness of the agent’s voice is of greater significance than the precise timing of the interruption.

Chapter 8 further highlighted the characteristics of interrupter’s nonverbal behaviour during the initiation of interruptions. It was revealed that interrupters exhibit distinct preparatory actions just before initiating interruptions. Due to the limited availability of interruption data, we employed fine-tuning techniques to specialize a pre-trained generative model, originally designed for general interaction nonverbal behaviours, to specifically generate interruption-related actions. Subsequent subjective evaluations demonstrated that the fine-tuned model generated more natural and expressive interruption behaviours compared to the original generative model.

## 9.2 Contributions

In essence, this thesis is dedicated to enhancing the interaction between Embodied Conversational Agents (ECAs) and humans by imbuing them with the capability to interrupt in an appropriate manner, thereby rendering interactions more seamless and natural. We introduced novel methodologies that empower ECAs to initiate interruptions more reasonably, while actively investigating additional factors that could render interruptions more natural and acceptable. We will now delve into a more detailed discussion of the contributions brought forth by this thesis.

### Annotation Schema

We proposed an annotation schema encompassing various conversation switches, including smooth turn exchanges, backchannels, and interruptions. This comprehensive schema facilitates the meticulous labelling of diverse switches that occur within conversations and further categorizes interruptions based on their completion and underlying motivations. Through this schema, we annotated two datasets: NoXi, representative of dyadic interactions, and a part of AMI, represen-

tative of multiparty interactions. This underscores the versatility of the schema, unhampered by the number of participants or interaction modes.

#### **Multimodal Analysis of Interruption**

In interactions, human communication is orchestrated through multiple modalities. By analyzing nonverbal behaviour features such as head movements, facial expressions, and intonation exhibited by both speakers and listeners, we discerned distinct factors distinguishing interruptions, smooth turn exchanges, and backchannels in human conversations. These identified factors not only prepared the ground for subsequent interruption generation but also illuminated pivotal considerations for designing virtual agent interruptions.

#### **Interruption Generation: Timing & Nonverbal behaviour**

Given the dynamic nature of dialogues, it's nearly impossible to determine all opportune and inopportune moments for initiating interruptions. To address this, we proposed employing a one-class Support Vector Machine (SVM) to overcome the challenge of limited negative samples. With real-time interactive applications in mind, our model relied solely on the speaker's nonverbal behaviour to determine whether an interruption could be initiated in the next instance. Subjective evaluations confirmed the acceptability of virtual agents interrupting human users. However, besides the timing of interruptions, factors such as the interrupter's voice quality, the content of the conversation, and the type of interruption exert notable influences.

During interruptions, we employed a fine-tuned nonverbal behaviour generation model to create virtual agent facial expressions and head rotations. Remarkably, our model's generated nonverbal behaviours received superior evaluations compared to those exhibited by actual interrupters.

### **9.3 Limitations and Future Work**

This thesis presents a multifaceted exploration aimed at refining the interaction dynamics between ECAs and humans. By introducing innovative strategies, we aim to enable ECAs to participate in interruptions naturally and appropriately, a feat that not only enriches interactions but also contributes to the distinctiveness of interruption behaviour. While we have made some strides in these core challenges, there are still some limitations that we will highlight in this section.

Presently, our research is rooted in nonverbal behaviour analysis. While nonverbal behaviour indeed plays a significant role in interactions, contextual analysis of dialogues is equally indispensable. For instance, when determining the timing of interruptions, incorporating an understanding of the dialogue's content could enhance accuracy. Moreover, our current capabilities are focused on predicting interruption timings and generating corresponding nonverbal behaviours. However,

as previously mentioned, the content of interruptions significantly influences their perception. While nonverbal behaviours aid in identifying suitable moments for initiating interruptions, the timing is closely tied to the content.

Furthermore, the process of annotating interruptions in conversations requires manual intervention, and even with a substantial amount of dialogue data, there is no guarantee of obtaining a sufficient quantity of interruptive instances to train a robust model. This significantly escalates the cost of research and imposes limitations based on the conversational scenario. To mitigate the issue of limited data and the risk of overfitting, we are constrained to employ relatively simplistic models, but this comes at the cost of achieving optimal performance. Also, since the notion of virtual agents interrupting human users is a relatively novel research direction, there is a scarcity of studies in this area. We employed simple models and a limited range of modalities, neglecting aspects like body posture that can convey valuable information. We firmly believe that incorporating more comprehensive and diverse information could optimize research outcomes.

During the process of subjective evaluation, we opted for presenting animations of ECA interruptions through videos, rather than engaging in real-time interactions with the ECA. There were two primary reasons for this approach. Firstly, the constraints of time prevented us from conducting real-time interaction experiments within the thesis timeline. Secondly, in real-time interactions, ensuring consistent and extended participation of human users in the role of the speaker is challenging, leading to increased variability that could affect experimental results. This complexity poses a significant challenge to the overall experiment design. Furthermore, the experience of observing interactions may differ from engaging in real interactions, potentially yielding distinct experimental outcomes. To obtain more precise experimental results, it is still necessary to design a real-time human-agent interaction system for testing interruptions. This system would allow us to evaluate interruptions in a controlled and real-world scenario, providing a more accurate assessment of their impact and effectiveness.

