

Robust Multimodal Turn-taking for Human-Machine Interaction



Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

Sam O'Connor Russell

School of Engineering

A dissertation submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I confirm that the work has been completed in full compliance with relevant university policies, including the Academic Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research. I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Sam O'Connor Russell, Monday 8th June, 2026

Abstract

During conversation, humans rapidly switch from speaking to listening and vice versa. This process is called turn-taking. Whilst second nature for humans, turn-taking presents a major challenge for emerging technologies such as voice assistants. At present, the majority of systems rely on detecting the silence after a speaking turn. This fails to replicate the fast-paced nature of human-human interaction. In recent years, predictive turn-taking models (PTTMs) have been proposed. Inspired by human turn-taking, PTTMs are neural networks which continuously predict whether or not a speaker change will occur in the near future. These models form an active area of research, and they have grown in sophistication and performance.

This thesis addresses a number of under-explored areas in PTTM development. Currently, a major obstacle is PTTMs rely upon manual transcription of conversations; a timely and costly process. The thesis therefore offers a comprehensive overview of automatic speech recognition (ASR) for automated conversational speech transcription. Insights include a gap in the performance of commercial and open-source ASR on the transcription of non-lexical aspects of speech. We then explore PTTM training with ASR transcription, finding no difference in the performance of PTTMs trained using word alignments from ASR and manually transcribed interactions. This enables PTTM training on larger, more diverse multimodal datasets for which manual transcription is typically unavailable.

PTTMs rely on speech alone to make their predictions, but turn human turn-taking is richly multimodal. The inclusion of visual cues such as gaze in PTTM has only been explored in a single, limited study. This major gap in the literature is addressed in this thesis by introducing the multimodal voice activity projection (MM-VAP) model, one of the first PTTMs to consider the inclusion of visual cues such as gaze, head pose and facial expression. We find that MM-VAP significantly outperforms the state-of-the-art audio-only approach: MM-VAP achieves an 83% balanced accuracy on the hold/shift prediction task, compared with 79% balanced accuracy in the audio-only state-of-the-art at the same noise level. Our evaluation is conducted across a comprehensive range of events, such as detecting the next speaker during silence and before overlapping speech. A further gap in the literature is that the performance of PTTMs in noise has not been considered, yet models will inevitably encounter noise once deployed. We address this in the thesis by adding environmental

noises such as babble and music noise to the Candor videoconferencing corpus. We find that our MM-VAP model is not inherently robust to noise, with performance falling from 83% to 54% balanced accuracy in +10 dB music noise. However, we find that when we adapt the training of the model to include examples of noise, MM-VAP outperforms the audio-only state-of-the-art audio-only model by a wide margin, at 75% balanced accuracy in +10 dB speech noise compared with 64% balanced accuracy audio-only state-of-the-art at the same noise level. We demonstrate that the performance increase arises from the exploitation of visual cues in the MM-VAP model. However, we find that the performance increase does not always generalise to new sources of noise, highlighting the importance of the training process in PTM development.

The thesis also explores the role of multimodal cues within the speech signal. Recent PTMs leverage self-supervised speech representations (S3Rs). These are learned representations as they capture prosodic, lexical and semantic aspects of the speech signal. However, it is difficult to establish which aspects of speech are exploited for prediction. We propose a vocoder-based methodology to selectively control the amount of prosodic and lexical information in speech. We find that the voice activity projection (VAP) model, an S3R-based turn-taking model from the literature, utilises both prosodic and lexical cues for prediction. However, when one feature is corrupted, the model can flexibly utilise the other without further training. A notable finding is that when speech is replaced with unintelligible noise following the prosodic contour of speech, performance remains above chance, at 69% balanced accuracy. This demonstrates that the prosodic contour alone is a powerful predictor of turn-taking patterns. We expand on this analysis by comparing the audio VAP model with large language model (LLM) based turn-taking models, which rely exclusively on text. We find that the LLM based models have a high number of false positive end-of-turn predictions. The audio-only turn-taking model does not suffer from the same errors, suggesting prosody plays a key role in turn-taking prediction. We therefore propose the Audio TurnGPT model, an LLM based turn-taking model which incorporates acoustic cues. We find that the model significantly improves prediction performance (90% vs. 85% balanced accuracy).

The thesis therefore demonstrates the complementary, additive role of multimodal cues in turn-taking prediction. The best-performing models exploit all cues, but there is a degree of redundancy, as individual cues can perform as turn-taking predictors. Multimodality not only improves performance but also increases robustness to noise, and means models retain performance even when cues are removed. The analyses, methodology, and models introduced in this thesis represent significant advances in the predictive turn-taking literature. It is hoped that this thesis leads to an increased awareness of multimodality when working human interaction data. All modalities that are available to interlocutors in a dataset should be considered, as should their interactions with another and their influence on unfolding interactions. We also offer a discussion on future work and the implications of our findings.

Acknowledgements

Completing a PhD has had lots of ups and downs, and many people to thank along the way.

First and foremost I want to thank my supervisor Prof Naomi Harte. Naomi first took me on as a summer intern when I was an undergraduate, giving me my first taste of research. I was grateful have the opportunity to return to TCD as a PhD student under her supervision. Naomi is a fantastic mentor. I admire her passion for research, sharp intellect and remarkable ability to remember details from meetings we had many months - or even years - earlier! Naomi is an excellent supervisor as she sets exceptionally high standards whilst encouraging and respecting her team's work-life balance. This makes for a fantastic working environment, which has made the process of doing the PhD enjoyable and fulfilling, and I'm very grateful for the experience.

I want to thank all my colleagues, my office mate Dr Samantha Kotey, Dr Justine Reverdy, Dr Ayushi Pandey, Dr Mark Anderson and Ed Storey. Thanks also to Dr Sébastien LeMageur for all the advice and best wishes to Zhaofeng Lin, Aris Papadopoulos, Victoria Ivanova, and Vu Hoang on the rest of their PhD journeys. Thanks also to Dr Delphine Chauarau, Dr Rishabh Jain, my internal examiner Prof François Pitié, Prof Anil Kokaram, and the rest of the video gang. Thanks also to all the technical staff in the Department, John Squires, Shane Hunt, and Eugene O'Rourke.

I want to also thank the housekeeping staff in the Department, in particular to Bernie and Liz, for all the cakes and morning natters these these last four years.

I owe a massive thanks to all the gang at my second home during the PhD, Capital Strength Weightlifting Club, especially for constantly checking in on me these last few months when I was writing up! Completing a PhD is at times a very isolating endeavour, so was great to have such an encouraging and social club to go to and switch off.

I want to thank my family, parents John and Dee, and brother Joe, for putting up with me these last four years, encouraging and supporting me, and letting me quit my job and move back home to go back to university for another four years.

Last but not least, thank you to Turgay for getting excited for every result, submission, review and paper, even when I couldn't find a way to get excited myself. Without your constant encouragement, none of this would have been possible.

Associated Publications

All technical work and writing of the publications was completed solely by the first author. Other authors were involved in a supervisory or advisory capacity.

Journal publications

- O'Connor Russell, Sam, Iona Gessinger, Anna Krason, Gabriella Vigliocco, and Naomi Harte. "*What automatic speech recognition can and cannot do for conversational speech transcription*". Research Methods in Applied Linguistics 3, no. 3. 2024: 100163.

Conference publications

- O'Connor Russell, Sam and Naomi Harte "*Visual Cues Support Robust Turn-taking Prediction in Noise*". Proc. Interspeech 2025, 1073-1077
- O'Connor Russell, Sam and Naomi Harte "*Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction*". Findings of the Association for Computational Linguistics: ACL 2025, pages 209–221, Vienna, Austria. Association for Computational Linguistics.
- O'Connor Russell, Sam, Delphine Charuau, and Naomi Harte "*The Role of Prosodic and Lexical Cues in Turn-Taking with Self-Supervised Speech Representations*" ICASSP 2026, Barcelona, Spain

Workshop publications

- O'Connor Russell, Sam and Naomi Harte "*Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction*". 2nd International Multimodal Communication Symposium, Frankfurt, Germany 91-92

Acronyms

ASR automatic speech recognition

AVSR audiovisual speech recognition

CPC contrastive predictive coding

FAU facial action unit

FTO floor-transfer offset

IPU interpausal unit

LLM large language model

MM-VAP multimodal voice activity projection

NLT non-lexical token

POS part-of-speech

PR precision-recall

PTTM predictive turn-taking model

RMS root-mean square

RNN recurrent neural network

ROC receiver operating characteristic

S3R self-supervised speech representation

SNR signal-to-noise ratio

SWB switchboard

TRP transition-relevant place

VAD voice activity detection

VAP voice activity projection

WER word error rate

List of figures

1.1	Multimodal cues utilised in human-human interaction	3
1.2	Examples of consumer-grade machines capable of multi-turn interaction with humans	4
1.3	A schematic of the training and development process for predictive turn-taking models	4
2.1	Illustrative example of an exchange in speaking turns between two interlocutors	14
2.2	Histogram of floor-transfer offsets values in Switchboard telephone speech corpus	16
2.3	Stills from the RoomReader corpus illustrating the use of Zoom videoconferencing software and recording setup of each corpus.	28
2.4	A still from a session of the Candor corpus illustrating the diverse nature of recording conditions, camera angles, and recording devices.	28
2.5	A schematic diagram depicting a shift and hold	33
2.6	A schematic diagram depicting an exchange between speakers with an overlap	33
2.7	Still from Candor with OpenFace features shown for the left participant	38
2.8	Visualisation of transformer attention	39
2.9	Schematic of reactive and predictive turn-taking models	41
2.10	A schematic of the recurrent neural network model proposed by Skantze (2017)	42
2.11	A schematic of the predictive turn-taking model presented by Roddy and Harte (2020)	44
2.12	Illustration of the voice activity projection model training objective	45
2.13	Voice activity projection model architecture	47
3.1	Bar chart of word error rates of six automatic speech recognition tools on speech corpora	65
3.2	Excerpt the Switchboard corpus comparing ground-truth and ASR transcription	69
3.3	Excerpt the Switchboard corpus comparing ground-truth and Amazon ASR word alignments	70

4.1	Validation set loss for each pair of learning rate and batch size hyperparameters	86
4.2	Mean training and validation loss for each training epoch	87
4.3	ROC curves for the voice activity projection turn-taking model, trained on the Switchboard corpus	96
4.4	Example of model output on an excerpt from the Switchboard corpus	97
4.5	ROC curves for the voice activity projection turn-taking model, trained on on the Candor corpus (Speechmatics ASR alignment)	99
4.6	Example of voice activity projection model output on a session from the Candor corpus	100
4.7	ROC curves for the voice activity projection turn-taking model, trained on either Switchboard (SWB) and deployed Candor (CND), or vice-versa	103
4.8	Cross-corpus analysis examples	104
5.1	Illustration of analysis windows used to compute prosodic and voice quality features	112
5.2	Mean normalised F_0 and intensity for holds and shifts in the Candor corpus .	114
5.3	Mean normalised jitter and shimmer for holds and shifts in the Candor corpus during windows preceding each hold and shift.	118
5.4	Schematic illustrating the process for the generation of prosody-matched noise and prosodic manipulation	123
5.5	Spectrogram of sample utterance from the Candor corpus with manipulations	124
5.6	Percentage of clean speech model accuracy achieved by when trained on manipulated speech	129
5.7	Example of model prediction during a shift from a session of the Candor corpus, with the output of models trained on clean and manipulated speech shown .	130
5.8	A comparison of voice activity projection and TurnGPT models on the same excerpt from session of the Switchboard corpus	137
5.9	A comparison of voice activity projection and TurnGPT models on the same excerpt of the Switchboard corpus	138
5.10	Comparison of the original Pairwise TurnGPT and our modified model trained on the Switchboard corpus	142
6.1	Histogram of facial action unit intensity values	149
6.2	A comparison of facial action units at key turn-taking events in the Candor corpus	150
6.3	A diagram depicting OpenFace gaze, head pose and facial landmark tracking in the Candor corpus	151
6.4	A diagram depicting OpenFace gaze, head pose and facial landmark tracking in the Candor corpus	152

6.5	Schematic of two versions of the proposed multimodal voice activity projection model	156
6.6	Training and validation losses for the multimodal voice activity projection early and late modality fusion models	159
6.7	Hold/shift prediction, before end of turn $FTO > +0$ ms	162
6.8	Shift prediction, before end of turn $FTO > 0$ ms	163
6.9	Balanced accuracy of models grouped by a minimum period of silence between turns	164
6.10	Transcription and qualitative comparison MM-VAP, VAP and video-only turn-taking models	173
6.11	Detail of next speaker behaviour during a shift	174
6.12	Comparison MM-VAP, VAP and video-only turn-taking models on a session from the Candor corpus	175
6.13	Detail of speaker behaviour during a shift	176
6.14	Voice activity projection model attention map	177
6.15	Voice activity projection model attention map	177
6.16	Multimodal voice activity projection model attention map	178
6.17	Multimodal voice activity projection model attention map	178
6.18	Multimodal voice activity projection model attention map	179
6.19	Multimodal voice activity projection model attention map	179
7.1	Average hold/shift prediction accuracy in each noise type	188
7.2	Average hold/shift prediction accuracy at each noise strength, expressed with the decibel signal-to-noise ratio	189
7.3	Model predictions during a shift in the Candor corpus at different noise conditions	191
7.4	Average speaker-level word error rate for each manipulation at each signal-to-noise ratio in decibels	194
7.5	Balanced accuracy of multimodal voice activity projection and voice activity projection models under different environmental noises, prosodic noises, and prosodic manipulation	198
7.6	An example showing a causal, cross-model attention mask used to align audio with text in our Audio-TurnGPT model	201
7.7	Architecture of the Audio-TurnGPT model	203
7.8	Predictions of Audio-TurnGPT on the Switchboard corpus	206
7.9	Comparing predictions of the text-only PairwiseGPT model, with our proposed text+audio Audio-TurnGPT model, and the audio-only VAP model	207
C.1	Transcription and qualitative comparison of turn-taking models on an excerpt from the Candor corpus	238

C.2	MM-VAP (early fusion) attention map depicting Speaker B audio self-attention, aggregated by word	239
C.3	MM-VAP (early fusion) attention map depicting Speaker B video self-attention, aggregated by word	239
C.4	MM-VAP (late fusion) attention map depicting Speaker B audio self-attention, aggregated by word	240
C.5	MM-VAP (late fusion) attention map depicting Speaker B video self-attention, aggregated by word	240
C.6	MM-VAP (late fusion) attention map depicting Speaker B audio→ Speaker A audio cross-attention (layer 1, aggregated by word)	241

List of tables

2.1	Summary of studies from our review of turn-taking	20
2.2	Summary of publicly available corpora of conversational interaction	27
2.3	Facial action unit codes, with descriptions and sample images	37
2.4	Overview of turn-taking models identified in our review	56
3.1	Descriptive statistics of the corpora used to evaluate ASR transcription	61
3.2	Sample transcription of LibriSpeech and Switchboard errors underlined	64
3.3	Word-error rates of six automatic speech recognition tools on speech corpora	64
3.4	Top five words in human and ASR transcripts of Switchboard and RoomReader corpora	65
3.5	The top five words most common correctly-transcribed words by ASR, as well as the top five deletion, substitution, and insertion errors.	66
3.6	List of non-lexical tokens alongside the source from which they were identified	67
3.7	Error analysis of non-lexical tokens in ASR transcriptions	68
3.8	Metrics for non-lexical tokens (non-lexical tokens (NLTs)). The classic WER penalises misspellings of non-lexical tokens (NLTs), whereas the adjusted WER (adj-WER) does not. The % recall is the combined percentage of non-lexical tokens (NLTs) that are correctly identified or misspelt.	68
3.9	Word alignment and absolute onset and offset error in seconds	69
4.1	Voice activity projection model architectural details	83
4.2	Overview of the three voice activity projection models trained in this chapter .	84
4.3	Number of sessions and mean session duration of cross-validation splits of the Switchboard and Candor corpora.	85
4.4	Epochs corresponding to the lowest validation loss	86
4.5	Statistics for shifts and holds obtained by varying the minimum floor-transfer offset	88
4.6	Turn-taking metrics and evaluation points	90
4.7	Average hold/shift prediction across on the Switchboard corpus	91
4.8	Average shift prediction across on the Switchboard corpus	92

4.9	Average hold/shift prediction across 5 folds on the Candor and Switchboard (ASR) corpora.	97
4.10	Results for shift prediction evaluation metrics across on Candor and Switchboard corpora	98
4.11	Cross-corpus hold/shift prediction results	102
4.12	Cross-corpus generalization performance, with the relative % change from within-corpus performance (Tables 4.8 and 4.10) shown in brackets.	105
5.1	Linear mixed-effects model for mean $F0$ in Hz in the Candor corpus	113
5.2	Linear mixed-effects model for mean intensity in decibels in the Candor corpus	115
5.3	Linear mixed-effects model for mean jitter in the Candor corpus	116
5.4	Linear mixed-effects model for mean shimmer in the Candor corpus	117
5.5	Overview of prosody-matched noise and prosodic manipulations of speech .	120
5.6	Candor corpus test set word error rate	123
5.7	Voice activity projection model performance on clean speech and on acoustic manipulations	125
5.8	VAP model performance when trained and tested on either clean or manipulated speech	128
5.9	Wav2vec2.0 model performance trained on clean speech and tested on acoustic manipulations	129
5.10	Special token counts and descriptions for TurnGPT (Ekstedt and Skantze 2020) and Pairwise TurnGPT (Leishman et al. 2024) extraction methods . . .	133
5.11	TurnGPT (Ekstedt and Skantze 2020) method of serialising speaking turns .	134
5.12	Pairwise TurnGPT (Leishman et al. 2024) method of token alignment	134
5.13	Performance metrics for the special tokens across TurnGPT and Pairwise (p/w) TurnGPT models on the test set	136
5.14	Token alignment illustrating a turn transition, using the Pairwise TurnGPT (Leishman et al. 2024) method	140
5.15	Large language model turn-taking model performance metrics across models and event types	141
6.1	Summary of visual features extracted with OpenFace	155
6.2	Multimodal voice activity projection model configuration parameters	157
6.3	Epochs corresponding to the lowest validation loss for each fold of cross-validation.	158
6.4	Average hold/shift prediction performance on Candor corpus	161
6.5	Average shift prediction scores across 5 folds of the Candor corpus	162
6.6	Multimodal voice activity projection model ablation study	165

7.1	Types of additive acoustic noise and descriptions	185
7.2	Test sets generated by adding noise to the Candor corpus and descriptions .	187
7.3	Overview of the voice activity projection and multimodal voice activity projection model training conditions on the Candor corpus training conditions	187
7.4	Hold/shift prediction accuracy of each model in different types of noise	190
7.5	Shift/hold prediction accuracy of the multimodal voice activity projection model under different training conditions	192
7.6	Shift prediction accuracy of each model in different types of noise	192
7.7	Acoustic disruptions grouped by their preservation of the lexical content, F_0 and intensity contours of the original speech signal	196
7.8	Performance metrics for special token prediction with Pairwise TurnGPT and the proposed Audio TurnGPT model	205
7.9	Performance metrics for special token prediction with Pairwise TurnGPT and the proposed Audio-TurnGPT mode	205
A.1	Baseline (always-hold) model F_1 scores and balanced accuracy for shifts and holds at different minimum floor-transfer offset (FTO) thresholds in the switchboard (SWB) and candor corpora.	225
B.1	One-sample t-tests comparing the VAP model trained on Switchboard (SWB) (ground-truth alignment) against an always-hold baseline model.	228
B.2	One-sample t-tests comparing the VAP model trained on Switchboard (SWB) (ASR alignment) against an always-hold baseline model.	229
B.3	One-sample t-tests comparing the VAP model trained on Candor (CND) against an always-hold baseline model	230
B.4	Paired t-tests comparing VAP models trained on Switchboard (SWB) (ground- truth) and SWB (ASR) alignments	231
B.5	Paired t-tests comparing CND $\rightarrow$ SWB and SWB VAP models cross-corpus performance	232
B.6	Paired t-tests comparing SWB $\rightarrow$ CND cross-corpus performance	233
B.7	Paired t-tests comparing VAP and MM-VAP (early fusion) models on the Candor corpus	234
B.8	Paired t-tests comparing VAP and MM-VAP (late fusion) models on the Candor corpus	235
B.9	Paired t-tests comparing MM-VAP (early fusion) and MM-VAP (late fusion) models	236

Table of contents

Abstract	x
Associated publications	x
Acronyms	xi
List of figures	xiii
List of tables	xvii
1 Introduction	1
1.1 What is conversation?	1
1.2 The challenges in designing conversational machines	2
1.2.1 Understanding human conversation	2
1.2.2 Achieving conversational interaction with machines	2
1.3 Predictive turn-taking models	4
1.3.1 Open challenges in predictive turn-taking	4
1.4 Thesis statement and scope	6
1.4.1 Research questions	6
1.4.2 Thesis scope	8
1.5 A high-level outline of this thesis	8
1.6 Contributions	10
2 Literature review	13
2.1 Turn-taking fundamentals	13
2.1.1 The timing of transitions between speakers	14
2.1.2 Speaker transitions in corpus of English telephone speech	15
2.1.3 Listeners anticipate speaker changes ahead of time	15
2.2 A review of the turn-taking literature	16
2.2.1 Corpus studies of multimodal turn-taking cues	16
2.2.2 Behavioural studies of multimodal turn-taking cues	18