In the future, with a sufficient amount of data, it will be possible to train interruption models based on the personalities and roles of the characters involved. This approach could imbue Embodied Conversational Agents (ECAs) with distinct personality traits, making their interruptions more contextually appropriate in various conversation scenarios. Furthermore, beyond knowing when and how to interrupt, ECAs should also be equipped to make decisions about whether to abandon an interruption midway, how to do so gracefully, and when to initiate this abandonment. These additional layers of decision-making will contribute to more nuanced and effective interactions in the realm of human-agent communication.

At the same time, it's important to note that our research primarily focuses on predicting potential interruption points based on the current or a few preceding turns. However, real-life human conversations largely rely on the entire context of the conversation rather than just the immediate past. Interruptions should also consider alignment with the overall conversation context and their necessity. Not every potential interruption point warrants an interruption; conversely, when in-

### 9.3. LIMITATIONS AND FUTURE WORK

---

Interruptions are necessary, it makes more sense to identify appropriate interruption timing and employ suitable interruption methods based on multimodal behaviour.

In practical terms, this means that effective interruptions should take into account the broader conversation flow and relevance, ensuring that they enhance, rather than disrupt, the overall discourse. It's not merely about identifying opportunities to interrupt but also about understanding when interruptions are warranted and how they can be seamlessly integrated into the ongoing dialogue.

# Bibliography

- Facial Expression Recognition (Face Recognition Techniques) Part 1.  
<http://what-when-how.com/face-recognition/facial-expression-recognition-face-recognition-techniques-part-1/>.
- Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019.
- Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International conference on multimodal interaction*, pages 74–84, 2019.
- Samer Al Moubayed, Jonas Beskow, and Gabriel Skantze. The furhat social companion talking head. In *INTERSPEECH*, pages 747–749, 2013.
- Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, volume 39, pages 487–496. Wiley Online Library, 2020.
- Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11333–11342, 2021.
- Jens Allwood. Dialog coding-function and grammar: Göteborg coding schemas. *rapport nr.: Gothenburg Papers in Theoretical Linguistics 85*, 2001.
- Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26, 1992.
- Sean Andrist, Iolanda Leite, and Jill Lehman. Fun and fair: influencing turn-taking in a multi-party game with a virtual agent. In *Proceedings of the 12th international conference on interaction design and children*, pages 352–355, 2013.
- Michael Argyle. *Bodily communication*. Routledge, 2013.
- Michael Argyle, Mansur Lalljee, and Mark Cook. The effects of visibility on interaction in a dyad. *Human relations*, 21(1):3–17, 1968.

## BIBLIOGRAPHY

---

- Peter Auer. Gaze, addressee selection and turn-taking in three-party interaction. *Eye-tracking in interaction: Studies on the role of eye gaze in dialogue*, pages 197–231, 2018.
- Ronald J Baken and Robert F Orlikoff. *Clinical measurement of speech and voice*. Cengage Learning, 2000.
- Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016a.
- Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016b.
- Timo Baumann and David Schlangen. Predicting the micro-timing of user input for an incremental spoken dialogue system that completes a user’s ongoing turn. In *Proceedings of the SIGDIAL 2011 Conference*, pages 120–129, 2011.
- Tobias Baur, Alexander Heimerl, Florian Lingens, Johannes Wagner, Michel F. Valstar, Björn Schuller, and Elisabeth André. explainable cooperative machine learning with NOVA. *KI - Künstliche Intelligenz*, Jan 2020. ISSN 1610-1987. doi: 10.1007/s13218-020-00632-3. URL <https://doi.org/10.1007/s13218-020-00632-3>.
- Geoffrey W Beattie. *Interruption in conversational interaction, and its relation to the sex and status of the interactants*. Walter de Gruyter, Berlin/New York Berlin, New York, 1981.
- Geoffrey W Beattie. Turn-taking and interruption in political interviews: Margaret thatcher and jim callaghan compared and contrasted. 1982.
- Linda Bell, Johan Boye, and Joakim Gustafson. Real-time handling of fragmented utterances. In *Proc. NAACL workshop on adaptation in dialogue systems*, pages 2–8, 2001.
- Adrian Bennett. Interruptions and the interpretation of conversation. In *Annual Meeting of the Berkeley Linguistics Society*, volume 4, pages 557–575, 1978.
- Casey C Bennett, Young-Ho Bae, Jun Hyung Yoon, Yejin Chae, Eunseo Yoon, Seeun Lee, Uijae Ryu, Say Young Kim, and Benjamin Weiss. Effects of cross-cultural language differences on social cognition during human-agent interaction in co-operative game environments. *Computer Speech & Language*, 81:101521, 2023.
- Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pages 1–10. IEEE, 2021.
- RL Birdwhistell. The language of the body: The natural environment of words. hillsday: Ed. A. *Silverstein Hillsdale*, 1974.
- Barry Bogin and Carlos Varea. Evolution of human life history. In *Evolutionary Neuroscience*, pages 753–767. Elsevier, 2020.