2.2.3	The interplay between verbal and non-verbal turn-taking cues	21
2.2.4	Summary of multimodal cues	22
2.3	Corpora for turn-taking research and model development	23
2.3.1	Conversational speech corpora	23
2.3.2	Available English language corpora for turn-taking model development	23
2.3.3	Selecting corpora for our work	25
2.4	Methods for turn-taking analysis and model evaluation	29
2.4.1	Transcription of conversational interactions	29
2.4.2	What is a transcription?	29
2.4.3	Automating transcription with ASR	30
2.4.4	Gaps in the ASR literature	31
2.4.5	Extraction of speaking turns and the voice activity signal	32
2.4.6	Extraction of turn-taking events	32
2.4.7	Extraction of prosody and voice quality features	34
2.4.8	Extraction of visual features	35
2.5	Machine learning architectures and evaluation	36
2.5.1	The transformer architecture	38
2.5.2	Self and cross attention	39
2.5.3	The GPT-2 architecture	40
2.6	Predictive turn-taking models	40
2.6.1	Predictive versus reactive turn-taking models	40
2.6.2	Predicting future speaking activity	41
2.6.3	Alternative approaches to predictive turn-taking	42
2.6.4	Incorporating visual cues in a predictive turn-taking model	43
2.6.5	The voice-activity projection turn-taking model	44
2.6.6	Self-supervised representations in predictive turn-taking models	50
2.6.7	The TurnGPT and PairwiseTurnGPT models	51
2.6.8	Summary of predictive turn-taking models	52
2.6.9	Evaluation of trained predictive turn-taking models	52
2.6.10	Turn-taking in multiparty interaction	53
2.7	Gaps identified from our literature review	54
2.7.1	The use of manual transcriptions	54
2.7.2	A lack of non-verbal cues	54
2.7.3	A limited understanding of the contributions of prosodic and lexical cues	55
2.7.4	A limited understanding of performance in real-world conditions	55
3	Transcribing conversations with ASR	57
3.1	Defining transcription	59
3.2	Transcribing with automatic speech recognition	59

3.2.1	Open source tools	60
3.2.2	Commercial tools	60
3.2.3	Corpora	60
3.3	Evaluating an ASR transcription	62
3.3.1	Insertions, deletions, substitutions and the error rate	62
3.3.2	Word timing and overlap	63
3.4	Conversational speech transcription with ASR	63
3.4.1	Transcribing what was said	63
3.4.2	Error analysis of ASR	64
3.4.3	<i>Uhm</i> or <i>ehm</i> ? ASR transcription of non-lexical tokens	66
3.4.4	Word timing and overlap	68
3.5	Discussion of results	70
3.5.1	Capturing what was said	70
3.5.2	Transcribing NLTs: <i>Uhm</i> or <i>uh-uhm</i> ?	71
3.5.3	Capturing overlapping speech and word alignment	71
3.6	How can ASR support transcription?	72
3.6.1	Caution on the use of the WER	72
3.6.2	ASR as an initial transcription aid	72
3.6.3	How much time can be saved by using ASR?	73
3.6.4	On replacing human transcribers	74
3.6.5	Coding of pragmatic and paralinguistic phenomena	74
3.6.6	Additional factors to consider in an ASR deployment	74
3.6.7	Limitations	76
3.6.8	Future research directions in conversational speech transcription . . .	77
3.7	General conclusion	78
3.7.1	Turn-taking model development with ASR transcription	78
4	Predictive turn-taking with audio	81
4.1	Training an audio-only turn-taking model	83
4.1.1	Model technical specification	83
4.1.2	Audio and label preparation	83
4.1.3	Cross validation	84
4.1.4	Hyperparameter selection	85
4.1.5	Model training procedure	85
4.1.6	Running the trained models and extracting p_{future}	86
4.2	Metrics for turn-taking model evaluation	87
4.2.1	Identifying shift and hold times	87
4.2.2	Hold/shift prediction metrics	88
4.2.3	Shift prediction metrics	89

4.2.4	Summary of key turn-taking metrics	90
4.2.5	Baseline model	90
4.3	Audio-only turn-taking on the Switchboard telephone speech corpus	90
4.3.1	Performance on hold/shift prediction metrics	91
4.3.2	Performance on shift prediction metrics	92
4.3.3	Comparing models trained on manual and ASR alignments	92
4.3.4	Performance relative to baseline model and the literature	92
4.3.5	Analysing performance with ROC and PR curves	93
4.3.6	Performance on an example from the corpus	94
4.3.7	Discussion	95
4.4	Audio-only turn-taking on the Candor videoconferencing corpus	95
4.4.1	Performance on hold/shift prediction metrics	95
4.4.2	Performance on shift prediction metrics	98
4.4.3	Performance relative to baseline model and the literature	98
4.4.4	Analysing performance with ROC and PR curves	99
4.4.5	Performance on an example from the corpus	99
4.4.6	Discussion	101
4.5	A cross-corpus analysis of audio-only turn-taking	101
4.5.1	Performance on turn-taking metrics	102
4.5.2	Performance analysis with ROC and PR curves	102
4.5.3	Example of the cross-corpus model predicting a speaker transition	105
4.5.4	Discussion	105
4.6	General conclusion	106
5	Prosodic and lexical cues in predictive turn-taking	109
5.1	Prosodic cues of holds, shifts, and mid-turn speech	111
5.1.1	Extracting acoustic features	111
5.1.2	Statistical analysis	111
5.1.3	Fundamental frequency F_0 and turn-taking	113
5.1.4	Intensity and turn-taking	114
5.1.5	Voice quality and turn-taking	115
5.1.6	Discussion	116
5.2	Isolating prosodic and lexical cues in the speech signal	119
5.2.1	Using a vocoder to control for prosodic and lexical cues in speech	119
5.2.2	Altering the speech signal with WORLD synthesis	120
5.2.3	Selective removal of prosodic and lexical cues from speech	122
5.3	Lexical and prosodic cues in audio-only turn-taking	124
5.3.1	Audio-only turn-taking when prosodic and lexical cues are removed	124

5.3.2	Quantifying the contribution of lexical and prosodic cues to turn-taking prediction	126
5.3.3	Sample prediction	127
5.3.4	Generalisation to other self-supervised speech representation (S3R) .	128
5.3.5	Discussion	130
5.4	Going purely lexical: turn-taking with large language models	131
5.4.1	Pre-processing of text	132
5.4.2	Serialisation of speaking turns (TurnGPT)	132
5.4.3	Pairwise alignment of tokens (Pairwise TurnGPT)	132
5.4.4	Training and validating large language models for turn-taking	133
5.4.5	Predictive turn-taking with large language models	135
5.4.6	Comparing audio and lexical turn-taking models	136
5.4.7	Proposed architectural modifications	139
5.4.8	Discussion	143
5.5	General conclusion	144
6	Incorporating visual cues in a predictive turn-taking model	145
6.1	Visual features during speaker transitions	148
6.1.1	OpenFace feature extraction	148
6.1.2	Facial action units	148
6.1.3	Facial action activity during speaker transitions	148
6.1.4	Analysis of facial action units during speaker transitions	149
6.1.5	Facial landmarks, eye gaze and head pose	151
6.1.6	Discussion	153
6.2	The multimodal voice activity projection model	154
6.2.1	Feature selection and pre-processing	154
6.2.2	Proposed multimodal turn-taking model architectures	154
6.2.3	Training and evaluation procedure	158
6.3	Assessing the performance of the multimodal voice activity projection model .	159
6.3.1	Comparing MM-VAP with VAP	159
6.3.2	Comparing early and late modality fusion versions of MM-VAP	160
6.3.3	Comparing MM-VAP with a video-only model	161
6.3.4	ROC and PR curves	161
6.3.5	The effect of transition length	162
6.3.6	Ablation study	163
6.3.7	Discussion	164
6.4	Qualitative analyses	167
6.4.1	Examples of speaker transitions	167
6.4.2	Discussion	168

6.5	Attention analysis	169
6.5.1	Aggregating word-level attention	169
6.5.2	The VAP audio-only model	170
6.5.3	MM-VAP, early modality fusion	170
6.5.4	MM-VAP, late modality fusion	171
6.5.5	Discussion	171
6.6	General conclusion	172
7	Towards real-world deployment of predictive turn-taking models	181
7.1	A framework for exploring predictive turn-taking in noise	184
7.1.1	Synthesising environmental noise	184
7.1.2	Prosody-matched noise and manipulation	185
7.1.3	Controlling the strength of additive noise	185
7.1.4	Evaluating PTM performance in noise	186
7.1.5	Augmenting the training set with noisy data	186
7.1.6	Impact of the noise on the WER	187
7.2	Turn-taking prediction in environmental noise	187
7.2.1	The impact of environmental noise on PTM performance	187
7.2.2	Augmenting the training set with noisy examples	188
7.2.3	The role of the word alignment in training	190
7.2.4	Predicting a shift between speakers	190
7.2.5	Generalisation to unseen noises	191
7.2.6	Impact on shift prediction	192
7.2.7	Discussion	192
7.3	Uncovering the role of acoustic and lexical cues in turn-taking prediction . . .	194
7.3.1	The impact of noise on speech intelligibility	194
7.3.2	Exploring turn-taking prediction performance in noise	196
7.3.3	Discussion	199
7.4	Adding speech cues to an LLM-based turn-taking model	199
7.4.1	Integrating audio cues into an LLM-based turn-taking model	200
7.4.2	Aligning modalities with masked cross-model attention	200
7.4.3	The Audio-TurnGPT Model	201
7.4.4	Training the Audio-TurnGPT model	202
7.4.5	Predictive turn-taking with Audio-TurnGPT	202
7.4.6	Comparing the predictions of Audio-TurnGPT with Pairwise TurnGPT .	204
7.4.7	Discussion	208
7.5	Deployment of turn-taking models to real-world interaction	209
7.5.1	Deployment of models from the literature to human-robot interaction	209
7.5.2	Deployment of MM-VAP to human-robot interaction	210

7.5.3	Remaining issues to be overcome in a real-world deployment	211
7.5.4	The role this thesis plays in achieving predictive turn-taking in human-robot interaction	211
8	Conclusion and future work	213
8.1	Summarising the key findings and insights of this thesis	213
8.1.1	Exploring machine transcription of conversational speech and demonstrating its utility in training data generation	213
8.1.2	Understanding the role of visual cues in predictive turn-taking	214
8.1.3	Turn-taking model evaluation in real-world deployments	215
8.1.4	Uncovering the role of prosodic and lexical cues in predictive turn-taking	216
8.2	Contemporary related work	217
8.3	Directions for future work	218
8.3.1	Develop predictive turn-taking models for multiparty interaction	218
8.3.2	Robust multimodal architectures	219
8.3.3	Real-world deployment of models to human-robot interaction	219
8.4	Concluding remarks	220
	Appendices	221
	A Baseline turn-taking model performance	223
	B Statistical significance tests	227
	C Extra examples	237
	Bibliography	242

Chapter 1

Introduction

1.1 What is conversation?

This thesis deals with computer algorithms for turn-taking, a process which occurs in every conversation between humans. As we spend much of our days interacting with one another, conversation is a ubiquitous, almost banal feature of everyday life. It has therefore been a focal point for numerous disciplines, including linguistics, psychology, and human–computer interaction (Clift 2016; Schegloff 2000a; Turnbull 2003; L. Clark, Pantidi, et al. 2019).

Turnbull defines conversation as *"a situation in which two socialised and co-present individuals talk to one another as they go about their everyday lives"*. Conversations occur every day in social, professional and educational contexts (Turnbull 2003). They can have specific goals, or they can be simple social chit-chat (L. Clark, Pantidi, et al. 2019). No matter the type of conversation, humans show a remarkable ability to coordinate speaking turns, ensuring an orderly, organised exchange of information between interlocutors (Schegloff 2000a; Sacks et al. 1974). Although conversation involves language, it encompasses more than just the words that are spoken (Turnbull 2003): if the question *"would you like to come to my house for dinner tomorrow?"* is followed by an extended silence, this would be awkward as the silence is interpreted as a 'no' (Kendrick and Torreira 2015).

Humans are highly attuned to non-lexical aspects of conversation, including silence (Turnbull 2003), body language (Knapp et al. 1978) and facial expression (Nota et al. 2023). This makes conversation richly multimodal (Holler and Levinson 2019). Recently, conversation has attracted interest from computer science, where the development of machines that can engage humans in naturalistic conversation is a key goal (Skantze 2021; L. Clark, Pantidi, et al. 2019). This makes the study of conversation highly interdisciplinary.

1.2 The challenges in designing conversational machines

Much work is being conducted to build machines capable of natural conversation with humans (L. Clark, Pantidi, et al. 2019). For this to be achieved, a deep knowledge of human conversation is required (Turnbull 2003; L. Clark, Pantidi, et al. 2019). We briefly review the ways in which humans converse, and discuss the latest consumer-grade voice assistants. This enables us to outline the present challenges in designing machines capable of conversational interaction.

1.2.1 Understanding human conversation

Throughout our daily lives, we engage in many different types of conversation, from social chit-chat to professional meetings (L. Clark, Pantidi, et al. 2019). Conversation is multimodal, as we simultaneously produce and respond to cues across verbal and non-verbal channels (Holler and Levinson 2019). Some of the cues involved in conversation are highlighted in Figure 1.1, which depicts a two-party in-person interaction. These multimodal cues are essential in regulating the way in which a conversation unfolds. This is because humans try to speak one at a time, minimising overlapping speech (Holler, Kendrick, et al. 2016; Sacks et al. 1974). Cues from speech, body language, gaze, in addition to complex social and cultural norms, are central to this organisation, governing who speaks when (Sacks et al. 1974; Levinson and Torreira 2015; Stivers et al. 2009). This process is referred to as turn-taking in the literature, see Skantze (2021) for a review.

1.2.2 Achieving conversational interaction with machines

Given the ubiquity of conversation in our lives, it is unsurprising that machines are being developed that are capable of voice-based interaction with humans. Amazon's Alexa and Apple's Siri are capable of limited interaction with humans (Akhtiamov et al. 2020). The consumer market for these smart speakers breached 15 billion USD in 2025 and is forecast to double by 2030¹, indicating a strong and growing demand (L. Clark, Doyle, et al. 2019). This demonstrates that we like using our voice for interacting with machines: users report higher satisfaction, enjoyment, and less effort when using a voice assistant compared with a chatbot (Rzepka et al. 2022), and speaking to a computer is 3 times faster than typing (Ruan et al. 2016; Ruan et al. 2018). Although user acceptance is generally positive, it is not universal (Tabassum et al. 2019) and varies with age (Gollasch and Weber 2021).

Most commercial voice assistants are only capable of basic, single-turn interaction, initialised by 'wake words' ("*Hey Siri!*"). This constrained form of interaction is not *conversation*, as it is limited to a single pair of speaking turns. A new frontier is the development of machines

¹<https://www.fortunebusinessinsights.com/smart-speaker-market-106297>

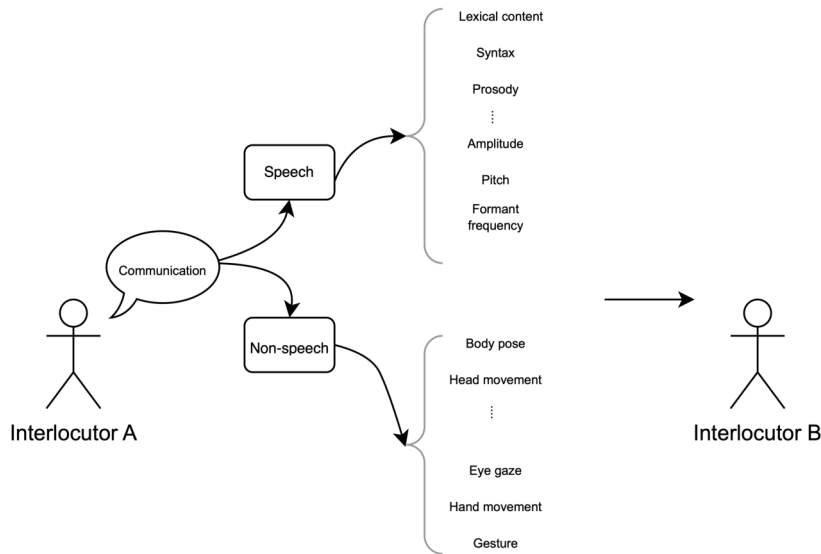


Figure 1.1 Human communication is multimodal involving an array of cues, some of which are depicted above. Humans effortlessly infer meaning from these signals, enabling successful and coordinated conversational interactions to take place.

capable of conversation, i.e. interactions across multiple turns, with humans (Akhtiamov et al. 2020; L. Clark, Doyle, et al. 2019). These conversational machines have potential applications in sectors such as e-health, education, and business (Alnefaie et al. 2021). Examples of machines capable of interaction across multiple turns include Futhat robot, ChatGPT voice mode, Gemini Live, shown in Figure 1.2.

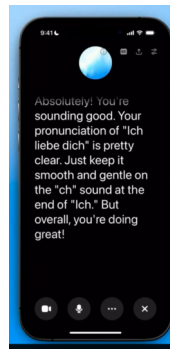
Despite being capable of multi-turn interaction, conversations with these machines are less spontaneous and natural than human-human interaction (Skantze 2021). A major contributing factor is that most turn-taking algorithms in use today are *reactive*, waiting for silence after a speaking turn and then deciding whether or not to speak (Skantze 2021). Humans do not take turns in this manner: in a typical human conversation, there is, on average, 200 milliseconds of silence between speaking turns (Stivers et al. 2009; Heldner and Edlund 2010). This is far faster than language formation, which takes at least 600 ms (Indefrey 2011).

Human turn-taking is therefore *predictive*, as we begin to formulate responses ahead of time, whilst continuing to listen (Holler, Kendrick, et al. 2016; Heldner and Edlund 2010; Sacks et al. 1974). The fast-paced nature of human turn-taking therefore cannot be replicated by reactive turn-taking models. This results in human-machine interactions that are less fluid² than those between humans (Li et al. 2022; Woodruff and Aoki 2003).

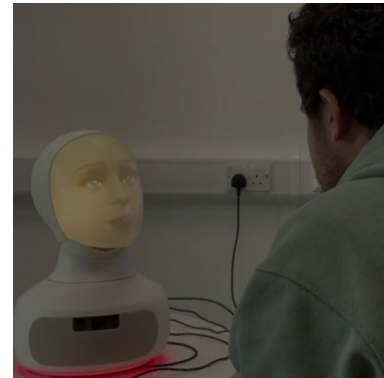
²It is important to highlight here that the latencies of automatic speech recognition, dialogue management and so on, also contribute to a lack of fluid interaction; however, the focus of this thesis is turn-taking models



(a) The interface of the Gemini live voice agent



(b) ChatGPT Voice interface



(c) Furhat robot

Figure 1.2 Some consumer-grade machines capable of multi-turn conversational interaction. Images sourced from Google (Gemini) and OpenAI (ChatGPT) show their smartphone interfaces, and the Furhat robot is an embodied agent (Furhat Robotics).

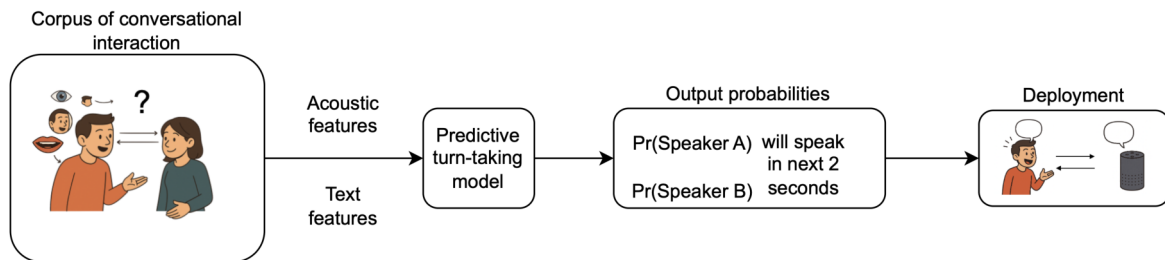


Figure 1.3 Diagram depicting the training and deployment process for predictive turn-taking models. The models are trained on corpora of human-human interaction. Typically, the models are trained to predict the future speaking activity of one or both of the interlocutors. (Pr=probability)

1.3 Predictive turn-taking models

PTTMs have been proposed to overcome the limitations of reactive turn-taking models. Inspired by human turn-taking, PTTMs are trained to predict upcoming speaking changes ahead of time on large corpora of human interaction (Skantze 2021). The assumption in the PTTM literature is that once deployed, PTTMs will support natural, fluid conversation between humans and machines. A diagram of the process for training a PTTM is shown in Figure 1.3.