## BIBLIOGRAPHY

---

- Galina B Bolden, Jenny Mandelbaum, and Sue Wilkinson. Pursuing a response by repairing an indexical reference. *Research on Language & Social Interaction*, 45(2):137–155, 2012.
- Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. A multimodal model for predicting conversational feedbacks. In *International conference on text, speech, and dialogue*, pages 537–549. Springer, 2021.
- Sheryl Brahnham, Chao-Fa Chuang, Randall S Sexton, and Frank Y Shih. Machine assessment of neonatal facial expressions of acute pain. *Decision Support Systems*, 43(4):1242–1254, 2007.
- Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics*, 28(3):449–483, 2017.
- Judee K Burgoon, Valerie Manusov, and Laura K Guerrero. *Nonverbal communication*. Routledge, 2021.
- Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multi-agent Systems*, pages 911–920, 2016.
- Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359, 2017.
- Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.
- Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, Hao Yan, et al. Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents*, pages 29–63, 2000.
- Justine Cassell, Tim Bickmore, Lee Campbell, Hannes Vilhjalmsson, and Hao Yan. More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-based systems*, 14(1-2):55–64, 2001.
- Justine Cassell, Alastair Gill, and Paul Tepper. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*, pages 41–50, 2007.
- Christian Cavé, Isabelle Guaitella, Roxane Bertrand, Serge Santi, Françoise Harlay, and Robert Espesser. About the relationship between eyebrow movements and fo variations. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP’96*, volume 4, pages 2175–2178. IEEE, 1996.

## BIBLIOGRAPHY

---

- Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–26, 2012.
- Mathieu Chollet, Magalie Ochs, and Catherine Pelachaud. A methodology for the automatic extraction and generation of non-verbal signals sequences conveying interpersonal attitudes. *IEEE Trans. Affect. Comput.*, 10(4):585–598, 2019.
- Noam Chomsky. *Aspects of the Theory of Syntax*. Number 11. MIT press, 2014.
- Nicole Chovil. Discourse-oriented facial displays in conversation. *Research on Language & Social Interaction*, 25(1-4):163–194, 1991.
- Shammur Absar Chowdhury and Giuseppe Riccardi. A deep learning approach to modeling competitiveness in spoken conversations. pages 5680–5684, 2017.
- Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi. Annotating and categorizing competition in overlap speech. pages 5316–5320, 2015.
- Shammur Absar Chowdhury, Evgeny A Stepanov, Morena Danieli, and Giuseppe Riccardi. Automatic classification of speech overlaps: feature representation and algorithms. *Computer Speech & Language*, 55:145–167, 2019.
- George Christodoulides and Mathieu Avanzi. Automatic detection and annotation of disfluencies in spoken french corpora. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- Adam Chýlek, Jan Švec, and Luboš Šmídl. Learning to interrupt the user at the right time in incremental dialogue systems. In *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 21*, pages 500–508. Springer, 2018.
- Herbert H Clark. *Using language*. Cambridge university press, 1996.
- Mark Cook and Mansur G Lalljee. Verbal substitutes for visual signals in interaction. 1972.
- Ruth E Corps, Martin J Pickering, and Chiara Gambi. Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience*, 34(5):615–627, 2019.
- Nigel Crook, Cameron Smith, Marc Cavazza, Stephen Pulman, Roger Moore, and Johan Boye. Handling user interruptions in an embodied conversational agent. In *Proceedings of the AAMAS International Workshop on Interacting with ECAs as Virtual Characters*, pages 27–33, 2010.
- Nigel Crook, Debora Field, Cameron Smith, Sue Harding, Stephen Pulman, Marc Cavazza, Daniel Charlton, Roger Moore, and Johan Boye. Generating context-sensitive eca responses to user barge-in interruptions. *Journal on Multimodal User Interfaces*, 6:13–25, 2012.
- Katharina Cyra and Karola Pitsch. Dealing with long utterances: How to interrupt the user in a socially acceptable manner? In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 341–345, 2017.

## BIBLIOGRAPHY

---

- Charles Darwin. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.
- Judy Davidson. *Subsequent versions of invitations, offers, requests, and proposals dealing with potential or actual rejection*, page 102–128. Cambridge University Press, 1985. doi: 10.1017/CBO9780511665868.009.
- Jim Davies. Program good ethics into artificial intelligence. *Nature*, 2016.
- Mike Demol, Werner Verhelst, and Piet Verhoeve. The duration of speech pauses in a multilingual environment. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.
- Soumia Dermouche and Catherine Pelachaud. Generative model of agent’s behaviors in human-agent interaction. In *2019 International Conference on Multimodal Interaction*, pages 375–384, 2019.
- David DeVault, Kenji Sagae, and David Traum. Can i finish? learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the SIGDIAL 2009 Conference*, pages 11–20, 2009.
- Chuang Ding, Lei Xie, and Pengcheng Zhu. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, 74:9871–9888, 2015.
- Mark Dingemanse and Simeon Floyd. Conversation across cultures. In *The Cambridge handbook of linguistic anthropology*, pages 447–480. Cambridge University Press, 2014.
- Allen T Dittmann and Lynn G Llewellyn. The phonemic clause as a unit of speech decoding. *Journal of personality and social psychology*, 6(3):341, 1967.
- Paul Drew. Quit talking while i’m interrupting: a comparison between positions of overlap onset in conversation. *Talk in interaction: Comparative dimensions*, pages 70–93, 2009.
- Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.
- Jens Edlund and Mattias Heldner. Exploring prosody in interaction control. *Phonetica*, 62(2-4):215–226, 2005.
- Olga Egorow and Andreas Wendemuth. On emotions as features for speech overlaps classification. *IEEE Transactions on Affective Computing*, 2019.
- Paul Ekman. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):63–69, 1992.
- Paul Ekman. Emotional and conversational nonverbal signals. In *Language, knowledge, and representation*, pages 39–50. Springer, 2004.
- Paul Ekman and Wallace V Friesen. Felt, false, and miserable smiles. *Journal of nonverbal behavior*, 6(4):238–252, 1982.

## BIBLIOGRAPHY

---

- Paul Ekman, Wallace V Friesen, and Silvan S Tomkins. Facial affect scoring technique: A first validity study. *Walter de Gruyter*, 1971.
- Mika Enomoto, Yasuharu Den, and Yuichi Ishimoto. A conversation-analytic annotation of turn-taking behavior in japanese multi-party conversation and its preliminary analysis. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 644–652, 2020.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.
- Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838, 2013.
- Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- Mireille Fares, Catherine Pelachaud, and Nicolas Obin. Transformer network for semantically-aware and speech-driven upper-face generation. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 593–597. IEEE, 2022.
- Will Feng, Anitha Kannan, Georgia Gkioxari, and C Lawrence Zitnick. Learn2smile: Learning non-verbal interaction through observation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4131–4138. IEEE, 2017.
- Nicola Ferguson. Simultaneous speech, interruptions and dominance. *British Journal of social and clinical Psychology*, 16(4):295–302, 1977.
- Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–10, 2019.
- Kerstin Fischer, Lakshadeep Naik, Rosalyn M Langedijk, Timo Baumann, Matouš Jelínek, and Oskar Palinko. Initiating human-robot interactions using incremental speech adaptation. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 421–425, 2021.
- Cecilia E Ford and Sandra A Thompson. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184, 1996.
- Peter French and John Local. Turn-competitive incomings. *Journal of Pragmatics*, 7(1):17–38, 1983.
- Peter French and John Local. Prosodic features and the management of interruptions 1. In *Intonation in discourse*, pages 157–180. Routledge, 2018.

## BIBLIOGRAPHY

---

- Harold Garfinkel. Studies of the routine grounds of everyday activities. *Social problems*, 11(3):225–250, 1964.
- Harold Garfinkel. Respecification: Evidence for locally produced, naturally accountable phenomena of order, logic, reason, meaning, method, etc. in and as of the essential haecceity of immortal ordinary society (i)—an announcement of studies. *Ethnomethodology and the human sciences*, pages 10–19, 1991.
- Harold Garfinkel. Studies in ethnomethodology. In *Social Theory Re-Wired*, pages 58–66. Routledge, 2023.
- Simon Garrod and Martin J Pickering. The use of content and timing to predict turn transitions. *Frontiers in psychology*, 6(751):1–12, 2015.
- Patrick Gebhard, Tanja Schneeberger, Gregor Mehlmann, Tobias Baur, and Elisabeth André. Designing the impression of social agents’ real-time interruption handling. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 19–21, 2019.
- Shazia Akbar Ghilzai and Mahvish Baloch. Conversational analysis of turn taking behavior and gender differences in multimodal conversation. *European Academic Research*, 3(9):10100–10116, 2015.
- Howard Giles. *Communication accommodation theory: Negotiating personal relationships and social identities across contexts*. Cambridge University Press, 2016.
- Erving Goffman. The neglected situation. *American anthropologist*, 66(6):133–136, 1964.
- Julia A Goldberg. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts. *Journal of Pragmatics*, 14(6):883–903, 1990.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Charles Goodwin. The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, 97:101–121, 1979.
- Charles Goodwin. Conversational organization. *Interaction between speakers and hearers*, 1981.
- Charles Goodwin and John Heritage. Conversation analysis. *Annual review of anthropology*, 19(1):283–307, 1990.
- Joseph Grafsgaard, Nicholas Duran, Ashley Randall, Chun Tao, and Sidney D’Mello. Generative multimodal models of nonverbal synchrony in close relationships. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 195–202. IEEE, 2018.
- Agustín Gravano and Julia Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634, 2011.