1.3.1 Open challenges in predictive turn-taking

Much progress has been made in PTTM development; however, several key areas remain unaddressed:

Predictive turn-taking models ignore visual cues Despite turn-taking and human conversation more broadly involving speech and visual cues, the vast majority of PTTMs algorithms are audio-only (Ekstedt and Skantze 2022b), text-only (Leishman et al. 2024), or utilise a mixture of audio and text features (Roddy, Skantze, et al. 2018a). This leads to vital cues that are highly connected with turn-taking, such as gaze aversion (Kendon 1967), being missed.

Telephone speech and manual transcription Many telephone speech corpora, such as Switchboard (Godfrey et al. 1992), exist and are provided with accurate manual transcriptions. Telephone speech has therefore been widely used to train PTTMs (Ekstedt and Skantze 2022b; Roddy, Skantze, et al. 2018a; Skantze 2017). These transcriptions provide the start and end times of each speech segment, which is necessary for PTTM training (Ekstedt and Skantze 2022b). The lexical content may also be used (Leishman et al. 2024).

However, this reliance on manual transcription constrains the availability of data for training, as many multimodal corpora are not provided with a manual transcription e.g. the 860 hour Candor corpus (Reece et al. 2023) of videoconferencing interaction is only provided with an ASR transcription. It is currently unknown whether ASR transcription is sufficiently accurate to support PTTM development.

Visual cues are missed by models The focus on telephone speech means models are trained on the audio taken from interactions where interlocutors could not see one another. This puts a speech-based model at an inherent disadvantage when deployed in a setting where interlocutors can see one another. This is because in a face-to-face interaction, human interlocutors rely on both verbal and visual cues. Yet, the model can only process the audio signal, missing vital visual cues linked with turn-taking, such as gaze aversion (Kendon 1967).

At the time of thesis planning, only a small number of corpora have made use of multimodal (i.e. audio + video) corpora (Roddy, Skantze, et al. 2018a), and tend to be very small in scale, e.g. the 11 hour Mahnob corpus (Bilakhia et al. 2015). There are also comparatively fewer multimodal corpora of conversation interaction that could be used for training, and transcription quality can vary. Hence, it is unknown whether including visual cues can improve turn-taking prediction where interlocutors can see as well as hear one another.

A lack of multimodal analysis Non-verbal communication itself is also highly multimodal. Prosody and other aspects of speech such as voice quality influence turn-taking in human interaction (Gravano and Hirschberg 2009; Bögels and Torreira 2015). There is, however, a lack of analysis of cues from speech and non-verbal communication are utilised by PTTMs. Earlier models used engineered features such as part-of-speech tags and GeMAPS features (Eyben et al. 2015), a standardised set of acoustic features. This enabled robust ablations to be conducted to determine the best cues for prediction (Skantze 2017; Roddy, Skantze,

et al. 2018a). This work revealed that in these models, prosodic, rather than linguistic cues, contributed most to prediction performance.

However, recently engineered features have been replaced with self-supervised learning in newer PTTMs (Ekstedt and Skantze 2022b). These are rich representations of speech, but they pose interpretability challenges (Gimeno-Gómez et al. 2025; Pasad et al. 2021). There has been limited analysis of lexical, prosodic cues in newer PTTMs. It is therefore unknown whether these models learn to exploit the same turn-taking cues as humans, or different ones.

Real-world testing conditions As PTTMs are developed in the lab, they are typically tested on the same corpora on which they were trained. An important aspect of a real-world deployment is the impact of noise on performance. PTTMs are intended to be deployed *in-the-wild*, where they will inevitably encounter diverse sources of noise such as music, background chatter, etc. Yet, to our knowledge PTTMs have to date only been tested on interactions free from background noise interference. The impact of noise on PTTM performance is therefore unknown and unquantified, despite robustness to noise being critical in a real-world deployment.

1.4 Thesis statement and scope

This thesis argues that understanding multimodality is central to the goal of developing machines that can engage in conversational interaction with humans. We focus on turn-taking, a major unsolved challenge in the development of conversational artificial intelligence. Our work significantly advances our understanding of multimodality in predictive turn-taking through a number of research questions set out as follows.

1.4.1 Research questions

Our work begins with exploring methods to automate transcription, a time-consuming and costly manual process that is necessary in any work involving conversation. ASR has the ability to automate transcription, but the performance of modern ASR on conversational speech transcription is understudied. We therefore ask:

RQ 1 *Can modern ASR systems support accurate transcription of conversational interaction?*

Most work in turn-taking has relied upon manually transcribed interactions, severely limiting the amount of available data for model training and development. Next, we assess whether ASR transcription is suitable for PTTM training, asking:

RQ 2 *Can state-of-the-art PTTMs from the literature be trained with automatically-transcribed data?*

We then shift our focus toward multimodality, and the central role it plays in turn-taking, initially focussing on speech. Rich multimodal cues from speech, including prosody (intensity, fundamental frequency) and lexical cues (syntax, grammar and so on), support turn-taking in human interaction. Yet, little is known about the role of these cues in recent PTTMs, which use self-supervised speech representations. We therefore ask:

RQ 3 *To what extent can prosodic and lexical cues support accurate and robust turn-taking with state-of-the-art PTTMs from the literature?*

We then shift our focus towards visual cues. The development of PTTM has been overly focussed on telephone speech, and the overwhelming majority of PTTM ignore visual cues. The role visual cues can play in prediction is therefore unclear. Accordingly, we ask:

RQ 4 *Are visual cues necessary for accurate turn-taking prediction for interactions where participants can see as well as hear each other?*

Given the strong connection between visual cues and turn-taking in the linguistics literature, we propose a novel, multimodal PTTM model which incorporates visual as well as speech-based cues. We explore its performance, asking:

RQ 5 *Can the incorporation of visual cues improve the performance of PTTM?*

Once deployed, PTTM will inevitably encounter noise, yet to the best of our knowledge, these models have only been tested in interactions recorded in clean, noise-free conditions. We therefore ask:

RQ 6 *How is the performance of PTTM impacted by noise?*

Finally, our last research question represents a deeper analysis of multimodality. Specifically, we ask whether multimodality is leveraged by PTTMs in similar ways to humans. For instance, in human interaction, prosody is used by humans to overcome lexical ambiguities, particularly when posing questions (Bögels and Torreira 2015). In face-to-face settings, when communicating in noise, being able to see a speaker makes speech more intelligible (MacLeod and Summerfield 1987).

We investigate whether PTTMs behave similarly, asking:

RQ 7 *Can PTTM utilise multimodal cues in the same way as humans?; For instance, by using*

(a) *prosodic cues to resolve lexical ambiguity*

(b) *visual cues to overcome noise*

In answering these questions, this thesis comprises a comprehensive analysis of multimodality and its role in predictive turn-taking, a key unsolved problem in human-robot interaction.

1.4.2 Thesis scope

As predictive turn-taking is an active and quickly developing field with many open challenges, we must precisely define the scope of this work. We focus on PTTMs for two-party interaction, where most work has been conducted. We train PTTM on pre-recorded corpora of human-human interaction, as is typical in the literature (Skantze 2021). Deployment to real-world interaction is therefore beyond the scope of this work.

1.5 A high-level outline of this thesis

This thesis is organised into a literature review and work chapters. We briefly summarise the contents and findings of each chapter. Our contributions are then outlined in Section 1.6.

Chapter 2: Literature review

This chapter offers an introduction to the core concepts and methodology of the thesis. Section 2.1 introduces turn-taking and Section 2.2 provides a comprehensive overview of the turn-taking literature. Section 2.3 reviews multimodal corpora and outlines the corpora used in this thesis for analysis and model training. Section 2.4 discusses transcription and outlines the feature extraction methodology, which facilitates our later analyses of turn-taking. The remainder of the literature review deals with machine learning models. Section 2.5 reviews the core machine learning architectures used in the thesis. Section 2.6 offers a comprehensive overview of state-of-the-art turn-taking models, and Section 2.7 summarises the gaps in the turn-taking literature identified in our review.

Chapter 3: Transcribing Conversations with ASR

Chapter 3 contains our study of transcription with automatic speech recognition. An important outcome of this chapter is a comprehensive validation of ASR, supporting the work of later chapters. Section 3.1 defines transcription, and Section 3.2 outlines ASR transcription platforms. Sections 3.3 and 3.4 outline our evaluation methodology and results. In Section 3.6, we offer a comprehensive discussion of the implications of our findings, aimed at a general audience of practitioners wishing to use ASR in their work. This chapter fills a gap in

the literature, as there is a strong demand for ASR, but little analysis aimed at researchers outside the ASR community.

Chapter 4: Predictive turn-taking with audio

This chapter analyses the performance of the VAP model, an audio-only turn-taking model from the literature (Ekstedt and Skantze 2022b). Sections 4.1 and 4.2 discuss the training procedure and evaluation protocol. We first analyse performance when trained and tested on the Switchboard corpus of telephone speech in Section 4.3.

In Section 4.4, we evaluate the VAP model's performance on the Candor corpus of videoconferencing interaction, a new domain. We find that, despite the interlocutors being able to see as well as hear one another, and the model not using any visual cues, performance is strong. In Section 4.5 we train the VAP model on Switchboard, deploy it to Candor, and vice-versa. We note a sharp drop in performance, raising questions about PTTMs' ability to generalise to new datasets and types of interaction.

Chapter 5: Prosodic and lexical cues in predictive turn-taking

This chapter presents a deep dive into the role of prosodic and lexical cues in PTTMs. In Section 5.1, we use linear mixed effects models and find patterns in speech which distinguish turn-taking from mid-turn speech in the Candor corpus, and could therefore be exploited in trained PTTMs. In Section 5.2 we develop a methodology to selectively preserve or mask prosodic and lexical cues in speech. In Section 5.3, we find that either prosodic or lexical cues can be removed from speech whilst the VAP model flexibly utilises the other, demonstrating both their role in turn-taking prediction. In the final part of the chapter in Section 5.4, we direct our attention to LLM based turn-taking models. We find they make a high number of false positive end-of-turn predictions; errors that are not found in an audio PTTM.

Chapter 6: Incorporating visual cues in a predictive turn-taking model

Having thoroughly explored prosodic and lexical cues from speech, we turn our attention toward visual cues in Chapter 6. We open the chapter with an analysis of visual features in the Candor corpus. In Section 6.1, we find that listeners exhibit patterns of mouth and facial movement which resemble speech before the onset of a speaking turn: visual cues which might be exploited for prediction.

In Section 6.2, we propose our MM-VAP model, a turn-taking model which integrates visual and audio features. In Section 6.3, we find MM-VAP achieves a higher performance than the audio-only VAP model on the Candor corpus. An ablation study reveals that facial expression contributes the most to the performance of MM-VAP. In Section 6.4, we show

examples of MM-VAP and VAP models on excerpts from the Candor corpus. We note MM-VAP predicts speaker transitions earlier and with higher confidence than VAP. We conduct a qualitative analysis of attention in MM-VAP and VAP models in Section 6.5, finding MM-VAP utilises more information from the previous turn than VAP.

Chapter 7: Towards real-world deployment of predictive turn-taking models

The final work chapter considers practical aspects of a real-world deployment of a trained PTTM. We first consider the impact of noise on performance. In Section 7.1, we outline a protocol for adding noise to the Candor corpus at controlled signal-to-noise ratios (SNRs). We use environmental noises that a model is likely to encounter once deployed. In Section 7.2, we show that both VAP and MM-VAP models are highly sensitive to noise. We find training with noisy data enables a multimodal PTTM, which includes visual features to better exploit visual cues, however, this does not always generalise to new types of noise. In Section 7.3, we conduct an analysis which considers whether the drop in performance in noise is attributable to the disruption of prosody, the masking of lexical information, or both.

A theme that emerges during the work of the thesis is that multimodality is exploited by turn-taking models. In Section 7.4 we return to the LLM models of Chapter 4, and propose a novel LLM based turn-taking model, Audio TurnGPT, which incorporates audio cues. We design a causal, multimodal fusion method to align the modalities of speech and text; highly heterogeneous cues with differing sample rates and dimensionalities. We find that Audio TurnGPT achieves a statistically significant performance over the state-of-the-art Pairwise TurnGPT model, illustrating the role of multimodality.

Finally, although a deployment to human-robot interaction is beyond the scope of this thesis, we use our extensive experience gained through working with PTTMs to discuss the practical challenges of a real-world deployment in Section 7.5.

Chapter 8: Conclusion and future work

This chapter concludes the thesis and brings together our findings from previous chapters. We outline the impact of our findings and identify a number of research directions for future work. We also discuss a number of works that also include visual cues in PTTMs.

1.6 Contributions

We summarise our contributions to knowledge as follows:

1. Conversational speech transcription

There is much recent interest in using ASR to automate transcription, driven by the wide availability of ASR platforms. Yet the literature lacks detail on conversational speech transcription with ASR and the practicalities of ASR use in research. Chapter 3 bridges this gap, answering Research Question 1. We note that commercial ASR outperforms open-source ASR on conversational speech, most notably on the transcription of disfluent speech. Our findings therefore deliver valuable insights to researchers wishing to use ASR in their work, and are published in our *Research Methods in Applied Linguistics* journal article (O'Connor Russell et al. 2024).

2. Training turn-taking models with ASR

In Chapter 4, we demonstrate that modern commercial ASR is suitably accurate for transcribing of data to train turn-taking models. This vastly expands the amount of available data, a finding noted in our *ACL Findings 2025* publication (O'Connor Russell and Harte 2025a). This not only supports ASR in our work, but also in future work machine learning work requiring transcribed conversations.

3. Visual cues enhance turn-taking prediction

As the overwhelming majority of work in PTTMs at the time of thesis planning utilised cues from speech, our novel multimodal model, MM-VAP, outlined in Chapter 6, comprises one of the first models to include visual cues. Earlier work considered small amounts of data, but our work comprises hundreds of hours of videoconferencing interaction, demonstrating a significant improvement in accuracy. An important finding was that the audio-only VAP model did perform quite well on videoconferencing speech, where interlocutors can see one another, despite not being able to access visual cues. These findings answer Research Question 4. Our MM-VAP model and analysis is presented in our *ACL Findings 2025* publication (O'Connor Russell and Harte 2025a).

4. Raising doubt about generalisation to new domains

Our cross-corpus analysis of the VAP model demonstrates that PTTMs trained on telephone speech performs poorly on videoconferencing speech and vice-versa. This raises doubt about the ability of these models to generalise to new interactions, which has not been noted in the literature. Given the reliance on telephone speech for training and the expectation that trained models will eventually be deployed to embodied human-robot interaction, this finding raises important questions. This work is presented in our *MMSYM 2024* workshop publication (O'Connor Russell and Harte 2024).

5. Demonstrating the role of multimodality in predictive turn-taking

We conduct a comprehensive analysis of prosodic and lexical cues in speech, which reveals that prosodic cues captured by the PTTMs play a vital role in their prediction performance. Our analysis in Chapter 4 shows the VAP model does not suffer from the errors originating from syntactic ambiguity experienced by LLM turn-taking models.

5.1. Exploring the role of prosodic and lexical cues

As there is difficulty in analysing self-supervised speech representations, which have become common in audio-based turn-taking. We therefore propose a new method of selectively isolating prosodic and lexical cues from speech in Chapter 5, which reveals that both support turn-taking. Our analysis of PTTMs performance in noise in Chapter 7 further uncovers the role of prosodic and lexical cues in prediction.

This work therefore addresses Research Question 3, and it appears in our *ICASSP 2026* publication (O'Connor Russell, Chaurau, et al. 2026).

The role that both prosodic and lexical cues play in turn-taking prediction is further reinforced by our Audio TurnGPT LLM in Chapter 7. Our analysis with this model reveals that incorporating speech into an LLM based turn-taking model improved performance, addressing Research Question 7a.

5.2. Visual cues help to overcome acoustic background noise

Our work in Chapter 7 represents what is, to our knowledge, the first consideration of PTTM performance in noise. These models will encounter diverse sources of noise once deployed. We show that MM-VAP can only exploit visual cues if some examples of noise are included during training. We also note that the enhanced performance does not necessarily generalise to new sources of noise, further adding to our aforementioned contribution on generalisation.

This work therefore addresses Research Question 7b. This work is presented in our *INTERSPEECH 2025* publication (O'Connor Russell and Harte 2025b)

Chapter 2

Literature review

This literature review introduces the key concepts in this thesis. Turn-taking is introduced in Section 2.1. A comprehensive review of studies of turn-taking in human conversation is provided in Section 2.2, where we identify a list of multimodal turn-taking cues that serve as a guide for analysis and modelling in later chapters. We discuss conversational speech corpora in Section 2.3. Our methodology for extracting speaking turns and multimodal turn-taking cues is described in Section 2.4.

In the second half of our review, we focus on modelling turn-taking with machine learning. Section 2.5 outlines the transformer architecture, which forms the backbone of modern predictive turn-taking models (PTTMs). Section 2.6 reviews the literature and state-of-the-art PTTMs. Finally, we summarise gaps in the literature that we have identified in our review in Section 2.7.

2.1 Turn-taking fundamentals

Conversation unfolds as a series of short speech segments called turns. Turn-taking is the process through which these conversations are organised. One of the first formalisations of turn-taking was by Sacks et al. (1974). In their framework, turns comprise multiple sentences or clauses. A pair of speaking turns is shown in Figure 2.1. The first speaker asks, ‘*John are you a student here at this university*’ and there is the response from Speaker 2, ‘*yes I am*’. There are multiple points where a turn is deemed to be complete by the listener in this example, which Sacks et al. termed transition-relevant places (TRPs). For instance, in this example, the following exchanges are also plausible:

1. John, are you a student? ⇒ Yes I am
2. John, are you a student here? ⇒ Yes I am
3. John, are you a student here at this university? ⇒ Yes I am

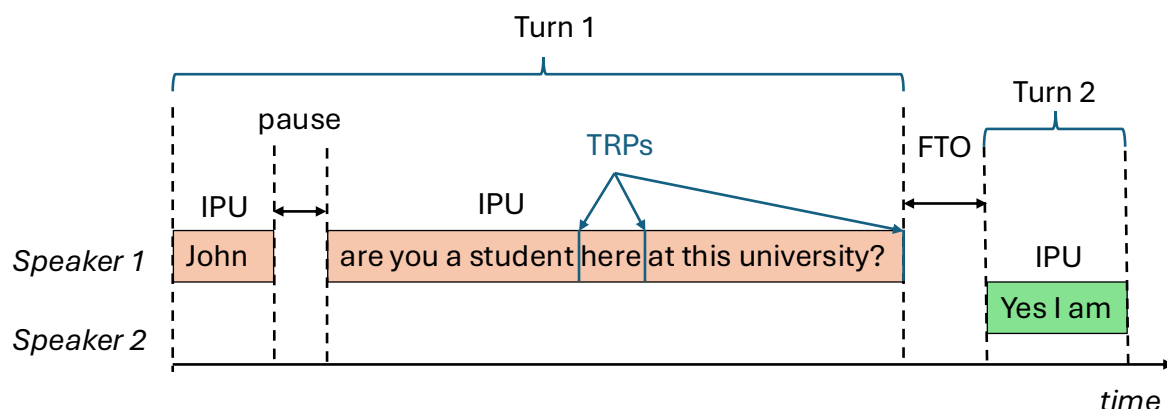


Figure 2.1 Illustrative example of an exchange in speaking turns. Turn 1 contains a pause, and three transition-relevant places (TRPs), where a transition between two speakers is possible due to grammatical completeness. The floor-transfer offset (FTO) indicates the time period between the end of the previous turn and the start of the next turn. The example is inspired by Bögels and Torreira (2015).

However, the listener speaks only after the word '*university*', despite it being grammatically correct for them to begin speaking earlier. The presence of a TRP therefore does not mean a speaker change will occur. According to Sacks et al. (1974), the presence of multiple TRP within a turn enables the speaker to *project* (i.e. predict) turn-endings ahead of time, with minimal silence (i.e. *gaps*) between speaking turns. Later studies of human interaction reveal a variety of multimodal cues from speech and non-verbal communication channels, such as prosody, intonation and facial expression facilitate this prediction (Bögels and Torreira 2015; N. Latif et al. 2018; Nota et al. 2023). This enables rapid and precise timing between speaking turns in conversation.

2.1.1 The timing of transitions between speakers

Given its ubiquity in human interaction, turn-taking has been widely studied. To motivate our review, in this section we first consider the timing of speaker transitions. This enables us to outline how turn-taking is a predictive process. We then review a number of corpus and behavioural studies of turn-taking in Section 2.2, and outline how prediction in turn-taking is supported by multimodal cues. Our review focusses on two-party interaction, where the vast majority of analysis and modelling work has been conducted.