## BIBLIOGRAPHY

---

- Agustín Gravano and Julia Hirschberg. A corpus-based study of interruptions in spoken dialogue. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- David Greenwood, Stephen Laycock, and Iain Matthews. Predicting head pose from speech with a conditional variational autoencoder. ISCA, 2017.
- P EKMAN-WV FRIESEN-JC HAGER. Facial action coding system. the manual on cd rom, 2002.
- Britta Hammarberg, Bernard Fritzell, J Gaufin, Johan Sundberg, and Lage Wedin. Perceptual and acoustic correlates of abnormal voice qualities. *Acta otolaryngologica*, 90(1-6):441–451, 1980.
- Jinni A Harrigan. Listeners’body movements and speaking turns. *Communication Research*, 12(2):233–250, 1985.
- Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 79–86, 2018.
- Peter A Heeman, Andy McMillin, and J Scott Yaruss. An annotation scheme for complex disfluencies. In *INTERSPEECH*. Citeseer, 2006.
- Rebecca Heins, Marita Franzke, Michael Durian, and Aruna Bayya. Turn-taking as a design principle for barge-in in spoken language systems. *International Journal of Speech Technology*, 2:155–164, 1997.
- Léo Hemamou, Ghazi Felhi, Jean-Claude Martin, and Chloé Clavel. Slices of attention in asynchronous video job interviews. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.
- Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.
- John Heritage. Current developments in conversation analysis. *D. Roger & P. Bull (Eds.), Conversation: An interdisciplinary perspective*, pages 21–47, 1989.
- John Heritage. *Garfinkel and ethnomethodology*. John Wiley & Sons, 2013.
- Katherine Hilton. *What Does an Interruption Sound Like?* Stanford University, 2018.
- Anna Hjalmarsson. The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35, 2011.
- Anna Hjalmarsson and Catharine Oertel. Gaze direction as a back-channel inviting cue in dialogue. In *IVA 2012 workshop on realtime conversational virtual agents*, volume 9, 2012.
- Elliott M Hoey. *When conversation lapses: The public accountability of silent copresence*. Oxford University Press, 2020.

## BIBLIOGRAPHY

---

- Judith Holler and Stephen C Levinson. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652, 2019.
- Judith Holler, Kobin H Kendrick, and Stephen C Levinson. Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic bulletin & review*, 25:1900–1908, 2018.
- Nusrah Hussain, Engin Erzin, T Metin Sezgin, and Yucel Yemez. Speech driven backchannel generation using deep q-network for enhancing engagement in human-robot interaction. *arXiv preprint arXiv:1908.01618*, 2019.
- Ian Hutchby. Power in discourse: The case of arguments on a british talk radio show. *Discourse & Society*, 7(4):481–497, 1996.
- Peter Indefrey and Willem JM Levelt. The spatial and temporal signatures of word production components. *Cognition*, 92(1-2):101–144, 2004.
- Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Analysis of respiration for prediction of "who will be next speaker and when?" in multi-party meetings. In *Proceedings of the 16th international conference on multimodal interaction*, pages 18–25, 2014.
- Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 6(1):1–31, 2016.
- Joseph Jaffe, Beatrice Beebe, Stanley Feldstein, Cynthia L Crown, Michael D Jasnou, Philippe Rochat, and Daniel N Stern. Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the society for research in child development*, pages i–149, 2001.
- Kathrin Janowski and Elisabeth André. What if i speak now? a decision-theoretic approach to personality-based turn-taking. 2019.
- Gail Jefferson. A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. 1973.
- Gail Jefferson. Notes on some orderlinesses of overlap onset. *Discourse analysis and natural rhetoric*, 500:11–38, 1984.
- Gail Jefferson. Notes on 'latency' in overlap onset. *Human Studies*, pages 153–183, 1986.
- Gail Jefferson. A sketch of some orderly aspects of overlap in natural conversation. *Pragmatics and beyond new series*, 125:43–62, 2004.
- Gail Jefferson and Emanuel Schegloff. A sketch of some orderly aspects of overlap in natural conversation. *Conversation Analysis*, pages 43–59, 1975.
- Kristiina Jokinen, Masafumi Nishida, and Seiichi Yamamoto. On eye-gaze and turn-taking. In *Proceedings of the 2010 workshop on eye gaze in intelligent human machine interaction*, pages 118–123, 2010.

## BIBLIOGRAPHY

---

- Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–30, 2013.
- Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let’s face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.
- Gudny Ragna Jonsdottir, Kristinn R Thorisson, and Eric Nivel. Learning smooth, human-like turntaking in realtime dialogue. In *Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008. Proceedings 8*, pages 162–175. Springer, 2008.
- Reshmashree Bangalore Kantharaju, Caroline Langlet, Mukesh Barange, Chloé Clavel, and Catherine Pelachaud. Multimodal analysis of cohesion in multi-party interactions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 498–507, 2020.
- Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.
- Adam Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63, 1967.
- Kobin H. Kendrick and Francisco Torreira. The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4):255–289, 2015. doi: 10.1080/0163853X.2014.955997.
- Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. Optimising turn-taking strategies with reinforcement learning. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 315–324, 2015.
- Thomas Kiderle, Hannes Ritschel, Kathrin Janowski, Silvan Mertes, Florian Lingenfelter, and Elisabeth André. Socially-aware personality adaptation. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2021.
- Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Gengage Learning, 2013.
- Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, 41(3-4):295–321, 1998.
- Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 133–140, 2019.