Though Sacks et al. (1974) discussed how the gaps between turns in conversation are very short, their work did not directly consider the timing of transitions between speakers. Several studies have explored timing in great detail, enabling rich insights into cognitive processes behind turn-taking (Heldner and Edlund 2010; Levinson and Torreira 2015). These studies explore timing by first identifying continuous segments of speech, called interpausal units (IPUs), from transcripts. Typically, small silences of less than 180-200 ms are ignored

(Skantze 2021; Heldner and Edlund 2010). Speaker turns are obtained by combining one or more adjacent IPU's from the same speaker, not separated by speech from another speaker. The floor-transfer offset (FTO) is then computed to analyse timing at corpus-level (J.-P. De Ruiter et al. 2006). Let the end of the first turn be denoted t_1 , and the start of the next turn be denoted t_2 . The FTO is then given by:

$$FTO(t_1, t_2) = t_2 - t_1 \quad (2.1)$$

The FTO is positive where there is a *gap* between turns (Figure 2.1). A negative FTO indicates overlapping speech. The average FTO in English language corpora is between +100 and +300 ms (Heldner and Edlund 2010; Stivers et al. 2009; Roberts et al. 2015; Cavalcanti and Skantze 2025); i.e. a very small period of silence. In an analysis of multiple languages, Heldner and Edlund (2010) reported a mean FTO of 111-130 ms in Dutch, English and Swedish corpora. Stivers et al. (2009) found differences between the FTO of 10 languages, e.g. 230 ms for English but only 7 ms for Japanese. However, they found several language-independent features, such as faster responses to yes/no questions when the answer was positive.

2.1.2 Speaker transitions in corpus of English telephone speech

To illustrate turn-taking in conversation, we extract FTO values from the 260-hour Switchboard corpus of telephone speech. Turns are extracted as described above. Details of the corpus are provided further on in Section 2.3. We extract 153,171 speaker transitions (and hence FTO values). The histogram of FTO values is shown in Figure 2.2 (mean=169 ms, std=43 ms mode=118 ms). Although the mean and modal FTO are positive, overlapping speech is very common; a third (32%) of speaker transitions in the corpus occur in overlap (FTO<0 ms). These statistics are in line with Roberts et al. (2015) who also used the Switchboard corpus.

2.1.3 Listeners anticipate speaker changes ahead of time

The fast speed at which speaker transitions occur has prompted much discussion. Heldner and Edlund (2010) conducted a detailed study of timing in human conversation, finding 41–45% of transitions occur after silences greater than 200 ms. Noting that silences greater than 200 ms are noticeable to a listener (Walker and Trimboli 1982), they conclude that approximately half of all speaker transitions result from a listener *reacting* to silence at the end of speaking turns.

However, the reactive interpretation of turn-taking has been disputed. Reaction time studies consistently demonstrate that at least 600 ms is needed to formulate basic, single-word utterances (Indefrey 2011; Indefrey and Levelt 2004). As turn-taking studies consistently note 100-300 ms between speaking turns, including our histogram in Figure 2.2, most researchers

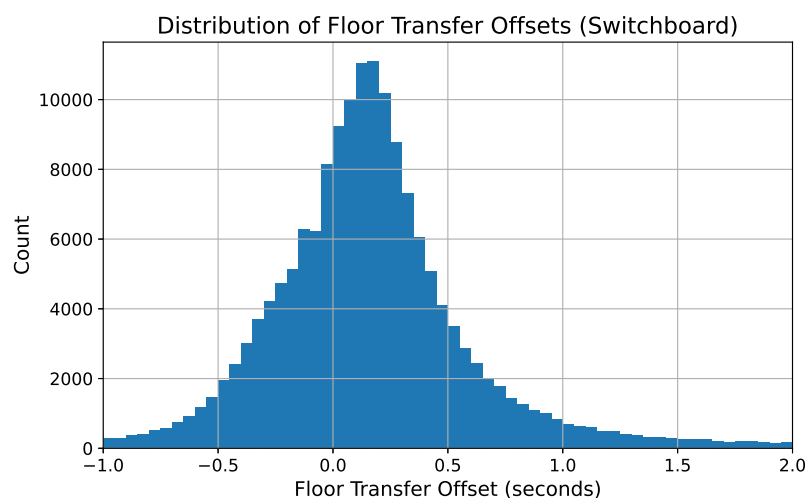


Figure 2.2 Histogram of floor-transfer offsets values computed from the Switchboard telephone speech corpus (Godfrey et al. 1992). Mean=169 ms, Std=43 ms Mode FTO=118 ms.

therefore conclude that turn-taking is predictive (Levinson and Torreira 2015; Holler, Kendrick, et al. 2016). In other words, in many cases, interlocutors must continually anticipate upcoming opportunities to speak ahead of time, as opposed to reacting to the cessation of speech.

It follows that robots must also predict their turns ahead of time in order to support fluid interaction. Skantze and Irfan (2025) found participants preferred to interact with a robot with predictive turn-taking capabilities than a basic model that spoke only after silences. PTTMs are outlined in Section 2.6 and form the basis of Chapters 4-7.

2.2 A review of the turn-taking literature

Given the remarkable speed at which speaker transitions occur, much research has been devoted to understanding turn-taking. We first review corpus studies of turn-taking, which analyse patterns in communication that commonly precede turn-taking events.

2.2.1 Corpus studies of multimodal turn-taking cues

An early corpus study of turn-taking was conducted by Duncan (1972), who analysed video recordings of interviews in US English. Falling intonation was found to precede a transition from one speaker to another, whereas a rising or level intonation indicated the current speaker would continue. Syntactically complete sentences occurred when the speaker yielded to the listener, whereas incomplete sentences were indicators that the current speaker would continue. Inhalation and changes in posture often preceded the onset of speech. Based

on these observations, Duncan outlined how conversation is not random, rather it is an organised and orderly process, governed by the following ‘*signals*’:

1. signals produced by the speaker to signal an intent to **yield** to the listener
2. signals produced by the speaker to signal an intent to **continue** speaking
3. signals produced by the listener to signal an intent to **start** speaking

Duncan therefore concluded that turn-taking arises from a coordinated exchange of these signals; i.e., patterns in verbal and non-verbal communication. However, later work by Hjalmarsson (2011) cautions that although turn-taking appears rules-based, signals associated with turn-taking are not necessarily produced consciously. She therefore prefers the term *turn-taking cues*, defined as “*perceivable phenomena relevant for turn-taking, regardless of whether they are conscious or not*” (Hjalmarsson 2011, pg. 25). We follow this terminology, as it more clearly outlines how patterns in speech and non-verbal behaviour associated with turn-taking are not necessarily produced knowingly by the speaker or the listener. There is a similar inconsistency in the literature surrounding the terminology used to refer to speaker changes. Some studies refer to changes in the active speaker as *shifts* (Skantze 2021), whereas others prefer the term *yields* (Heldner and Edlund 2010; Gravano and Hirschberg 2009; Duncan 1972). We choose the term *shift* as it does not assume that the speaker is intentionally handing over the floor. As most studies refer to silences with no change in speaker as *holds* (Gravano and Hirschberg 2009), we also use the term *hold* in our work.

Further corpus studies of turn-taking have identified numerous turn-taking cues from speech. Wennerstrom and Siegel (2003) noted a rise in intonation, level intonation, and syntactic completion were associated with shifts in the active speaker. In an analysis of US English, Gravano and Hirschberg (2011) found in the last 500-1000 ms of a turn, fundamental frequency (F_0) was decreased relative to mid-turn speech. The decrease was more prior to shifts than before holds. Intensity (elevated before holds, decreased before shifts) and voice quality features (elevated jitter and shimmer before shifts) also showed distinct patterns. Transcriptions of conversations reveal that sentences are almost always complete before a shift, but not necessarily before holds (Wennerstrom and Siegel 2003; Gravano and Hirschberg 2009). Certain lexical features, such as filled pauses like ‘*uhm*’ and ‘*ehm*’, are almost exclusively found before holds (Wennerstrom and Siegel 2003).

A number of studies have also identified non-verbal turn-taking cues. Kendon (1967) noted in English conversation, speakers tend to avert their gaze while speaking, especially at the start of a turn. On the other hand, speakers tend to look back at the listener at the end of a turn. He interpreted these as cues which regulate the interaction, being indicators of an intent to continue speaking or defer to the listener (Kendon 1967). Similarly, Duncan found eye contact was a strong signal of an upcoming shift, whereas averted gaze was strongly associated with a current speaker continuing (Duncan 1972). Studies have also explored

facial expression. Nota et al. (2021) analysed Dutch conversations, finding blinks, squints, gaze shifts and eyebrow frowns occurred at the start of questions. They therefore concluded they serve as an early visual indicator of an upcoming question, which necessitates a listener response (Nota et al. 2021). Similarly, Flecha-García (2010) found eyebrow raises mark the beginning of utterances in English (Flecha-García 2010), echoing earlier work by Cavé et al. (2002) who found eyebrow raises in French occur at the beginning of turns (Cavé et al. 2002).

2.2.2 Behavioural studies of multimodal turn-taking cues

Corpus studies therefore reveal the role of prosody and syntactic cues in turn-taking. Certain facial expressions and non-verbal behaviours, such as gaze, also indicate that the current speaker will continue speaking or defer to the listener. However, as corpus studies assess correlations in recorded interactions, they cannot reveal whether or not these cues are actually relied upon by humans. Behavioural studies address this by establishing whether specific cues are actually used by humans during turn-taking. Experiments typically involve asking participants to judge whether an utterance is complete. Verbal and non-verbal behaviour is then manipulated, and any impact on the listener's response is recorded.

Prosody has been an important focus for behavioural studies of turn-taking. In an early study, Cutler and Pearson (1985) gave participants short utterances in English and asked them to read them aloud. Subjects then listened to the recordings and were to judge whether the samples came from the middle or the end of a turn. A fall in pitch was judged by the participants as the end of a turn, whereas a rise in pitch was associated with judgements of mid-turn speech. However, they acknowledged that their use of read speech was a major limitation, as it was unlikely to elicit the same prosodic variation as conversational speech (Cutler and Pearson 1985, pg. 152). Later work by J.-P. De Ruiter et al. (2006) used spontaneous, conversational speech by recording pairs of Dutch-speaking participants. After transcription, 108 turns were selected. Participants reliably predicted the end of utterances by pressing a button. When the F_0 contour was removed from the utterance by flattening, there was no significant effect on prediction accuracy. The recordings were then low-pass filtered to remove intelligibility, whilst preserving F_0 . As participants could no longer reliably judge when a turn ended, they concluded that prosody is not sufficient for turn-taking, whereas syntax is both necessary *and* sufficient (J.-P. De Ruiter et al. 2006, pg. 531).

However, Bögels and Torreira challenged these findings, arguing that prosody plays an essential role in turn-taking (Bögels and Torreira 2015). Dutch speakers were recorded producing sentences deliberately constructed to be syntactically ambiguous (Section 2.1). Long and short versions of the sentences were recorded (e.g. short *So you are a student?*; long *So you are a student here at Radboud University?*) and then spliced together to manipulate the prosodic contour. Participants pressed a button when they judged the utterance complete. When the manipulated utterances were played, participants often

pressed the button early, at the point of syntactic completion. This did not occur in the condition where the full, unmodified utterance was played. This shows participants exploit prosody to overcome the ambiguity in the lexical channel. These findings do not contradict J.-P. De Ruiter et al. (2006); rather, they suggest that certain cues become necessary under particular conditions; in this case, syntactic ambiguity. They therefore argue that studies of turn-taking must carefully consider the interplay of multiple cues Bögels and Torreira (2015, pg. 55). Finally, in a behavioural study of Swedish listeners, (Hjalmarsson 2011) found falling intonation and semantic completeness were associated with shifts, whereas flat intonation, semantic incompleteness and inhalation were associated with holds, consistent with the findings of other studies.

These studies reveal listener sensitivity to a wide range of turn-taking cues previously identified from corpus research, namely prosody (F_0 , intensity), syntax, disfluencies and other phenomena such as inhalation. Behavioural studies have also investigated the role of visual turn-taking cues. N. Latif et al. (2018) captured audio and video recordings from friends engaging in a casual English language conversation. They played recordings to participants and asked them to press a button as soon as they judged an utterance complete. Participants were significantly more accurate at identifying end-of-turns when presented with both audio and video. This was followed by a second experiment, where participants were shown clips with either the video or audio removed from the last portion of an utterance (200-600 ms). Participants who had access to both audio and video correctly identified shifts up to 600 ms earlier (N. Latif et al. 2018, pg. 36–38), illustrating a role in visual cues in supporting turn-taking.

While N. Latif et al. considered all facial signals together, Nota et al. (2023) conducted a fine-grained analysis by using avatars to manipulate facial expression. The avatars spoke utterances sourced from a Dutch corpus that were either questions or statements, and participants were asked to press a button if they thought the utterance was a question. The results showed questions with eyebrow frowns were identified both more accurately and more quickly than the same utterances without eyebrow movement. Earlier and longer eyebrow movements lead to faster responses Nota et al. (2023, pg. 15-16). Hence, eyebrow frowns served as early turn-taking cues, priming the listener that they are about to be posed a question and hence should respond in the near future. This reinforces their earlier corpus study, which identified eyebrow movement early in an utterance as a salient turn-taking cue (Nota et al. 2021).

Therefore, both audio and video cues support turn-taking, and behavioural studies corroborate the findings from previously discussed corpus studies (Section 2.2.1). Visual cues play a complementary role, enabling speakers to make faster and more accurate turn-taking decisions, with information early on in the utterance speeding up prediction.

Table 2.1 Summary of studies from our review in Section 2.2 that identify turn-taking cues. Due to differences across studies, they are grouped by the specific turn-taking event that the study analyses, namely: hold (speech prior to silence with no speaker change), shift (speech prior to silence with a speaker change), question/statement (whether the utterance is a question or statement), mid-turn speech, and the end and start of an utterance. Behavioural indicates a listener judgement experiment was used. Languages EN=English, SV=Swedish, NL=Dutch and FR=French.

Language + Event Analysed	Method	Study	F0	Intensity	Voice quality	Syntax	Misc.	Visual	Inhalation
HOLD									
EN-US	Conv. analysis	Duncan (1972)	rising/level					gaze aversion	✓
SV	Behav.+stat.	Wennerstrom & Siegel (2003)				synt. inc.			
EN-US	Statistical	Gravano & Hirschberg (2011)	fall (small)	increase	↓ jitter/shimmer	synt. inc.	uhm, eh		
EN-GB	Conv. analysis	Kendon (1967)						gaze aversion	
SV	Behavioural	Hjalmarsson (2011)	flat			sem. inc.			✓
SHIFT									
EN-US	Conv. analysis	Duncan (1972)	falling					eye contact	
SV	Behav.+stat.	Wennerstrom & Siegel (2003)	rising/level + rising F0			synt. comp.			
EN-US	Statistical	Gravano & Hirschberg (2011)	fall (large)	decrease	↑ jitter/shimmer	synt. comp.		eye contact	
EN-GB	Conv. analysis	Kendon (1967)							
EN-US	Behav.+stat.	Cutler et al. (1985)	falling pitch			synt. comp.			
NL	Behavioural	De Ruiter et al. (2006)				synt. comp.			
NL	Behavioural	Bögels et al. (2015)	F0 contour			synt. comp.			
SV	Behavioural	Hjalmarsson (2011)	falling			sem. comp.			
QUESTION / STATEMENT									
NL	Corpus	Nota et al. (2021)						blinks, squints, gaze, brows	
NL	Behav. (avatar)	Nota et al. (2023)						eyebrows	
UPCOMING RESPONSE (YES/NO)									
NL	Behav. (avatar)	Emmendorfer et al. (2025)						brows, gaze, head	
MID-TURN									
EN-US	Behav.+stat.	Cutler et al. (1985)	rising F0						
END-OF-UTTERANCE									
—	Classifier	Ferrer et al. (2003)	F0				pause duration		
EN-US	Statistical	Gravano & Hirschberg (2011)	falling F0					visual cues	
EN-US	Behavioural	Latif et al. (2018)							
START-OF-UTTERANCE									
—	—	Flecha-García (2010)						eyebrow raise	
FR	—	Cavé et al. (2002)						eyebrow raise	
FR	—	Cavé et al. (1996)						eyebrows	

2.2.3 The interplay between verbal and non-verbal turn-taking cues

A number of questions arise from corpus and behavioural studies of turn-taking, discussed here.

Defining turn-taking modalities Although the turn-taking literature is in agreement that turn-taking is a multimodal phenomenon encompassing cues for across speech, prosody, and visual communication, there is no standard terminology to refer to these modalities. Prosody refers to the non-lexical aspects of speech such as the F_0 contour that encode a wide range of non-verbal information such as emotion (Biron et al. 2025). Prosody is also important for turn-taking (Bögels and Torreira 2015). In this thesis, we adopt the term *verbal* to refer to the words spoken, i.e. lexical information. We adapt the term *non-verbal* in reference to both visual turn-taking cues such as gaze, but also in reference cues which arise from speech but cannot be considered words, such as as breathing, F_0 and intensity.

Non-verbal turn-taking cues: a natural consequence of speech? Given the presence of both verbal and non-verbal turn-taking cues, a question that arises is whether non-verbal turn-taking cues simply arise from the mechanics of speech production. Certain cues, such as inhalation, are intrinsically linked with the onset of speech. Therefore, inhalation arises as an inevitable consequence of speech production; however, it also serves as an indicator of an upcoming turn-taking event, as it typically precedes a turn (Duncan 1972).

There is evidence that other turn-taking cues, including eyebrow movement, are partially independent of speech production. Cavé et al. (1996) found that although eyebrow movement and variation in F_0 were linked in the majority (71%) of cases. In a third of cases, eyebrow movement was not related to F_0 variation, and rapid eyebrow movement also occurred during silences. They therefore concluded that eyebrow movement can function as an independent turn-taking cue that does not arise purely as a consequence of speech production or F_0 modulation. Therefore, certain non-verbal turn-taking cues, such as inhalation, occur as a consequence of speech production whilst also functioning as turn-taking cues and other cues, such as eyebrow movement, may function as independent indicators of upcoming turn-taking events. The interplay between verbal and non-verbal turn-taking cues is therefore richly complex.

The interplay between verbal and non-verbal cues Another important question is how they interact with and influence each other. J. P. Trujillo, Levinson, et al. (2022) recorded pairs of Dutch speakers engaged in a videomediated conversation. A blur was applied to reduce the participants' ability to see one another. As the blur intensity increased, gesturing and body movement became less pronounced. This suggests participants were subconsciously

aware gesture was ineffective, so they gestured less. However, vocal intensity also increased, demonstrating that verbal communication is altered to compensate for a lack of visual cues. There were, however, no reliable changes in $F0$ and speech rate (J. P. Trujillo, Levinson, et al. 2022). Verbal and non-verbal cues are therefore partially interrelated. Multimodal communication is therefore flexible, as interlocutors adapt to the availability and reliability of cues.

Does more cues not lead to more cognitive load? Lastly, the presence of multiple, co-occurring cues suggests there is a lot of information for the speaker to process. Many of these turn-taking cues might be redundant, increasing cognitive load. However, empirical findings show that the more cues that are present, the easier it is for interlocutors to predict turn-taking. Hjalmarsson found that when there were two or three turn-taking cues were significantly faster than when there was only a single turn-taking cue (Hjalmarsson 2011). Emmendorfer and Holler (2025) found in an avatar study that facial movements early in utterances, namely eyebrow movements, gaze shifts, and head movements, help listeners to anticipate upcoming information. Importantly, combinations of facial signals produced the most robust differences in responses, as opposed to responses with individual turn-taking cues. Therefore, the presence of multiple, co-occurring turn-taking cues across multiple modalities facilitates, rather than hinders, turn-taking.

2.2.4 Summary of multimodal cues

Our review has outlined that turn-taking is a fast-paced process. Listeners subconsciously exploit various turn-taking cues from the speaker and display their own cues prior to speaking. Turn-taking system is richly multimodal, involving acoustic, prosodic, lexical and non-verbal cues. These cues are partly interdependent, yet they can have independent functions in the turn-taking process. They therefore must be considered together, within the wider framework of multimodal conversational interaction. Furthermore, the system is highly adaptive, with speakers conveying and responding to cues differently depending on the interaction environment. Multiple cues create a high degree of redundancy in the system, and certain cues serve complementary roles until ambiguity arises: e.g. the role of prosody in resolving syntactic ambiguity.

Table 2.1 summarises the findings of our review, listing each turn-taking cue reported in the surveyed studies. The rich multimodality of turn-taking has important implications for PTTMs, which typically do not utilise any non-verbal cues. A further issue is that they are often trained on telephone speech but deployed in embodied interactions where multimodal cues are available. We return to these questions in Section 2.6, where PTTMs are outlined. This list is also not exhaustive, and does not cover multiparty interaction, where multimodal cues such as gaze redirection and shifts in posture are used to select the next speaker

(Petukhova and Bunt 2009). The complexities of multiparty turn-taking are briefly discussed further on in Section 2.6.10.

2.3 Corpora for turn-taking research and model development

Having reviewed turn-taking, we now turn our attention to corpora. Conversational corpora are essential for analysing turn-taking and training turn-taking models. We begin by providing context on corpus research, highlighting some important considerations in corpus studies.

2.3.1 Conversational speech corpora

The methodology of Conversation Analysis originated in the 1960s, and concerns the study speech produced within the surrounding context of previous and upcoming speaking turns (Goodwin and Heritage 1990). Linguistic corpora are essential for this work. In Conversation Analysis, corpora should be collected from natural interactions, as discreetly as possible. This methodology has been employed to understand diverse phenomena from mother-child (Nomikou 2023) to patient-clinician interactions (Shaw et al. 2020). However, as the data come from real-life interaction, corpora from these studies are therefore typically not made publicly available.