## BIBLIOGRAPHY

---

- Ladislav Kunc, Zdenek Míkovec, and Pavel Slavík. Avatar and dialog turn-yielding phenomena. *International Journal of Technology and Human Interaction (IJTHI)*, 9(2):66–88, 2013.
- Emina Kurtic, Guy J Brown, and Bill Wells. Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration. pages 2550–2553, 2010.
- Emina Kurtić, Guy J Brown, and Bill Wells. Resources for turn competition in overlapping talk. *Speech Communication*, 55(5):721–743, 2013.
- Debi Laplante and Nalini Ambady. On how things are said: Voice tone, voice intensity, verbal content, and perceptions of politeness. *Journal of language and social psychology*, 22(4):434–441, 2003.
- Chi-Chun Lee and Shrikanth Narayanan. Predicting interruptions in dyadic spoken interactions. pages 5250–5253, 2010.
- Chi-Chun Lee, Sungbok Lee, and Shrikanth S Narayanan. An analysis of multi-modal cues of interruption in dyadic spoken interactions. 2008.
- Gene H Lerner. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in society*, 32(2):177–201, 2003.
- Gene H Lerner. Conversation analysis. *Conversation Analysis*, pages 1–312, 2004.
- Gene H Lerner. When someone other than the addressed recipient speaks next: Three kinds of intervening action after the selection of next speaker. *Research on Language and Social Interaction*, 52(4):388–405, 2019.
- Stephen C Levinson. Turn-taking in human communication—origins and implications for language processing. *Trends in cognitive sciences*, 20(1):6–14, 2016.
- Stephen C Levinson and Judith Holler. The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 2014.
- Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 2015.
- Rivka Levitan and Julia Bell Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech*, 2011.
- Han Z Li. Cooperative and intrusive interruptions in inter-and intracultural dyadic discourse. *Journal of language and social psychology*, 20(3):259–284, 2001.
- John K Local, John Kelly, and William HG Wells. Towards a phonology of conversation: turn-taking in tyneside english1. *Journal of Linguistics*, 22(2):411–437, 1986.
- Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, page 11. Plymouth, MA, 2000.

## BIBLIOGRAPHY

---

- Birgit Lugrin. Introduction to socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pages 1–20. 2021.
- Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021.
- Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.
- Douglas W Maynard and Steven E Clayman. Ethnomethodology and conversation analysis. *Handbook of symbolic interactionism*, pages 173–202, 2003.
- David H McFarland. Respiratory markers of conversational interaction. 2001.
- Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech & Language*, 28(4):903–922, 2014.
- Michael Moerman and Harvey Sacks. On “understanding” in the analysis of natural conversation. *Talking culture: Ethnography and conversation analysis*, pages 180–186, 1988.
- Michael Moerman and Harvey Sacks. Appendix b. on “understanding” in the analysis of natural conversation. In *Talking Culture*, pages 180–186. University of Pennsylvania Press, 2010.
- Lorenza Mondada and Florence Oloff. Gestures in overlap. the situated establishment of speakership. In *Integrating gestures. The interdisciplinary nature of gesture*, pages 321–338. Benjamins, 2011.
- Michael T Motley. Facial affect and verbal context in conversation: Facial expression as interjection. *Human Communication Research*, 20(1):3–40, 1993.
- Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. Multimediate: Multi-modal group behaviour analysis for artificial mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4878–4882, 2021.
- Shu Nakazato. Japanese dialogue corpus of multi-level annotation. In *1st SIGdial Workshop on Discourse and Dialogue*, pages 1–8, 2000.
- Daniel Neiberg and Khiet P Truong. Online detection of vocal listener responses with maximum latency constraints. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5836–5839. IEEE, 2011.
- Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400, 2009.

## BIBLIOGRAPHY

---

- AC Norwine and Oliver J Murphy. Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, 17(2):281–291, 1938.
- Catharine Oertel, Marcin Włodarczak, Jens Edlund, Petra Wagner, and Joakim Gustafson. Gaze patterns in turn-taking. In *Thirteenth annual conference of the international speech communication association*, 2012.
- Catherine Pelachaud. Greta: an interactive expressive embodied conversational agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 5–5, 2015.
- Catherine Pelachaud, Norman I Badler, and Mark Steedman. Generating facial expressions for speech. *Cognitive science*, 20(1):1–46, 1996.
- Janet Pierrehumbert. The meaning of intonational contours in the interpretation of discourse janet pierrehumbert and julia hirschberg. *Intentions in communication*, 271, 1990.
- Anita Pomerantz. *Pursuing a response*. Cambridge University Press Cambridge, UK, 1983.
- Antoine Raux and Maxine Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 629–637, 2009.
- Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. Doing research on a deployed spoken dialogue system: One year of let’s go! experience. In *Ninth International Conference on Spoken Language Processing*, 2006.
- Brian Ravenet, Angelo Cafaro, Beatrice Biancardi, Magalie Ochs, and Catherine Pelachaud. Conversational behavior reflecting interpersonal attitudes in small group interactions. In *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15*, pages 375–388. Springer, 2015.
- Carina Riest, Annett B Jorschick, and Jan P de Ruiter. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:89, 2015.
- Y Rim. Personality variables and interruptions in small discussions. *European Journal of Social Psychology*, 1977.
- Jeffrey D Robinson. Revisiting preference organization in context: A qualitative and quantitative examination of responses to information seeking. *Research on Language and Social Interaction*, 53(2):197–222, 2020.
- Amélie Rochet-Capellan and Susanne Fuchs. Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658):20130399, 2014.
- Matthew Roddy, Gabriel Skantze, and Naomi Harte. Investigating speech features for continuous turn-taking prediction using lstms. *arXiv preprint arXiv:1806.11461*, 2018a.

## BIBLIOGRAPHY

---

- Matthew Roddy, Gabriel Skantze, and Naomi Harte. Multimodal continuous turn-taking prediction using multiscale rnns. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 186–190, 2018b.
- Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced social interaction with agents: 8th international workshop on spoken dialog systems*, pages 247–258. Springer, 2019.
- Zsófia Ruttkay and Catherine Pelachaud. *From brows to trust: Evaluating embodied conversational agents*, volume 7. Springer Science & Business Media, 2004.
- Harvey Sacks. Lectures on conversation: Volume i. *Malden, Massachusetts: Blackwell*, 1992.
- Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.
- Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6169–6173. IEEE, 2018.
- Putu Pande Novita Sari, Ni Luh Putu Sri Adnyani, and I Made Suta Paramarta. Conversation analysis: Turn taking on indonesia lawyer club talk show. *Lingua Scientia*, 28(1):47–57, 2021.
- Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyooki Aikawa. Learning decision trees to determine turn-taking by spoken dialogue systems. In *INTERSPEECH*, 2002.
- Emanuel A Schegloff. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology*, 97(5):1295–1345, 1992.
- Emanuel A Schegloff. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63, 2000.
- Emanuel A Schegloff. Accounts of conduct in interaction: Interruption, overlap, and turn-taking. In *Handbook of sociological theory*, pages 287–321. Springer, 2001.
- Emanuel A Schegloff. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press, 2007.
- Emanuel A Schegloff and Harvey Sacks. *Opening up closings*. Walter de Gruyter, Berlin/New York Berlin, New York, 1973.
- Dominik Schiller, Katharina Weitz, Kathrin Janowski, and Elisabeth André. Human-inspired socially-aware interfaces. In *International Conference on Theory and Practice of Natural Computing*, pages 41–53. Springer, 2019.

## BIBLIOGRAPHY

---

- David Schlangen. From reaction to prediction: Experiments with computational models of turn-taking. *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking*, 2006.
- David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111, 2011.
- Ethan Selfridge and Peter A Heeman. Importance-driven turn-bidding for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 177–185, 2010.
- Ethan Selfridge, Iker Arizmendi, Peter A Heeman, and Jason D Williams. Continuously predicting and processing barge-in during a live spoken dialogue task. In *Proceedings of the SIGDIAL 2013 Conference*, pages 384–393, 2013.
- Margret Selting. On the interplay of syntax and prosody in the constitution of turn-constructive units and turns in conversation. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 6(3):371–388, 1996.
- Syamimi Shamsuddin, Luthffi Idzhar Ismail, Hanafiah Yussof, Nur Ismarrubie Zahari, Saiful Bahari, Hafizan Hashim, and Ahmed Jaffar. Humanoid robot nao: Review of control and motion exploration. In *2011 IEEE international conference on Control System, Computing and Engineering*, pages 511–516. IEEE, 2011.
- Elizabeth Shriberg, Andreas Stolcke, and Don Baron. Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech. 2001a.
- Elizabeth Shriberg, Andreas Stolcke, and Don Baron. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Seventh European Conference on Speech Communication and Technology*, 2001b.
- Rein Ove Sikveland and Richard Ogden. Holding gestures across turns: moments to generate shared understanding. *Gesture*, 12(2):166–199, 2012.
- Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, 2017.
- Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178, 2021.
- Gabriel Skantze and Anna Hjalmarsson. Towards incremental speech generation in dialogue systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 1–8, 2010.
- Gabriel Skantze and David Schlangen. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 745–753, 2009.
- Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, 65:50–66, 2014.