In the last several decades, corpora recorded in more controlled settings have become publicly available. These corpora are designed to ensure maximum utility for future research (Ruhi and Schmidt 2014). In the RoomReader corpus, participants follow a low-stakes, casual game-show task (Reverdy et al. 2022), reducing the privacy concerns of public dissemination. However, the use of task-based interaction such as games inevitably reduces authenticity, as were it not for the experiment, the interaction would not have happened (Goodwin and Heritage 1990). Furthermore, the use of scenario-based tasks shapes the turns, actions, and behaviours that unfold (Goodwin and Heritage 1990). At the same time, task-based interactions are easier to study as real-world dialogue is highly unstructured (A. H. Anderson et al. 1991).

2.3.2 Available English language corpora for turn-taking model development

With these considerations in mind, we conduct a review of publicly available corpora, with attention both to general-purpose conversational datasets and to those that have been used specifically in turn-taking. We then select three corpora used in all the work of this thesis. We focus on English language corpora in our review.

Task-based versus unstructured interaction Noting that natural dialogue is challenging to constrain, A. H. Anderson et al. used the Map Task scenario (G. Brown et al. 1983) to

elicit controlled, yet spontaneous interaction (A. H. Anderson et al. 1991). 128 unscripted two-party interactions (64 participants) were recorded in English. In each interaction, both participants receive a map. Only one participant receives a route and must guide the other participant along the route. 20 hours of transcribed audio recordings are publicly available, and the corpus has been extensively utilised in turn-taking prediction (Roddy, Skantze, et al. 2018b; Roddy, Skantze, et al. 2018a; Skantze 2017). However, only audio is provided. An influential audio-visual corpus using a task-based scenario the AMI Meeting corpus (Kraaij et al. 2005). Groups of 4 interlocutors take part in mock English language meetings (over 100 hrs). Participants design a new product, playing the roles of project manager, UX designer, and so on. Audio and video are recorded. Another example is the RoomReader corpus (Reverdy et al. 2022), recorded in English on the Zoom videoconferencing platform. Unlike AMI, the participants play a game where one participant guides 2 others through a collaborative task. Participants must discuss and agree on the top three most frequent responses to a question previously asked by 100 people, ranked by popularity. For example:

- **Question:** “Name something people are often chased by in movies.”
- **Correct responses (ranked):** monsters, cars, cops

In the corpus, interactions are unscripted, and collaboration is encouraged. The dialogue is therefore less constrained by the task than in other corpora such as Map Task. The RoomReader corpus scenario is based on the earlier Multisimo corpus, which uses the same task for prompting interaction, but the interactions take place in-person, with participants seated around a table (Koutsombogera and Vogel 2018). An issue with the use of these tasks for corpus elicitation is that, although the dialogue is unscripted, the task still exerts a strong influence over the dialogue. The interaction is likely quite predictable from lexical information related to the task itself, rather than turn-taking cues from speech. Furthermore, the total amount of data that is available is limited, as participants were required to be physically present in a controlled environment.

In contrast, the Switchboard corpus of telephone speech is much larger, and dialogue is less constrained (Godfrey et al. 1992). Pairs of participants join a call and discuss a given topic in English, e.g. *nursing homes*, *air quality*. 260 hours of transcribed audio is provided across 2,438 interactions (mean session length=6 mn). The use of a prompt elicits spontaneous speech. In comparison with Map Task, conversations are less structured. Other English language telephone corpora include the Fisher corpus (Cieri et al. 2004, 2,000 hours). These telephone speech corpora have been extensively used to study turn-taking (Stivers et al. 2009; Heldner and Edlund 2010; Cavalcanti and Skantze 2025) and also in turn-taking model development (Ekstedt and Skantze 2022b; Roddy and Harte 2020). The provided manual transcriptions make them readily usable, but the use of a telephone means interlocutors must rely on turn-taking cues from speech, and naturally, no video recordings are available.

Multimodal conversational corpora Multimodal corpora enable speech and non-verbal interaction to be studied. In the NoXi corpus, pairs of interlocutors in separate rooms interacted via large screens enabling body language, facial expression and other non-verbal cues to be captured (Cafaro et al. 2017). Interlocutors discuss prescribed topics in English, German or French. A similar corpus is the Mahnob mimicry corpus (Bilakhia et al. 2015), consisting of 54 dyadic interactions recorded in-person in English. Both Noxi (Onishi et al. 2024) and Mahnob (Roddy, Skantze, et al. 2018b) have been used in multimodal turn-taking model development; however, there is far less data available for training of PTTMs in comparison with audio-only models: e.g. 11-10 hrs in NoXi and Mahnob versus 260 hrs in Switchboard.

Videoconferencing enables more data to be collected as participants do not have to be physically present in the same room. Müller, Sood, et al. (2020) scraped 90 hours of interviews conducted via videoconference in English from YouTube. The corpus has been used to study gaze aversion (Müller, Sood, et al. 2020), but has not been used for turn-taking. An issue with interviews is that turns are typically assigned by the host, through questions (Greatbatch 1988), making turn-taking very deliberate and controlled. Another corpus using videoconferencing interaction is the Candor corpus. Pairs of English-speaking interlocutors join a videoconferencing call (Reece et al. 2023). The interactions are completely unprompted, and 50 hours of audio and video recording are captured across 1,657 sessions.

Audio-only PTTMs have been trained on large corpora such as Switchboard (Ekstedt and Skantze 2022b; Inoue, Bing'er Jiang, et al. 2024; Roddy and Harte 2020). On the other hand, the development of multimodal models has been constrained to smaller corpora such as the 11 hr Mahnob corpus, in very limited studies (Roddy, Skantze, et al. 2018b). The Candor corpus has not been used for training PTTMs.

2.3.3 Selecting corpora for our work

Based on our review, we select three corpora for use in this thesis. For our study of transcription in Chapter 3, we require corpora with manual transcriptions to assess automatic speech recognition (ASR). We choose **Switchboard** as it is a large corpus of US English conversational interaction provided with a manual transcription and a word-level alignment. To include diversity of accents, we also choose to include **RoomReader** as most speakers in the corpus are native speakers of Irish English, and the corpus is provided with a manually corrected ASR transcription.

Next, we select corpora to train and test PTTMs. These models are covered in detail in Section 2.6, though for now we note these models are based on deep learning, a data-intensive task. Hence, a key concern is that corpora must be of two-party interaction and suitably sized to support deep learning. As work has been conducted with Switchboard (Ekstedt and Skantze 2022a; Inoue, Bing'er Jiang, et al. 2024; Roddy and Harte 2020), a 260 hr long corpus, we use 260 hrs as a rough guideline for an appropriately sized for modern

PTTM development. This leaves **Switchboard** and **Candor** as the corpora that are suitable for this task. The presence (Candor) and absence of (Switchboard) of non-verbal turn-taking cues in the corpora make them ideally suited to answering our research questions in Chapter 1. They support our work in analysis and model development across Chapters 4-7. Although the Mahnob corpus has been used for PTTM training (Roddy, Skantze, et al. 2018a), its size is small at 11 hrs and therefore unlikely to support PTTM training with modern deep learning models, which are typically trained on hundreds of hours of conversation (Ekstedt and Skantze 2022b; Li et al. 2022).

We provide the key details of these corpora below. A summary of the corpora we have covered in our review is provided in Table 2.2 along with further corpora not discussed for space limitations. We also note that since the time this work has been conducted, new corpora have been made available, including the 4,000 hr *Seamless Interaction Dataset* corpus from Meta (Agrawal et al. 2025). We discuss recent advances in the field in Chapter 8.

RoomReader The RoomReader corpus (Reverdy et al. 2022) consists of audio and video recordings of conversational interaction which take place via the Zoom¹ videoconferencing platform. A still from the corpus is shown in Figure 2.3a. There are 30 small-group tutor-led English language sessions, typically with 4–5 participants (mode = 4). The participants follow the previously described game and join the Zoom videoconferencing call from a quiet location of their choosing using personal devices. The corpus provides audio, video, transcriptions, and diarised audio, totalling 8.5 hours. The mean session duration is 17 minutes. The mean participant age is 23 years (std = 5.2). In total, the corpus includes 115 participants (50 male, 65 female), of whom 89 (37 male, 52 female) are native speakers of English. Irish-English is the dominant native dialect, spoken by 63 participants (24 male, 39 female). There are two tutor participants (one male, one female), both native Irish-English speakers, each of whom leads 15 of the 30 sessions.

Speaker-diarised audio (48 kHz) and video (60 fps) are provided. Transcripts were obtained from Microsoft ASR and then corrected by a linguist, who ensured all lexical information was corrected and preserved, including disfluencies and filled pauses.

Candor The Candor corpus (Reece et al. 2023) consists of 1,656 two-party videoconferencing interactions, each comprising casual, unprompted conversation in US English. There is 850 hours of recordings in total (mean session length = 34 mns). Although interactions are not in-person, interlocutors can see one another. Videoconferencing provides a front-facing camera angle that is ideal for visual feature extraction. Our review did not identify a corpus of in-person interaction of this size.

¹<https://zoom.us/meetings>

Table 2.2 Summary of publicly available corpora of conversational interaction

Corpus	Hours	Multiparty	Multimodal	Recording setup	Sessions	Participants	Language	Scenario	Notes	Paper
RoomReader	8.5	X	X	Videoconference	30	18	EN	FamilyFeud	Spontaneous, small-group	Reverdy et al. (2022)
Multisino	2.7	X	X	In-person	18	828	EN	FamilyFeud	Spontaneous, small-group	Koutsombogera and Vogel (2018)
Candor Corpus	850	X	X	Videoconference	1656	828	EN	Unscripted	Unscripted, prompted	Reece et al. (2023)
Switchboard	240	X	X	Telephone	2400	543	EN	Unscripted, prompted	Unscripted, task-based, different rooms	Godfrey et al. (1992)
AMI	100	X	X	In-person	30	26	EN	Product design	Unscripted, task-based, different rooms	Kraaij et al. (2005)
Cardiff Conversation	5	X	X	Rig	40 (EN)	26	EN	None	Unscripted conversation	Aubrey et al. (2013)
NoXI	11 (EN)	X	X	Videomediated	63	216	EN, FR, DE	Novice-expert	Rig recording, unscripted conversation	Cafaro et al. (2017)
Teams corpus	47	X	X	In-person	22	78	EN	Board game	Cooperative	Litman et al. (2016)
MPiGroupInteraction	7.3	X	X	In-person	121	2 + guest	DE	Discussion points	Multimediate challenge paper	Müller, Huang, et al. (2018)
DEyeADic	74	X	X	Videoconference	128	54	EN	Interview	YouTube	Müller, Sood, et al. (2020)
Map Task	20	X	X	In-person (with and without eye contact)	128	54	EN	Map Task	10 hours with eye contact	A. H. Anderson et al. (1991)
Mahndb Mimicry	11.6	X	X	In-person	54	43	EN	Unscripted, prompted	Dinner party	Blakchia et al. (2015)
CHiME	30.74	X	X	In-person	43	43	EN	Unscripted	Dinner party	Christensen et al. (2010)



(a) RoomReader

Figure 2.3 Stills from the RoomReader corpus illustrating the use of Zoom videoconferencing software and recording setup of each corpus.

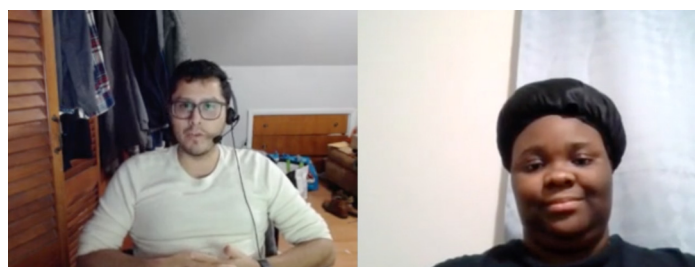


Figure 2.4 A still from a session of the Candor corpus illustrating the diverse nature of recording conditions, camera angles, and recording devices.

Interlocutors were free to use a device and location of their choosing (e.g. home offices) and to discuss whatever topics they wished. As the corpus was collected during the COVID-19 pandemic, participants frequently discussed aspects of their daily lives, including the pandemic, work, and sports. The recording conditions were uncontrolled, resulting in diverse environments, camera angles, and devices. Recordings are in 60 fps video and 48 kHz stereo audio with one channel per speaker, and an ASR transcription is provided.

Switchboard The Switchboard corpus consists 260 hour Switchboard corpus of two-party US English interactions (Godfrey et al. 1992). In each interaction, the participants discuss a prescribed topic, e.g. office attire (mean session length 6 min 23 s). Stereo audio is provided at 16 kHz, with one channel per speaker. There is a slight channel bleed, with the other speaker faintly audible on the channel of one speaker. A manual transcription is provided, with extensive labels of linguistic phenomena. Word and phoneme label alignments are also provided.

2.4 Methods for turn-taking analysis and model evaluation

In this section, we outline the methods that we will use for analysis of turn-taking and the training of predictive turn-taking models (PTTMs). We first outline transcription of corpora, an important step in PTTM development. We then describe how we automatically extract speaking turns from transcribed interactions. Finally, we describe how we automatically extract key audio-visual features associated with the turn-taking cues for analysis.

2.4.1 Transcription of conversational interactions

A transcript is a written record of who said what during an interaction. Typically, transcriptions are time-aligned at the utterance and sometimes word-level, providing rich detail on the timing of turns and speaker transitions. Today, most transcription is conducted manually, by experts. This can take up to 14 times longer than the conversation itself (Umair et al. 2022). Given the time and cost involved in manual transcription, there is much interest in using ASR to automate transcription across diverse research domains from healthcare (K. Latif et al. 2023) to criminology (Loakes 2024).

PTTM development and training are heavily dependent on transcription. Naturally, PTTMs, which utilise text, leverage the lexical content of the transcription itself (Ekstedt and Skantze 2020; Leishman et al. 2024). However, where the PTTM uses audio alone and does not use text, transcription is still relied upon to provide the timing of the start and end of speaking turns for model training and development (Ekstedt and Skantze 2022b; Roddy and Harte 2020). To the best of our knowledge, PTTM model development has thus far relied on manual transcription. This vastly constrains the scale and amount of data available for training and development, particularly multimodal datasets, where manually transcribed data are scarce. Automating transcription with ASR therefore has the potential to generate large amounts of data for PTTM development, but has not been explored to date.

We therefore offer a discussion on the various fields in which ASR is gaining traction as a tool for automated transcription of conversation, identifying a number of gaps in the literature. In Chapter 3, we assess how state-of-the-art ASR can support transcription. Our work addresses these gaps in the literature, offering timely and essential insights to researchers who need to transcribe dialogue in a research setting. Furthermore, this work is vital in supporting the work in Chapters 4-7, where we use ASR to transcribe corpora in order to analyse turn-taking, and to develop and analyse PTTMs.

2.4.2 What is a transcription?

As transcripts are used in many different fields, requirements are highly domain-specific. Some researchers work with transcripts purely to analyse what was said, e.g. criminologists

analysing interview audio in a police investigation (Loakes 2024). Computer scientists also work with transcriptions to train machine learning algorithms in diverse areas from audiovisual speech recognition (Shi et al. 2022) to turn-taking (Roddy, Skantze, et al. 2018b). Here, transcription requirements are primarily lexical, i.e. writing down what was said and when. Another group of researchers who use transcription include linguists who are interested in aspects of conversation itself, e.g. conversational repair (Schegloff 2000b). This requires careful coding of aspects of conversation such as disfluent speech, backchannels, and so-on.

Given the diversity of fields in which transcriptions are used, specific requirements vary from one domain to another (Psathas and T. Anderson 1990). At a minimum, a transcript should capture what was said and by whom. Non-lexical tokens (NLTs) - short words such as ‘*uhm*’ and ‘*ehm*’ - should be preserved, as they often carry important meaning. For instance, K. Latif et al. (2023) note in clinical settings, patients often reply with NLTs like *mh-hum* to questions like “*Are you allergic to any antibiotics*”? However, NLT transcription is challenging as there is no finite set of NLTs (Ward 2006). Depending on the conversation modality (e.g. telephone or in-person), non-verbal utterances such as head nodding to signal agreement may need to be transcribed (Parks and Bennett 2024). Repair and hesitation markers may also need to be included (Widodo 2014), as may the start and end times of words.

2.4.3 Automating transcription with ASR

Given the time involved in manual transcription, there is much interest in using ASR to automate the task, at least in part (Umair et al. 2022). Commercial providers have been around for some time, though cost may have limited uptake in research. In recent years, open-source ASR tools such as Whisper (Radford, Kim, et al. 2023) have made state-of-the-art speech recognition widely available, leading to interest in automating transcription.

ASR is receiving much attention for medical transcription (K. Latif et al. 2023; Han et al. 2024; Ciampelli et al. 2023). Han et al. (2024) used Whisper to automatically transcribe patient-clinician interactions, enabling medical notes to be automatically queried. Ciampelli et al. (2023) demonstrated the use of ASR transcriptions in mental health diagnoses. Qualitative research is another area where ASR is becoming commonly used, bypassing manual transcription (Wollin-Giering et al. 2024; Kashefi et al. 2024). In law enforcement, Field et al. (2023) explored using ASR to transcribe body camera recordings. Other areas where ASR is finding application include the transcription of archival television footage (Y. Yang 2023). Finally, in corpus linguistics, corpora have been released with ASR (Reece et al. 2023) or hand-corrected ASR transcripts (Reverdy et al. 2022).

The widespread application of ASR suggests that researchers have a high expectation of its transcription capability. This is reinforced by commercial ASR providers who promote ASR as a highly accurate, general-purpose tool for automated transcription (Speechmatics, Ltd. 2024). But is ASR sufficiently accurate to support these types of transcription? ASR

accuracy is typically quoted using the word error rate (WER), which rewards correct spelling and word placement. A lower WER indicates superior performance. State-of-the-art ASR systems such as Whisper achieve a 2-3% WER on audiobook speech (Radford, Kim, et al. 2023), outperforming humans (Lippmann 1997). However, ASR performance is poorer on conversational speech, with WER rates between 10-30% (Baumann et al. 2017; Liesenfeld et al. 2023). The higher WER reflects the challenging nature of conversational speech. A diverse range of factors also impact the WER such as background noise, non-native speakers, and different accents (Alharbi et al. 2021; Sanabria et al. 2023; Loakes 2022).

An ASR transcription is generally considered *ready-to-use* when the WER is less than 10% (Microsoft, Inc. 2024). However, the WER metric does not detail the types of errors that occur. Liesenfeld et al. (2023) found that ASR removed overlapping speech, a common feature of casual conversation. Stolcke and Droppo (2017) found that ASR confuses *uh* and *uh-huh* whereas humans transcribers are good at recognising this difference. Lopez et al. (2022) found that NLTs are common in human transcripts of conversation, but absent from ASR transcripts. The mistranscription of NLTs may be a deal-breaker: K. Latif et al. (2023) found that NLTs such as *uhmm* were transcribed with 94-98% error by Google and Amazon ASR in patient-doctor interactions. As these words can carry important clinical meaning, ASR cannot be used to automate medical notetaking tools, irrespective of the overall WER.

2.4.4 Gaps in the ASR literature

There is evidence that commercial ASR outperforms open-source ASR on conversational speech (Morbini et al. 2013; Speechmatics, Ltd. 2024). However, to the best of our knowledge at the time of writing, there is no study which compares the transcription of NLTs across commercial and open-source ASR platforms. As much of the current ASR literature focusses on the WER, other aspects of transcription quality remain understudied. Furthermore, ASR transcription quality should be considered within the context of the available resources, skill set and budget, yet current literature lacks these insights.

ASR has not been utilised in PTTM model development training and development. The use of manually transcribed corpora is typical, e.g. with Switchboard (Roddy and Harte 2020; Ekstedt and Skantze 2022b; Inoue, Bing'er Jiang, et al. 2024). For most multimodal corpora, manual transcriptions are not available. Hence, ASR transcription has the potential to vastly expand the amount of data available for PTTM training and development, but to our knowledge, this has not been at the time of thesis planning. We return to these questions in Chapters 3 and 4.

2.4.5 Extraction of speaking turns and the voice activity signal

Having outlined transcription, we turn our attention to PTTM training and evaluation. We first use the word timings from either a manual or ASR transcription to obtain IPU by combining segments of speech from a single speaker that are separated by no more than 200 ms of silence (Heldner and Edlund 2010). We then combine IPUs from the same speaker, not separated by speech from the other speaker, to form speaking turns. Earlier work by Walker and Trimboli (1982) demonstrated that silences that are less than 200 ms are not noticeable to listeners, and this threshold has become standard in the turn-taking literature (Heldner and Edlund 2010). Note that turns from *different* speakers may have silences less than 200 ms between them, as only IPUs from the same speaker are combined.