## BIBLIOGRAPHY

---

- Cameron Smith, Nigel Crook, Daniel Charlton, Johan Boye, Raul Santos De La Camara, Markku Turunen, David Benyon, Björn Gambäck, Oli Mival, Nick Webb, et al. Interaction strategies for an affective conversational agent. *Presence*, 20(5):395–411, 2011.
- Tanya Stivers and Federico Rossano. Mobilizing response. *Research on Language and social interaction*, 43(1):3–31, 2010.
- Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.
- Nikko Ström and Stephanie Seneff. Intelligent barge-in in conversational systems. In *INTERSPEECH*, pages 652–655, 2000.
- Deborah Tannen. *Gender and discourse*, new york. NY: Oxford University, 1994.
- Deborah Tannen et al. *You just don't understand: Women and men in conversation*. Virago London, 1991.
- Louis Ten Bosch, Nelleke Oostdijk, and Jan Peter De Ruiter. Turn-taking in social talk dialogues: temporal, formal and functional aspects. In *9th International Conference Speech and Computer (SPECOM'2004)*, pages 454–461, 2004.
- Mark Ter Maat and Dirk Heylen. Turn management or impression management? In *International Workshop on Intelligent Virtual Agents*, pages 467–473. Springer, 2009.
- Mark Ter Maat, Khiet P Truong, and Dirk Heylen. How turn-taking strategies influence users' impressions of an agent. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 441–453. Springer, 2010.
- Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2020.
- Kristinn R Thórisson. Real-time decision making in multimodal face-to-face communication. In *Proceedings of the second international conference on Autonomous agents*, pages 16–23, 1998.
- Kristinn R Thórisson. Mind model for multimodal communicative creatures and humanoids. *Applied Artificial Intelligence*, 13(4-5):449–486, 1999.
- IR Titze. *Principles of Voice Production*. Prentice-Hall Inc., 1994.
- Michael Tomasello, Brian Hare, Hagen Lehmann, and Josep Call. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of human evolution*, 52(3):314–320, 2007.
- FJ Torreira, Sara Bögels, and Stephen C Levinson. Breathing for answering. the time course of response planning in conversation. 2016.

## BIBLIOGRAPHY

---

- Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.
- Naomi Truan and Laurent Romary. Building, encoding, and annotating a corpus of parliamentary debates in xml-tei: A cross-linguistic account. *Journal of the Text Encoding Initiative*, 2021.
- Khiet P Truong. Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee. In *Interspeech*, pages 1404–1408, 2013.
- Nguyen Tan Viet Tuyen and Oya Celiktutan. Context-aware human behaviour forecasting in dyadic interactions. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 88–106. PMLR, 2022.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Anna Vatanen. Responding in early overlap: Recognitional onsets in assertion sequences. *Research on Language and Social Interaction*, 51(2):107–126, 2018. doi: 10.1080/08351813.2018.1413894.
- Dirk Nicolas Wagner et al. Augmented human-centered management. human resource development for highly automated business environments. *Journal of Human Resource Management*, 23(1):13–27, 2020.
- Nigel Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody 2004, International Conference*, 2004.
- Nigel G Ward. *Prosodic patterns in English conversation*. Cambridge University Press, 2019.
- Nigel G Ward, Anais G Rivera, Karen Ward, and David G Novick. Root causes of lost time and user stress in a simple dialog system. 2005.
- Nigel G Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. Turn-taking predictions across languages and genres using an lstm recurrent neural network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 831–837. IEEE, 2018.
- Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.
- Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Asap: Endowing adaptation capability to agent in human-agent interaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 464–475, 2023.
- Yi Xu. Prosody, tone and intonation. *The Routledge handbook of phonetics*, pages 314–356, 2019.

## BIBLIOGRAPHY

---

- Haimin Yang, Zhisong Pan, Qing Tao, et al. Robust and adaptive online time series prediction with long short-term memory. *Computational intelligence and neuroscience*, 2017, 2017.
- Li-chiung Yang. Visualizing spoken discourse: Prosodic form and discourse functions of interruptions. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001.
- Liu YANG, Catherine ACHARD, and Catherine PELACHAUD. Annotating interruption in dyadic human interaction. *LREC*, 2022.
- Victor H Yngve. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578, 1970.
- Davey Young et al. A conversation analysis of the acquisition and use of turn-taking practices in an english discussion class. 2015.
- Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 346–364. Springer, 2020.
- Margaret Zellers, David House, and Simon Alexanderson. Prosody and hand gesture at turn boundaries in swedish. *Proc. Speech Prosody 2016*, pages 831–835, 2016.
- Tiancheng Zhao, Alan W Black, and Maxine Eskenazi. An incremental turn-taking model with active system barge-in for spoken dialog systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 42–50, 2015.
- Shuo Zhou, Timothy Bickmore, Amy Rubin, Catherine Yeksigian, Molly Sawdy, and Steven R Simon. User gaze behavior while discussing substance use with a virtual agent. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 353–354, 2018.
- Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.
- Elisabeth Zima, Clarissa Weiß, and Geert Brône. Gaze and overlap resolution in triadic interactions. *Journal of Pragmatics*, 140:49–69, 2019.