Once speaking turns have been identified, we use the start and end times of speaking turns to extract a binary stereo voice activity detection (VAD) signal with the same number of samples as the audio track, where 1=speaking and 0=not speaking. The VAD signal is then used in the training and evaluation of PTTMs, and the extraction of turn-taking events as detailed in Section 2.4.6. For a speaker s , the voice activity at frame n is given by:

$$v_s[n] = \begin{cases} 1, & \text{if } n \in [\lfloor f_s \cdot t_{i,\text{start}}^{(s)} \rfloor, \lfloor f_s \cdot t_{i,\text{end}}^{(s)} \rfloor] \text{ for any IPU } i \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

Where s is the speaker index ($s \in [0, 1]$), f_s is the sampling rate and $t_{i,\text{start}}$ and $t_{i,\text{end}}$ are the start and end times of any IPU in extracted from either a manual or ASR transcription.

2.4.6 Extraction of turn-taking events

In the predictive turn-taking literature, models are evaluated on a subset of turn-taking events within a corpus such as holds and shifts, rather than evaluating the model at every frame of speech (Skantze 2017; Roddy, Skantze, et al. 2018a; Ekstedt and Skantze 2022b). We follow this methodology by extracting *holds* and *shifts* transcribed corpora, enabling the analysis and the evaluation of trained models in later chapters.

Holds To extract holds we first identify all silences > 250 ms from the VAD signal (Equation 2.2). Next, we identify silences where the active speaker is the same before and after the silence. To ensure that false starts are not counted as holds, we further limit the holds by ensuring that there is only one active speaker 1 second before and after the start of the silence (Ekstedt and Skantze 2022b). Figure 2.5a depicts a hold in a two-party interaction. These mid-turn silences are sometimes referred to as *pauses* (Heldner and Edlund 2010).

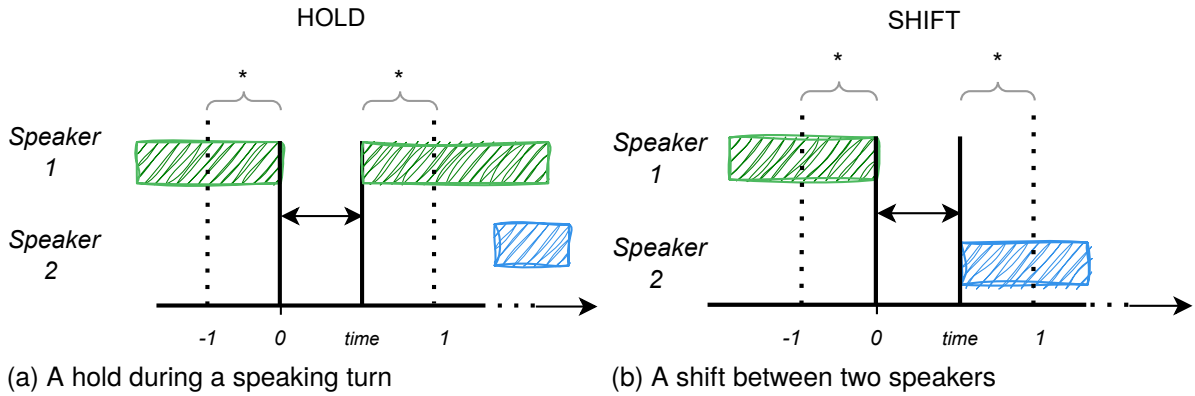


Figure 2.5 Shift and hold schematic, where *=regions before/after the event where only a single speaker can be active.

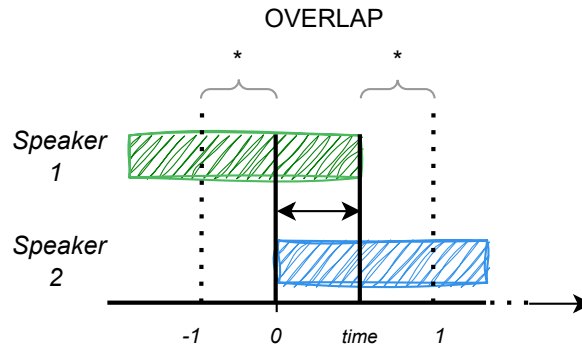


Figure 2.6 An overlap between two speaking turns, where *=regions before/after the event where only a single speaker can be active.

Shifts Next, to extract shifts we identify speaker transitions or shifts which take place with at least 250 ms of silence between adjacent speaking turns; i.e. an FTO > 250 ms. We limit our analysis to shifts where, in the 1 second prior to the start of the shift, and in the 1 second after the silence, only 1 speaker is active. This ensures rapid changes in speaker and short interjections are not captured. This is because we wish to capture sustained transitions from one speaker to another and not short interjections such as backchannels (Ekstedt and Skantze 2022b). Figure 2.5a depicts a shift in a two-party interaction.

Overlap Another important evaluation to consider is overlapping speech (Roddy, Skantze, et al. 2018a), which frequently occurs in natural conversation (Section 2.1.2), but is often omitted from PTTM evaluation (Ekstedt and Skantze 2022b). We identify speaker transitions which occur with periods of overlap. The overlaps must be at least 250 ms in duration; i.e. an FTO < -250 ms. Again, we ensure that only a single speaker is active in the 1 second immediately before and after the overlap. We show a schematic of a speaker transition with an overlap in Figure 2.6.

2.4.7 Extraction of prosody and voice quality features

Having outlined transcription and the extraction of turn-taking events, we now describe how we extract prosodic and voice quality features from audio recordings of conversational corpora. Our review of turn-taking in Section 2.2 identified intensity, $F0$, and jitter and shimmer as important verbal turn-taking cues. We describe their extraction from audio recordings here, which is relevant for Chapter 5, where we conduct analysis of audio turn-taking cues.

Intensity Intensity is a physical property of a waveform, reflecting the short-term sound energy of the signal. Perceptually, listeners experience intensity as loudness, which correlates with intensity (Uppenkamp and Röhl 2014). Intensity is expressed in decibels (dB), and can be measured directly from speech, whereas loudness requires listener tests. Using audio recordings from a corpus of conversational speech, we extract intensity in decibels with the `to_intensity` method of the Python Parselmouth library (Jadoul et al. 2018). Intensity is computed separately for each speaker (i.e., each channel of the stereo recordings) using a time step of 5 ms and a minimum pitch of 75 Hz.

Fundamental frequency, $F0$ Pitch is an important part of sound perception, but it is not directly measurable from a waveform. Pitch is the perceptual correlate of fundamental frequency ($F0$), which can be estimated from speech (Oxenham 2012). F_0 is the lowest frequency component of a periodic sound and can be directly measured from a waveform. It is reported in Hz. We measure $F0$ from speech using the `to_pitch_ac` method of Parselmouth (Jadoul et al. 2018). We use a floor of 75 Hz and a of 600 Hz, and a time step of 5 ms. It is defined only for voiced frames of speech, which contain vowels and other voiced sounds.

Jitter Next, we consider voice quality features. Voice quality captures aspects of speech related to phonation, such as breathy and creaky voice, which have important pragmatic roles (Gobl and Chasaide 2003), yet are often overlooked (Lameris et al. 2025). There is evidence that they play a role in turn-taking (Gravano and Hirschberg 2009). The first voice quality feature we extract is jitter, which is the per-cycle variation in the period of $F0$, defined as follows (Farrús et al. 2007):

$$\text{Jitter} = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N} \sum_{i=1}^N T_i} \quad (2.3)$$

Where i is the i_{th} frame, T_i is the period, equal to the i_{th} estimate of $F0$, It is a unitless quality. We extract jitter using $F0$, which is extracted as described above. Hence, it is defined only for voiceless frames of speech. We compute the mean jitter within short windows of speech located at key turn-taking events, as outlined in Chapter 5.

Shimmer The second voice quality feature that we consider is shimmer, which captures variation in the amplitude. The formula is (Farrús et al. 2007):

$$\text{Shimmer} = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N} \sum_{i=1}^N A_i} \quad (2.4)$$

Where i is the i_{th} frame and A is the amplitude, which is computed from the decibel intensity ($A = 10^{\frac{I}{20}}$, where I is the intensity). Like jitter, we compute shimmer within short windows of speech. Our analysis is outlined in Chapter 5.

2.4.8 Extraction of visual features

Our review of the turn-taking literature identified a number of visual cues associated with turn-taking, such as gaze direction and facial movement. We extract visual features from recordings of conversational interaction, which enables us to analyse their association with turn-taking, and to incorporate them into a multimodal PTTM in Chapter 6. We briefly review the available methods for feature extraction, prior to outlining the specific features we extract with the OpenFace toolkit (Baltrušaitis, Robinson, et al. 2016).

Methods for analysing visual cues from video To analyse non-verbal turn-taking cues from video, the movement of the face and body must be quantified. Classic computer vision methods focus on capturing motion across frames of video, e.g. optical flow (Horn and Schunck 1981). These methods capture motion across frames of video and can therefore be used to track movement of the body (Lennon et al. 2006). However, they are challenging to interpret as motion is captured in a general sense, e.g. motion of the head or body.

Methods which enable a more fine-grained analysis include active appearance models that localise and individually track the face, jaw, lips and brow. They outperform classic methods, even when parts of the face are occluded (Sterpu and Harte 2017). An open-source tool which performs facial landmark detection and tracking is OpenFace (Baltrušaitis, Robinson, et al. 2016). The model detects and tracks head pose, facial landmarks, and also estimates facial action units (FAUs). FAUs are a set of objective descriptions of facial movement (e.g. jaw drop) traditionally labelled by trained experts (Ekman and Friesen 1978). Other tools include MediaPipe (Lugaresi et al. 2019) and OpenPose (Qiao et al. 2017), which detect full body pose and key landmarks e.g. elbow and hand position. These methods have been used to analyse non-verbal interaction such as gesture (J. P. Trujillo, Vaitonyte, et al. 2019). They are also useful features for machine learning tasks such as lip-reading (Sterpu and Harte 2017), although more recently they have been superseded by learned representations such as AVHubert, which operate on raw pixels (Shi et al. 2022). Learned representations are now the dominant approach in modelling, but interpretability issues mean

determining which cues are utilised by a trained model is highly challenging (e.g. gaze, facial expression, or head movement).

We wish to analyse non-verbal turn-taking cues and understand how they are exploited in a trained model. We therefore choose a landmark detection-based approach over learned representations as this balances the ease of interpretability with the richness of the extracted features. We chose OpenFace as it provides a rich set of facial features for analysis and machine learning. Although it does not detect other regions of the body such as the hand and torso, in our work, we focus on the face. This is because we work with videoconferencing data where regions of the body beyond the face are not consistently visible (Figures 2.4 and 2.7).

Feature extraction OpenFace extracts a wide range of features from the face and head (Baltrušaitis, Robinson, et al. 2016). We select a subset of these features for use in our work:

- **Facial action units (FAUs)** These are 17 features which numerically describe facial movements e.g. jaw drop, on a scale from 0.0 to 5.0. There is one gaze vector per eye, which is a 3 dimensional unit vector.
- **Head position** We also extract head position (X, Y, Z in mm) and head rotation (roll, pitch and yaw in radians).
- **Facial landmarks** We select 15 facial landmarks located on the brow, jaw, nose and lips (x and y in pixels)
- **Confidence** This is a scalar confidence score which ranges from 0 to 1 and measures the degree of confidence of the model tracking

We extract one vector per frame of video, with 60 dimensions in total. We show a sample visualisation of OpenFace features in Figure 2.7. Head pose is depicted as a blue cube centred on the head, pointing in the estimated direction of head pose. Eye gaze vectors are depicted in green. Facial landmarks are shown in red, though we only use a subset of 15 in this work. It is not straightforward to visualise FAUs, so these are omitted from the visualisation. We provide a complete list of the FAUs in Table 2.3, along with short descriptors.

2.5 Machine learning architectures and evaluation

In this section, we outline the transformer architecture, which forms the backbone of modern PTTMs. This lays the groundwork for the final sections of the review, PTTMs and their evaluations are described.

Table 2.3 Facial action unit codes, with descriptions and sample images. Images sourced from (Tian et al. 2001; Valstar et al. 2017). descriptions from (Baltrušaitis, Robinson, et al. 2016).

Action Unit Number	Descriptor	Image
AU01	Inner brow raiser	
AU02	Outer brow raiser	
AU04	Brow lowerer	
AU05	Upper lid raiser	
AU06	Cheek raiser	
AU07	Lid tightener	
AU09	Nose wrinkler	
AU10	Upper lip raiser	
AU12	Lip corner puller	
AU14	Dimpler	
AU15	Lip corner depressor	
AU17	Chin raiser	
AU20	Lip stretched	
AU23	Lip tightener	
AU25	Lips part	
AU26	Jaw drop	
AU45	Blink	

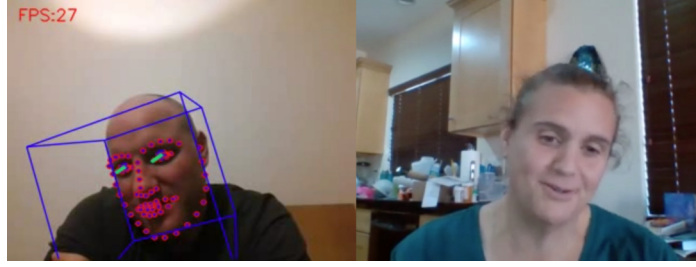


Figure 2.7 Still from Candor with OpenFace features shown for the left participant

2.5.1 The transformer architecture

The transformer (Vaswani 2017) is a machine learning architecture for sequence modelling. It has been at centre of recent advancements in speech and language processing (Raiaan et al. 2024), including turn-taking (Leishman et al. 2024; Ekstedt and Skantze 2022b) where it has replaced earlier models with recurrent architectures (Skantze 2017; Roddy and Harte 2020).

At the heart of the transformer model is the attention mechanism. Consider a sequence $X = X_0, X_1 \dots X_{N-1}$, for instance, of words in a sentence or frames of speech. In the transformer, the sequence is projected into a query (Q), key (K), and value (V) tensors:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \quad (2.5)$$

Where W_Q, W_K and W_V are learned during training. Next, the attention is computed between the query and keys:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \quad (2.6)$$

where d_k is the dimensionality of the key vectors and the softmax is applied row-wise. The output of the attention mechanism is then given by

$$Y = AV, \quad (2.7)$$

The output Y is a weighted sum of the value vectors V . The attention weights A are in turn computed from the similarity between queries and keys. A visualisation of the attention mechanism is shown in Figure 2.8. Note how the attention facilitates information flow from the input X to the output Y . Note that in this example, the attention is causally masked. This means that at time t , the output Y_t cannot utilise future inputs X_i where $i > t$. This is typical in many tasks, including turn-taking, where only information from the past should be utilised to make predictions about current or upcoming behaviour (Ekstedt and Skantze 2022b).

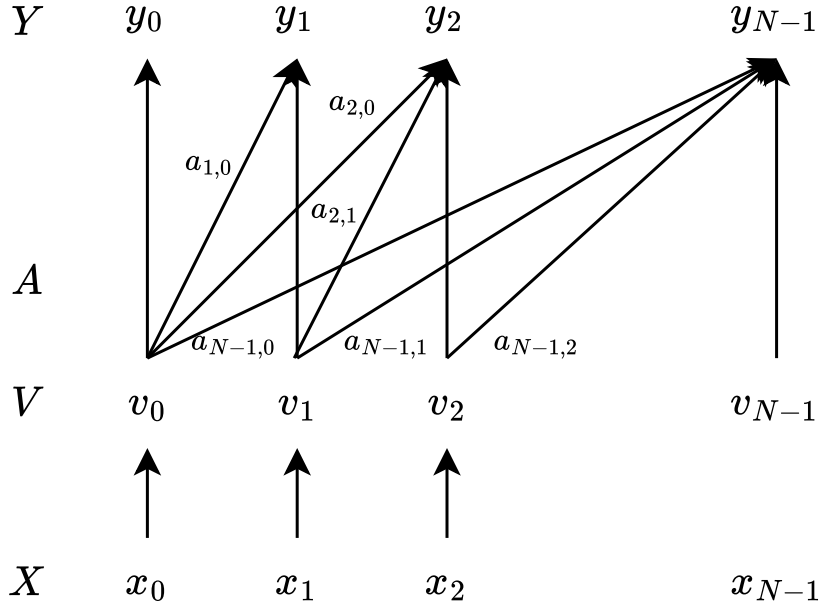


Figure 2.8 Visualisation of transformer attention

Like all modern machine learning algorithms, the transformer is trained via backpropagation over batched input and output pairs. Both supervised and unsupervised learning can be used (Vaswani 2017; Riviere et al. 2020; Baevski et al. 2020). Most PTTMs predict the upcoming speaking activity of interlocutors, described in detail in Section 2.6.

2.5.2 Self and cross attention

The original transformer model comprised an encoder and a decoder (Vaswani 2017). The encoder processed an entire sequence of input, and the decoder generated an output, e.g. a translation of an English input into a French output. There are two types of attention in the model: self-attention and cross-attention.

Self-attention learns patterns within a single sequence. In this example, words in the English sentence or words in the French sentence. In a self-attention layer, the query (Q), key (K) and value (V) are equal, i.e. $X = Y$ in Equation 2.6. Cross-attention learns patterns from one sequence to another, e.g., from the French translation to the original English sentence. The query (Q) is the English sentence, and the key (K) and value (V) are the French sentence. In PTTMs, self and cross attention have been exploited to learn patterns in speech, and to combine information from the two speakers in an interaction, as described further on (Inoue, Bing'er Jiang, et al. 2024; Leishman et al. 2024).

2.5.3 The GPT-2 architecture

An influential variant of the transformer is GPT-2 (Radford, Wu, et al. 2019), which only uses the decoder part of the original transformer model. In the original GPT-2 paper, both the input and output are the same sequence of tokens, with the output being the input shifted forward by 1 time step. In other words, the model is a language model as it continually predicts the next token. Each element in the output is only allowed to use earlier tokens in the input; i.e. it is causally masked, as we have previously described. Other minor adjustments of the original transformer model include layer normalisation and the scaling of weights in residual connections (Radford, Wu, et al. 2019).

The GPT-2 model refers to the large language model described in the paper (Radford, Wu, et al. 2019), which is open-source and publicly available². The GPT-2 architecture is the basic structure of the model, which has become widely used in PTTM architectures.

2.6 Predictive turn-taking models

In this section of our literature review, we discuss PTTMs. We review the modelling literature, defining the distinction between reactive and predictive models. We then outline the voice activity projection (VAP) model (Ekstedt and Skantze 2022b) (audio) and TurnGPT models (Ekstedt and Skantze 2020; Leishman et al. 2024) (text). We choose these models as they are state-of-the-art audio and text-based turn-taking models identified in our review. They directly support the work of Chapters 4 and 5 and lay the necessary groundwork for our multimodal voice activity projection (MM-VAP) and Audio TurnGPT models, introduced in Chapters 6 and 7. We also provide details on evaluation metrics for trained PTTMs.

2.6.1 Predictive versus reactive turn-taking models

As highlighted in our introduction in Chapter 1, most commercial systems react to silence after speaking turns. This leads to human-robot interactions which fail to replicate the fast-paced nature of human-human conversation. Predictive models anticipate speaker transitions ahead of time, like humans. Predictive and reactive models are outlined in Figure 2.9. Consider a point in time t during an interaction between two interlocutors I_a and I_b . If I_a has been speaking at some point in time before time t :

- A **reactive model** predicts whether interlocutor I_a (or I_b) has stopped speaking at the current time t using the conversational history H up to and including t .

²<https://huggingface.co/openai-community/gpt2>

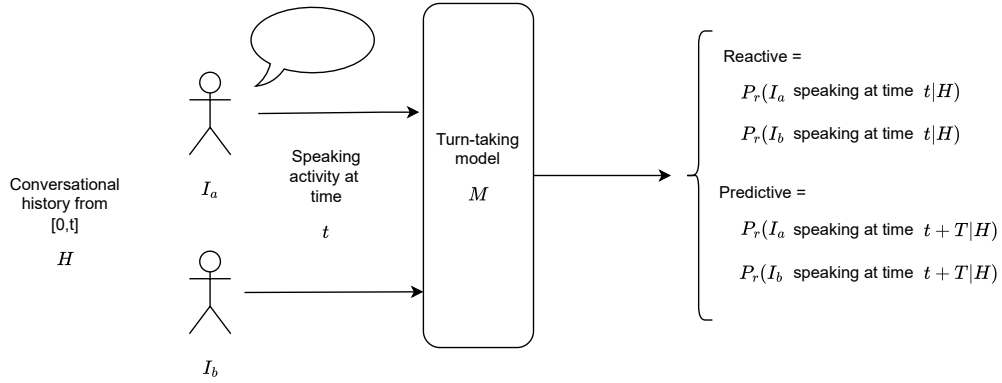


Figure 2.9 Schematic of reactive and predictive turn-taking models (PTTMs). A conversational interaction between two interlocutors I_a and I_b is depicted. At time t , the model M makes predictions using information from conversation history up to time t denoted H . A reactive model outputs the probability that an interlocutor I_a has stopped speaking at time t . A predictive model can make several possible predictions about upcoming changes between speakers at some future point in time $t + T$.

- A **predictive model** outputs the probability that interlocutor I_a (and/or I_b) will be speaking at future point in time $t + T$ where $T > 0$ using the conversational history up to and including t .

There are slight variations from PTM to PTM (Roddy and Harte 2020; Li et al. 2022), but most predictive models follow this basic format of predicting upcoming speaking activity at or within a window of future time (Skantze 2021).

2.6.2 Predicting future speaking activity

An early PTM was proposed by Skantze (2017). Recognising the limitations of reactive models, Skantze proposed to use an recurrent neural network (RNN) to predict upcoming speaking activity continuously while speech is ongoing, and to not wait for the silence at the end of a turn. The model is shown in Figure 2.10. It predicts at each time t the probability of one of the two interlocutors speaking in the next 3 seconds. Predictions were made at a rate of 20 Hz (every 50 ms).

The model was trained on the Map Task corpus, which we previously discussed in Section 2.3.2. The model employed feature engineering, where a set of features was extracted from speech. Features from audio included the binary VAD labels (see Section 2.4.5), $F0$ and intensity. Part-of-speech (POS) tags were extracted from the transcription, capturing syntactic aspects of speech linked to turn-taking. The VAD labels were used to generate the binary training labels, which capture future speaking activity and are shown in Figure 2.10.

The model's ability to predict holds and shifts was evaluated during pauses at least 500 ms long. It achieved an $F1$ score of 0.763, outperforming a majority hold baseline

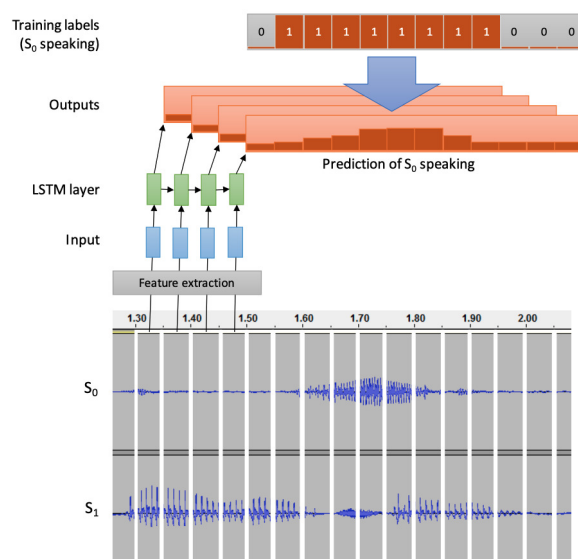


Figure 2.10 A schematic of the recurrent neural network (RNN) model proposed by Skantze (2017), reproduced from their paper. The model continuously predicts upcoming speaking activity if a single interlocutor is in a two-party interaction.

(0.421). The model was also evaluated at the onset of speech segments, and performed well in distinguishing between short (< 3 second) and long (> 3 second) segments of speech ($F1=0.786$). This shows that the model was able to capture turn-taking behaviours such as next speaker detection and whether or not the current speaker will continue speaking; important considerations for spoken dialogue systems (Skantze 2017; Skantze 2021). By training models on subsets of features, Skantze found prosody (i.e. $F0$, intensity and spectral stability) contributed significantly more to prediction performance than POS (POS) tags.

The model was further developed by Roddy, Skantze, et al. (2018a) on the Map Task corpus. An extra word embedding feature was included, and the more comprehensive set of GeMAPS acoustic features was used (Eyben et al. 2015), which includes loudness, shimmer, $F0$, jitter and spectral features inducing MFCCs. They similarly found that acoustic features outperformed linguistic features, showing that prosodic cues are the predominant cue learned by the model. The evaluation was expanded to include overlapping speech, which is very common in human conversation (Section 2.1.2)

2.6.3 Alternative approaches to predictive turn-taking

Predicting future speaking activity is just one approach taken in predictive turn-taking. Other approaches include the work of Li et al. (2022), who proposed a model to directly predict the time until the next speaker change on the Switchboard corpus. Their model was based on an RNN that combines audio features and word embeddings, trained to output a probabilistic

distribution over lead-in times to the next turn. The audio features include features extracted from the pre-trained wav2vec (version 1.0) model (Schneider et al. 2019), along with root mean square energy, pitch, and $F0$. Word embeddings were obtained from the pre-trained GPT-2 model (Radford, Wu, et al. 2019). The mean average precision was reported, which makes their work challenging to interpret and compare with other studies.

Their work echoes earlier work by Roddy and Harte (2020). The Switchboard corpus was modelled as a human-machine interaction, where the first turn was treated as the *user turn* and the subsequent turn as the *system response*. An RNN-based model was trained to generate distributions over lead-in times until the next response. The model and output are shown in Figure 2.11, where the predicted distribution over response times in Switchboard is compared with the true distribution. Acoustic features included voice activity, energy, a log-Mel filter-bank, the $F0$ track, and acoustic energy. Linguistic features consisted of embeddings encoding both the user turn and the response. A notable difference between their work and that of Li et al. (2022) is that the model is not causal: the prediction assumes the text of the system response is known. This is highly relevant for prediction, as there are strong timing differences between different types of response e.g., in yes/no questions, a ‘yes’ response is faster on average than a ‘no’ response (Kendrick and Torreira 2015). However, this is a strong assumption that may not be possible to achieve in a human-robot interaction deployment, particularly given the latency of spoken dialogue systems.

2.6.4 Incorporating visual cues in a predictive turn-taking model

Few published PTTMs have used non-verbal cues. Roddy, Skantze, et al. (2018b) incorporated gaze features extracted with OpenFace into a multimodal PTTM. The PTTM was based on a similar task of predicting upcoming speaking activity (Skantze 2017; Roddy, Skantze, et al. 2018a); however, architectural modifications were introduced to enable the model to handle the differing frame rates of acoustic and gaze features. This highlights a common challenge in multimodal machine learning, namely the fusion of heterogeneous features with differing temporal resolutions in a single model (Baltrušaitis, Ahuja, et al. 2018). The Mahnob mimicry corpus was used in their work (Section 2.3).

Their results showed that a model incorporating both acoustic and gaze features outperformed an acoustic-only model. However, this was only true when the model utilised their proposed architectural modifications, which handled different frame rates of acoustic and non-verbal cues; a naive concatenation of gaze and acoustic features performed the same as an audio-only model. The magnitude of the increase was, however, very slight (0.88 vs. 0.87). A further caveat is that the Mahnob mimicry corpus is very small, at 11 hours. Furthermore, only gaze was used, with other non-verbal features such as head pose, facial expression and body movement ignored.

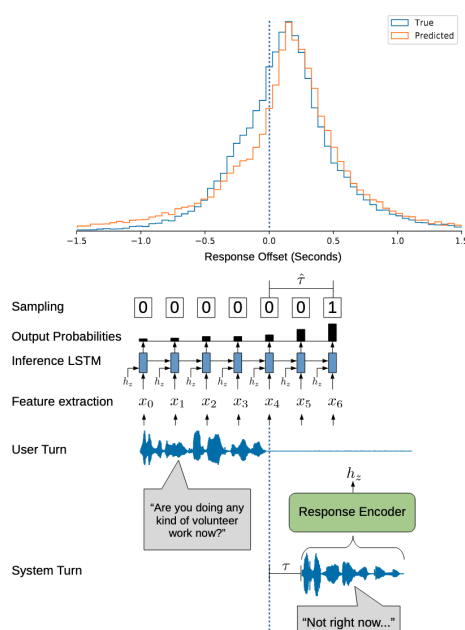


Figure 2.11 A schematic of the predictive turn-taking model presented by Roddy and Harte (2020). The model uses an RNN and outputs a probabilistic distribution over possible speaker transition times given acoustic and linguistic features. The model assumes that the text of the next turn (the response) is known.

At the time of thesis planning, this work was the only work including visual cues into a PTM. Given the small size of the corpus used, it is therefore unknown whether visual cues can support interaction at scale. Furthermore, it is also unknown if the numerous non-verbal turn-taking cues features beyond gaze identified in our review, such as facial expression, can support prediction (Section 2.2).

2.6.5 The voice-activity projection turn-taking model

A recent audio-only PTM is the VAP model proposed by Ekstedt and Skantze (2022b). The model is based on the GPT-2 architecture. Unlike earlier models, the VAP model jointly predicts the speaking activity of both participants in a two-party interaction. In this section, we outline the model architecture and the training labels used. The VAP model is evaluated in Chapter 4, and the basic architecture is utilised in our multimodal voice activity projection model introduced in Chapter 6.

2.6.5.1 Training objective

Like models discussed earlier in this section, the VAP model continually predicts upcoming speaking activity. However, the speaking activity of both interlocutors is predicted by repre-

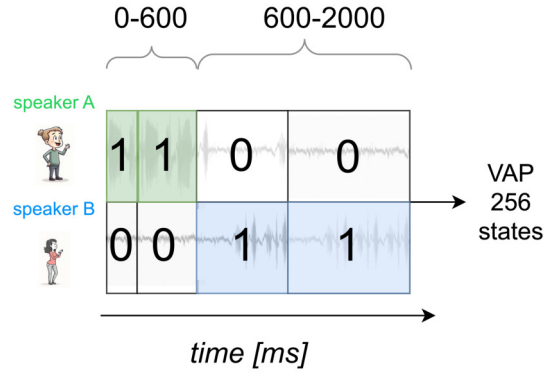


Figure 2.12 Illustration of the voice activity projection (VAP) model training objective, showing 1/256 possible states of possible speaking activity in the next 2 seconds

senting the upcoming speaking activity of both interlocutors as one of 256 possible states, as illustrated in Figure 2.12. At each point, we use the VAD signal from Section 2.4.5 to generate a training label. The label consists of bins representing the next 2 seconds of speaking activity. The bins span the next 0-200, 200-600, 600-1200 and 1200-2000 milliseconds. If 50% of frames within a bin contain speech, we set the bin label to 1, otherwise the bin is labelled 0.

The Switchboard corpus is used for training and evaluation. The word-level phonetic alignment provided with the manual transcription is used to generate the VAD labels (Ekstedt and Skantze 2022b). Note that the model does not use the tokens from the transcription in any training or evaluation of the VAP model, only the time-alignment is used for label generation. As there are 8 bins, there are $2^8 = 256$ possible VAP states.

2.6.5.2 The contrastive predictive coding model

The VAP model architecture is depicted in Figure 2.13. It takes as input a stereo audio file and assumes there is one speaker per channel, as is the case in the Switchboard corpus. The audio is 16 kHz stereo.

Features are extracted from the audio using a pre-trained feature extractor, called the CPC model, which is from Riviere et al. (2020). The model weights are loaded from their checkpoints, and they are not updated further during the training of the VAP model. The CPC model is a sequence-to-sequence neural network model. The model's encoder is a convolutional neural network which extracts information from the input audio signal $\mathbf{x}_t$, where t represents a frame in the input audio. The sequence is then modelled with a 1 layer gated recurrent unit neural network, generating an embedding $\mathbf{z}_t$. The model is then trained by unsupervised learning. The task for supervised learning is given the current embedding $\mathbf{z}_t$, predict the next K embeddings $\mathbf{z}_{t+1}, \mathbf{z}_{t+2}, \dots, \mathbf{z}_{t+K}$. In this way, the model learns to model

the temporal and phonetic information contained within in the input speech signal, without requiring any labels (Riviere et al. 2020). The CPC model is pre-trained on the LibriSpeech audiobook corpus (Panayotov et al. 2015b; Riviere et al. 2020). It is therefore an example of a self-supervised speech representation (S3R) model. The pre-trained weights are loaded, and the front-end is subsequently frozen during training of the VAP model; i.e. the audio front-end is not optimised in any way during training. There is one 256 dimensional feature vector extracted in the CPC front-end for every 20 ms of audio (i.e. 320 samples or 50 Hz).

The use of the CPC model for feature extraction means that unlike earlier work, which used engineered features such as $F0$ and log-Mel spectrograms (Roddy, Skantze, et al. 2018a), representations are learned. Through training on large volumes of unlabelled data, which leads to useful representations being generated for many downstream tasks, including turn-taking. These representations have become widespread in many speech and language processing tasks due to their ability to capture rich representations of acoustic and linguistic phenomena (Pasad et al. 2021; Gimeno-Gómez et al. 2025).

2.6.5.3 Model architecture

In its initial formulation, the model relied on ground-truth voice activity during inference, effectively allowing it to know when turns had ended (Ekstedt and Skantze 2022b)). This dependency was subsequently removed in later work, which also extended the model to handle stereo audio (Inoue, Bing'er Jiang, et al. 2024). This is the version used in our experiments and is depicted in Figure 2.13. We re-implement the model with reference to the authors' publicly available code and analyse its performance in detail in Chapters 4, 5, and 7.

After the initial audio feature extraction (CPC in Figure 2.13), there are transformer layers organised into blocks. The first block is a self-attention block (SelfAttnBlock). The self-attention block consists of a single transformer layer given by the following equation:

$$\text{SELFATTNBLOCK}(x) = \text{TRANSFORMERLAYER}(Q = xW_Q, K = xW_K, V = xW_V) \quad (2.8)$$

In self-attention, the same input sequence x is used to compute the learned query, key and value representations, i.e. $Q = xW_Q$, $K = xW_K$, and $V = xW_V$. The model therefore learns temporal patterns in the input signal through self-attention, as previously described in Section 2.5. In the model shown in Figure 2.13, there is one block per speaker, and the weights are shared (tied) between speakers.

The self-attention blocks are followed by cross attention blocks (CrossAttnBlock). There is one self-attention block per speaker, but weights are shared (tied) between speakers. The cross attention blocks integrate information across both speaker channels x_1 and x_2 as follows:

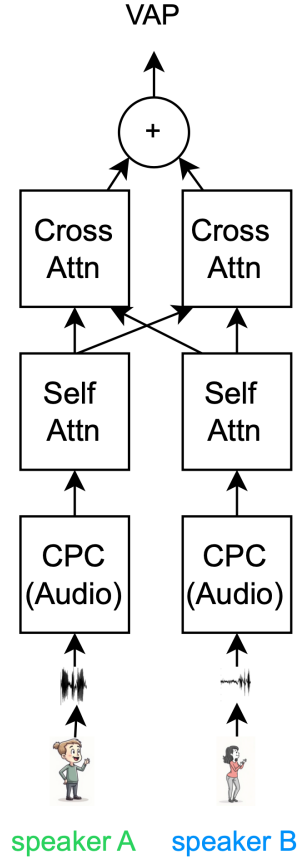


Figure 2.13 Voice activity projection (VAP) model architecture from (Inoue, Bing'er Jiang, et al. 2024), consisting of self and cross attention (attn) transformer layers stacked into blocks and a frozen CPC audio feature extractor. The transformer layer weights are shared between speakers. The model outputs a probability distribution over the 256 possible states in the VAP training objective, depicted in Figure 2.12.

$$\text{CROSSATTNBLOCK}(x_1, x_2) = \left[\begin{array}{l} Q = x_1 W_Q, \\ \sigma \left(\text{TRANSFORMERLAYER} \left(\begin{array}{l} K = x_2 W_K, \\ V = x_2 W_V \end{array} \right) \right) \\ Q = x_2 W_Q, \\ + \text{TRANSFORMERLAYER} \left(\begin{array}{l} K = x_1 W_K, \\ V = x_1 W_V \end{array} \right) \end{array} \right] N. \quad (2.9)$$

Where x_1 is the output of the self-attention block for the left speaker, x_2 for the right, and $\sigma(\cdot)$ is a linear layer followed by GeLU non-linearity. The cross-attention block comprises $N = 3$ stacked transformer layers; the output of each forms the input to the next layer. In each layer, there are two transformer layers. In the first transformer layer, the query q is computed

from the left speaker x_1 , and the key and value are computed from the right speaker x_2 . This means that the transformer learns temporal patterns in x_1 via self-attention, and temporal patterns between x_1 and x_2 via cross-attention. In the second transformer layer, x_1 and x_2 are reversed. The outputs from both are added before a GeLU non-linearity, $\sigma(\cdot)$, which ensures symmetry.

2.6.5.4 Prediction head

After the cross-attention block, there is a prediction layer (VAP in Figure 2.13). The output is a 256-dimensional vector over the 256 VAP probability states (shown in Figure 2.12), which describe speaking activity in the next 2 seconds. The layer is given by:

$$\text{VAP}(x_1, x_2) = \text{SOFTMAX}(\sigma(x_1) + \sigma(x_2)) \quad (2.10)$$

Where $\sigma(\cdot)$ is a GeLU activation, and $\sigma(x_1)$ and $\sigma(x_2)$ are the outputs of the cross-attention block for speakers A and B. Additionally, the model also predicts the probability that the *current* 20 ms frame contains speech, i.e. the voice activity signal:

$$\text{VAD} = [\sigma(x_1); \sigma(x_2)] \quad (2.11)$$

Where $[\cdot; \cdot]$ is a tensor stacking operation.

2.6.5.5 Loss function

During training of the VAP model, the following loss function is minimised (Inoue, Bing'er Jiang, et al. 2024):

$$L_{tot} = L_{vap} + L_{vad} \quad (2.12)$$

Where L_{vap} is the cross-entropy of the ground-truth VAP labels and the VAP prediction of the model (Equation 2.9) and L_{vad} is the binary cross-entropy of the ground-truth VAD labels and the VAD prediction of the model.

2.6.5.6 Causality in the VAP model

Both the embedding and the model are causal, meaning information from future frames of audio cannot be used in the prediction of upcoming speaking activity. This ensures that the model cannot naïvely copy information from future speaking turns, forcing it to use prior context like human turn-taking. This is achieved by using Alibi attention, which enforces causality in the self- and cross-attention blocks (Press et al. 2021; Inoue, Bing'er Jiang, et al. 2024).

2.6.5.7 Studies exploring the performance of the VAP model

A number of studies have explored the performance of the VAP model. In the original paper Ekstedt and Skantze (2022b) trained and evaluated the VAP model³ on the Switchboard corpus. They evaluated the model's ability to distinguish shifts from holds during the silence between adjacent speaking turns (shifts) and during mid turn silences (pauses). They found the joint modelling of both speakers' future speaking activity (Figure 2.12) was significantly higher than earlier approaches (Skantze 2017), which model both speakers' future speaking activity independently (e.g. 0.889 vs. 0.896 $F1$, Ekstedt and Skantze (2022b)). The model was further evaluated on other turn-taking tasks, such as the ability to distinguish mid-turn speech from speech at the end of a turn before a speaker transition. Notably, overlapping speech was not included in the evaluation.

The model has also been explored in further studies which sought to uncover which turn-taking cues it relies upon to make its predictions. Ekstedt and Skantze (2022a) investigated the role of prosody in the model's predictions. They conducted various transforms, such as the flattening of the $F0$ or intensity contour, and found that the model's performance decreased but remained above a baseline model that always predicted a hold. This demonstrates that the model utilises prosody in its predictions. However, as either $F0$ or intensity was flattened, the model may rely on the unmodified cue. Furthermore, the study does not address the potential for the model to be reliant upon lexical cues implicitly captured in self-supervised speech representations (Pasad et al. 2021), such as the CPC front-end, which are not addressed.

In another study, Bing'er Jiang et al. (2023) demonstrated that filled pauses such as 'uhm' strongly impact the VAP model's predictions. Where a filled pause is present, the model strongly predicts a hold, echoing the turn-holding cues of filled pauses in human turn-taking (Wennerstrom and Siegel 2003; Gravano and Hirschberg 2009). This suggests the model is sensitive to lexical cues. However, filled pauses are sustained voiced segments: this may mean that vowel elongation, which is a turn-taking cue (Duncan 1972), is exploited by the model. This illustrates the complexity of analysing the behaviour of models utilising self-supervised representations. Furthermore, a study by Inoue, Bing'er Jiang, et al. explored the multilingual capabilities of the VAP model. For example, a model trained on English achieved a balanced accuracy of hold/shift prediction of 59.4% when tested on Japanese and 68.6% when tested on Mandarin, compared to 79.6% when evaluated on English (Switchboard corpus). Performance was therefore above chance in unseen languages, but significantly degraded. This suggests that the model is able to exploit prosodic variation in unseen languages, as neither the model nor the audio front end CPC model was exposed to Japanese or Mandarin during training.

³The mono-audio original version of the VAP model

As noted in Chapter 1, the performance of PTTMs, including the VAP model have not been explored in noise. This is surprising as once deployed in a human-robot interaction scenario, the models will inevitably encounter noise. We address this in Chapter 7, where we explored PTTM performance in acoustic noise and disentangle the impact of additive noise on prosody and intelligibility.

2.6.6 Self-supervised representations in predictive turn-taking models

As noted, in the VAP model, features are extracted from the audio using a pre-trained feature extractor, called the CPC model (Riviere et al. 2020). This is highlighted in Figure 2.13. The model is loaded from pre-trained weights and frozen during training. The CPC model extracts self-supervised representations from the audio: i.e. extracted features that are learned by pre-training on large volumes of unlabelled data. This generates useful representations across downstream tasks (Pasad et al. 2021; Gimeno-Gómez et al. 2025).

The ability of self-supervised speech representation to capture rich representations of acoustic and linguistic phenomena makes these features substantially more expressive than those used in earlier models, such as $F0$ tracks and POS tags (Pasad et al. 2021). Hence, self-supervised representations have largely replaced feature engineering (e.g. the extraction of $F0$ contours, spectral features) in PTTMs. Other embeddings have been used, such as wav2vec audio features (Li et al. 2022) GPT-2 word embeddings in text-based models Li et al., and other types of word embedding (Roddy and Harte 2020).

However, the richness of features derived from self-supervised representations introduces a major challenge: interpretability (Pasad et al. 2021; Gimeno-Gómez et al. 2025). The richness of speech is collapsed into a single representation; it is difficult to disentangle whether prosody or lexical information drives turn-taking prediction, which was the case in earlier work (Roddy, Skantze, et al. 2018a; Skantze 2017). Furthermore, audio S3R can learn useful lexical information implicitly (Pasad et al. 2021). This means that a model which is not trained using any textual word embeddings could potentially exploit both prosodic and lexical cues. As a result, it remains unclear which of the rich prosodic (e.g. $F0$, intensity) and lexical cues identified in our earlier review in Section 2.2 are most informative for prediction. It is also unknown whether acoustic cues continue to dominate performance, like earlier work using engineered features (Skantze 2017; Roddy, Skantze, et al. 2018a).

As noted, studies show that the VAP model is sensitive to prosody (Ekstedt and Skantze 2022a). Furthermore, the multilingual capabilities of the model, which achieved above-chance performance in unfamiliar, unrelated languages, similarly suggesting that prosody plays a major role in prediction (Inoue, Bing'er Jiang, et al. 2024). However, there has been no systematic analysis of which prosodic and lexical features are exploited by VAP models. More broadly, a growing body of work has proposed interpretability techniques for S3R speech models. These efforts have largely focused on phonetic and lexical representations

(Pasad et al. 2021; Parsons et al. 2025). This leaves an important gap in understanding how self-supervised representations support conversational interaction tasks, such as turn-taking.

2.6.7 The TurnGPT and PairwiseTurnGPT models

While the VAP model deals with audio, large language model (LLM) based turn-taking models have been proposed, which utilise text alone to make their predictions. The advent of transformer-based LLMs has led to interest in utilising LLMs for turn-taking, both as extraction tools for word embeddings (Li et al. 2022) but also as turn-taking models themselves (Ekstedt and Skantze 2020; Leishman et al. 2024).

As audio models using S3R can potentially exploit both prosodic and lexical cues through self-supervised speech representations, we use LLM based turn-taking models to fully isolate the contribution of lexical cues to turn-taking prediction. We detail two models from the literature used in our exploration of text-based turn-taking prediction in Chapters 5 and 7.

2.6.7.1 TurnGPT

The TurnGPT model was proposed by Ekstedt and Skantze (2020). The model is essentially a pre-trained GPT-2 LLM from Radford, Wu, et al. (2019) with an additional token inserted where speaker transitions occur. The Switchboard corpus was serialised into turns with overlapping speech ignored. As the model is an LLM, a single stream of text is required. Punctuation is removed, and then tokens are encoded with byte-pair encoding (Sennrich et al. 2016). The transcription was therefore transformed into a single stream of text. Where speaker transitions occurred, an extra turn-shift <TS> was inserted, with overlapping speech ignored. An example of this using the pair of turns from the start of the chapter in Figure 2.1 is shown below (punctuation removed):

- *John are you a student here at this university <TS> yes I am*

The pre-trained GPT-2 LLM was loaded (Radford, Wu, et al. 2019), and the vocabulary size was increased to allow for this extra token. The model was evaluated across different corpora, including Switchboard (Ekstedt and Skantze 2020) and achieved high accuracy in distinguishing holds from shifts (82%). However, the model was evaluated in later work by Bing'er Jiang et al. (2023) and found to suffer from early false positives arising from syntactic ambiguity, examples of which were highlighted in the example in Section 2.1 at the start of the chapter. This was resolved by additionally providing the model with the ground-truth text of the response (Bing'er Jiang et al. 2023). This is similar to Roddy and Harte (2020) who also used the text of the response in the prediction of the timing of responses. However, this makes the model non-causal as future information is assumed to be known and used for turn-taking prediction.

2.6.7.2 Pairwise TurnGPT

The TurnGPT model was expanded by Leishman et al. (2024). Their Pairwise TurnGPT model utilises two GPT-2 models, initialised from the pre-trained weights from Radford, Wu, et al. (2019). There is one stream of text per speaker, and they are combined by cross-attention prior to prediction. Like TurnGPT, byte-pair encoding is used. Speaker tokens are aligned into pairs using the ground-truth timing information from the Switchboard corpus, and the model introduces additional special tokens to handle overlapping speech, backchannels and pauses. We provide a detailed discussion and example of the alignment procedure in Chapter 7. For now, we note that Leishman et al. (2024) found that the model significantly improved performance with a balanced accuracy of 92% for end-of-turn prediction compared with a balanced accuracy of 85% in the TurnGPT model. A full overview of the models is provided in Chapter 5, where we discuss turn-taking with lexical cues.

2.6.8 Summary of predictive turn-taking models

An overview of the PTTMs discussed here is provided in Table 2.4. As this is a rapidly developing field, a number of PTTMs have been developed independently of our own work since thesis planning, including multimodal models (Onishi et al. 2024). We defer discussion of these newer models until our future work section in Chapter 8.

2.6.9 Evaluation of trained predictive turn-taking models

Next, we describe how the performance of trained PTTMs is reported. For this discussion, assume a trained model is available that predicts the probability of a shift. A ground-truth label is also available, extracted from a transcript as described in Section 2.4.6. This makes turn-taking a binary prediction problem. PTTMs are then typically evaluated using the F_1 and balanced accuracy scores (Skantze 2017; Roddy, Skantze, et al. 2018a; Inoue, Bing'er Jiang, et al. 2024).

To compute the metrics, let a shift be a *positive* prediction, and a hold be a *negative* prediction. We then compute the number of true and false positives, as well as the number of true and false negatives in the corpus as predicted by the model. The F_1 score for the shift class is then:

$$F_1(\text{SHIFT}) = 2 \times \frac{\text{Precision}_{+ve} \times \text{Recall}_{+ve}}{\text{Precision}_{+ve} + \text{Recall}_{+ve}} \quad (2.13)$$

Where the precision for the positive class (shifts) is:

$$\text{Precision}_{+ve} = \frac{TP}{TP + FP}, \quad \text{Recall}_{+ve} = \frac{TP}{TP + FN} \quad (2.14)$$

We then compute the F_1 score for the hold class:

$$F_1(\text{HOLD}) = 2 \times \frac{\text{Precision}_{-ve} \times \text{Recall}_{-ve}}{\text{Precision}_{-ve} + \text{Recall}_{-ve}} \quad (2.15)$$

Where the precision for the negative class (holds) is:

$$\text{Precision}_{-ve} = \frac{TN}{TN + FN}, \quad \text{Recall}_{-ve} = \frac{TN}{TN + FP} \quad (2.16)$$

A higher F_1 score indicates a higher number of correctly identified shifts and holds, and hence, a higher model accuracy. However, an issue with linguistic corpora is that pauses in speech outweigh the number of speaker transitions (see Chapter 4). The weighted F_1 score accounts for any class imbalance. The weighted F_1 score formula is:

$$F_1(\text{WEIGHTED}) = \frac{n_{\text{shift}} \times F_1(\text{SHIFT}) + n_{\text{hold}} \times F_1(\text{HOLD})}{n_{\text{shift}} + n_{\text{hold}}} \quad (2.17)$$

Where n_{shift} and n_{hold} are the numbers of shifts and holds in the corpus. Finally the balanced accuracy score has also been used in predictive turn-taking models development (Inoue, Bing'er Jiang, et al. 2024) and is given by:

$$\text{Balanced accuracy (\%)} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \times 100 \quad (2.18)$$

A random model has a balanced accuracy of 50%, making it easy to compare model performance against a baseline. Further detail on evaluation metrics is provided in Chapter 4, where we outline our specific evaluation procedure for PTTMs in this thesis.

2.6.10 Turn-taking in multiparty interaction

Although our work focusses on turn-taking in two-party interaction, we briefly highlight some of the complexities of multiparty interaction here. Turn-taking in multiparty interaction is considerably more complex as interlocutors must coordinate who should speak next at the end of each turn. Speaker changes can take place explicitly by directly naming the next speaker ("*John, you go ahead*") or implicitly through the use of multimodal cues such as eye gaze (O'Conaill et al. 1993). Further complexity is introduced by the multiple roles that participants can play. In a dyadic interaction, an interlocutor is either a speaker or a listener. On the other hand, in multiparty interaction, an interlocutor may be a speaker or a listener, a bystander to the main conversation, or an overhearer playing no role in the interaction (A. H. Anderson 2006).

Given this complexity, most PTTMs focus on two-party interaction. Comparatively less PTTMs for multiparty interactions have been published e.g. Malik et al. (2020). Beyond modelling complexity, a further complication that also limits development is that multiparty

datasets are limited in size in comparison with two-party datasets (see review in Section 2.3). We therefore focus on modelling two-party interaction in our work. Although much work is needed in multiparty interaction, there are ample challenges which remain in predictive turn-taking for two-party interaction, as highlighted in our review and the following section.

2.7 Gaps identified from our literature review

From our literature review, we have identified a number of gaps in the literature, briefly summarised here. These are addressed in the work chapters of this thesis, and form the basis of our research questions outlined in Chapter 1.

2.7.1 The use of manual transcriptions

All PTTMs highlighted here are trained on corpora for which manual transcriptions are provided and used. This includes models which utilise the transcriptions for timing information only (Ekstedt and Skantze 2022b; Inoue, Bing'er Jiang, et al. 2024) but also models utilising word embeddings which directly use lexical information from the transcripts (Roddy and Harte 2020; Ekstedt and Skantze 2020; Leishman et al. 2024).

As manual transcription is costly and time-consuming, this limits the amount of data available for training. Our review notes ASR could vastly increase the amount of data available for PTTM training and development, however no study has considered whether ASR is sufficiently accurate for model training and development. Furthermore, the literature demonstrates a strong general interest in ASR for transcription. However, the literature lacks detail on the considerations a researcher wishing to deploy ASR in their work should follow, and a discussion of the pros and cons of open-source and commercial ASR.

2.7.2 A lack of non-verbal cues

Our review notes the role of non-verbal cues in human turn-taking has been well-studied. Yet, most PTTMs utilise cues from speech alone. Many models are trained on telephone speech, where interlocutors cannot respond to non-verbal cues such as gaze. At the time of thesis planning, only a single study explores the incorporation of non-verbal cues into a PTTM (Roddy, Skantze, et al. 2018b). However, the model only utilised gaze, and the findings were very limited due to the use of an 11-hour corpus of in-person interaction.

Hence, it is unknown whether the inclusion of non-verbal cues can support PTTMs at scale, on a large corpus of human interaction. Moreover, it is unknown which non-verbal cues can support turn-taking, e.g. from the mouth, head pose or eye gaze. Furthermore, the use of telephone speech to train PTTMs is widespread (Roddy and Harte 2020; Li et al. 2022; Ekstedt and Skantze 2022b). However, the extent to which these models are suitable

for deployment to interactions where interlocutors can see as well and hear one another, and hence utilise non-verbal cues, remains unknown. As these models are intended to be deployed in real-world interactions to support human-robot interaction, where non-verbal cues will play a role, this is an important question to answer.

2.7.3 A limited understanding of the contributions of prosodic and lexical cues

Early audio-only PTTMs employed feature engineering. This enabled detailed ablation studies to be conducted, which revealed that prosodic cues contributed significantly more to prediction than lexical cues (Skantze 2017; Roddy, Skantze, et al. 2018a). However, in these models, the lexical representations were very limited, e.g. part-of-speech tags.

Recently, models have exploited self-supervised speech representations (Ekstedt and Skantze 2022b; Inoue, Bing'er Jiang, et al. 2024; Li et al. 2022). These representations are far more powerful than engineered features, as they capture prosodic, syntactic and semantic cues (Pasad et al. 2021). However, the richness of speech is collapsed into a single representation, making interpretability challenging (Gimeno-Gómez et al. 2025). Therefore, these models capture a range of speech features relevant to turn-taking, such as prosody and lexical cues.

Our review notes a range of prosodic and lexical cues from speech are used for humans during turn-taking. However, precise cues that these PTTMs exploit are unclear. A classic ablation study involving the selecting, inclusion, and removal of cues as done in prior work using features engineering (Skantze 2017; Roddy, Skantze, et al. 2018a) is not possible with self-supervised representations. Furthermore, recently LLM based models, which use text alone to make their prediction, have been proposed. Further work is needed to understand the role that each of these cues plays in turn-taking prediction to inform better modelling choices in deployed PTTMs.

2.7.4 A limited understanding of performance in real-world conditions

Finally, the evaluation of models is conducted on corpora collected in controlled experimental conditions. Although corpora such as Candor (Reece et al. 2023) enforce little control over where interlocutors are recorded, there is no specific source of noise introduced or measured. Hence, it is unknown how PTTMs trained on corpora of human interaction will handle noise which will inevitably be encountered in-the-wild, once deployed. As these models are expected to be deployed on-device to support human-robot interaction Skantze; Skantze and Irfan, it is important to establish how these models handle acoustic interference.

Table 2.4 Overview of turn-taking models identified in our review

Author	Name	Model Architecture	Audio	VAD	Acoustic Features	Lexical Features	Other Features	Corpus	Predicts
Skantze (2017)		LSTM	Mono	Y	F_0 , Intensity, spectral stability	POS tags		Map Task	Speaking activity of one interlocutor (next 3 s)
Roddy, Skantze, et al. (2018a)		LSTM	Mono	N	Lininess, shimmer, F_0 , jitter, MFCCs	Word embeddings		Map Task	Speaking activity of one interlocutor (next 3 s)
Li et al. (2022)		LSTM + GMM	Mono	N	Intensity, spectral stability	way2vec 1.0		Switchboard	Response time distribution
Priddy, Skantze (2020)		LSTM	Mono	N	RMS energy, F_0 , frequency features	Response embeddings		Switchboard	Response time distribution
Roddy, Skantze et al. (2018b)		LSTM	Mono	N	VAD, energy, log-Mel filterbank, F_0 , acoustic energy	Word embeddings		Mainob Mimicry, Map Task (audio)	Speaking activity of one interlocutor (next 3 s)
Roddy, Skantze (2022b)	VAP	Transformer (GPT-2)	Mono	Y	GallAPS feature set	CPC		Switchboard	Speaking activity of two interlocutors (next 2 s)
Erkoc, Bingöl, Jiang, et al. (2024)	VAP	Transformer (GPT-2)	Stereo	N				Various + Switchboard	Next token
Erkoc and Skantze (2020)	TurnGPT	Transformer (GPT-2)				BPE tokens		Various + Switchboard	Next token
Bingöl, Jiang et al. (2023)	Response-conditioned TurnGPT	Transformer (GPT-2)				BPE tokens		Various + Switchboard	Next token
Leishman et al. (2024)	Pairwise TurnGPT	Transformer (GPT-2)				BPE tokens		Switchboard	Next tokens for interlocutor pair

Chapter 3

Transcribing conversations with ASR

Most work involving conversation requires a transcript. Transcripts are also essential for training predictive turn-taking models, which comprise much of the work in this thesis. They are also heavily used in many fields, from linguistics (Schegloff 2000b) to medicine (K. Latif et al. 2023) and criminology (Loakes 2022). Yet typically most transcription is typically done manually, by experts; a laborious process which can take 14 times the length of the audio (Umair et al. 2022). There is therefore much interest in using ASR to automate the task, at least in part (García Molina et al. 2024). As highlighted in Chapter 2, ASR is a promising technology for transcription. The technology has seen many recent advances, and is now widely available and can be run on a researcher's machine, for free. This has led to the adoption of ASR to support transcription across diverse fields.

Despite this widespread adoption, a number of questions remain regarding the use of ASR for transcription. The performance of ASR is worse on conversational speech than on read speech (Stolcke and Droppo 2017). However, as much of the current ASR literature focusses on the performance in terms of the WER, other aspects of transcription quality are understudied; most notably, the transcription of NLTs. NLTs are short words such as *uhm* and *ehm* that serve various functions, including backchannels and answers to yes-no questions. Accurate transcription of these words is a critical requirement for transcription in many applications.

Although there is evidence that commercial ASR outperforms open-source ASR on conversational speech (Morbini et al. 2013; Georgila et al. 2020; Speechmatics, Ltd. 2024), to the best of our knowledge at the time of writing, the literature lacks a comprehensive comparison of open-source and commercial tools. There is therefore a need for a study which goes beyond the WER, notably exploring the transcription of NLTs and accuracy of word alignment in commercial and open-source ASR platforms. Lastly, a deployment of ASR for conversational speech transcription should be evaluated within the context of the available resources, researcher skill-set and budget. The current literature lacks these necessary insights.

In this chapter, we address these open issues in the use of ASR for transcription, addressing **Research Question 1** of this thesis. We review six state-of-the-art ASR technologies, three commercial (Amazon, Inc. (2024), Google, Inc. (2024) and Speechmatics, Ltd. (2024)) and three open-source (wav2vec2.0 (Baevski et al. 2020), Hubert (Hsu et al. 2021) and Whisper (Radford, Kim, et al. 2023)). We assess their capabilities as automatic transcription tools, using the Switchboard (Godfrey et al. 1992), RoomReader (Reverdy et al. 2022) corpora of conversational speech and the LibiSpeech corpus (Panayotov et al. 2015a) of audiobooks (read speech). We conduct a thorough evaluation and make a number of recommendations concerning the use of ASR in research involving transcription.

The work has broad application in various domains, including linguistics and machine learning. It also has relevance for the training of PTTMs, which is the basis of subsequent chapters of the thesis. This is because PTTMs rely on transcription to extract word alignments for training. The work in this chapter is therefore vital in supporting the work of later chapters.

Chapter structure

The remainder of this chapter is laid out as follows. Section 3.1 defines transcription and sets out the evaluation criteria for ASR in this chapter. In Section 3.2, open-source and commercial ASR platforms are outlined, and suitable corpora are identified. The evaluation methodology for our study of ASR is outlined in Section 3.3.

The results are outlined and discussed in Sections 3.4 and 3.5. Finally, in Section 3.6, we discuss the implications of our findings; identify key factors to consider when considering the use of ASR in research; and identify limitations and directions for future work.

Research questions addressed

The work in this chapter addresses the following research question:

- **Research Question 1** *Can modern ASR systems support accurate transcription of conversational interaction?*

Publications

The work in this chapter is presented in the following journal publication:

- O'Connor Russell, Sam, Iona Gessinger, Anna Krason, Gabriella Vigliocco, and Naomi Harte. "What automatic speech recognition can and cannot do for conversational speech transcription". *Research Methods in Applied Linguistics* 3, no. 3. 2024: 100163.

3.1 Defining transcription

We begin the chapter by outlining a definition of transcription and setting out our evaluation criteria. Put simply, a transcript is a written record of an interaction, necessary in research involving conversation (Psathas and T. Anderson 1990). The specific requirements of a transcription, however, vary by domain as highlighted in our literature review in Chapter 2.

Given the differing views on what exactly transcription entails, we adopt a simple definition in this chapter. We define transcription as writing what was said from audio, also known as *speech-to-text* (Google, Inc. 2024; Amazon, Inc. 2024) or *orthographic* transcription (Goldman 2011). Furthermore, for the purposes of our study, we set out a number of attributes that the transcription should fulfil based on our literature review, namely the ability to:

1. Capture an accurate representation of what was said
2. Capture non-lexical tokens such as *uhm* and *eh-hem*
3. Accurately record the timing of words, and
4. Accurately capture overlapping speech

As we work with orthographic transcription in this chapter, specific actions which cannot be transcribed with ASR, e.g. non-verbal cues such as head nodding to indicate a yes response, are omitted. Such cues may sometimes be included in transcriptions, notably in the Conversation Analysis domain (Clift 2016). However, non-verbal cues clearly cannot be identified from audio alone and hence ASR. We also exclude the annotation of paralinguistic features, most notably laughter, from our analysis. The inclusion of such paralinguistic cues is not typical in ASR systems, including those used in this chapter.

3.2 Transcribing with automatic speech recognition

We choose a range of both commercial and open-source providers as the literature lacks detail on which type of ASR should be used in research, as outlined in Chapter 2. Open-source is attractive in a research setting as it is free-to-use, and can be run locally on a researcher's machine. This may be a condition of ethical approval, as commercial services involve uploading recordings to remote commercial servers, possibly in different jurisdictions. In recent years, open-source ASR tools such as OpenAI's Whisper (Radford, Kim, et al. 2023) have made state-of-the-art speech recognition widely available. However, the literature lacks detail on whether they are suitable for conversational speech transcription.

Commercial ASR providers have been around for some time, although cost may have limited uptake in research. Self-reported industry metrics demonstrate improvement above open-source models, including Whisper (Speechmatics, Ltd. 2024). However, it is unclear

whether any performance advantage means that they are more suited to the transcription of conversational speech.

3.2.1 Open source tools

We select three open-source ASR models from the Huggingface machine learning repository (Huggingface 2024) which can run on an Nvidia 3080 GPU, a consumer-grade GPU that can be installed on a desktop computer. The models are Whisper (*openai/whisper-medium.en*), HuBERT (*facebook/hubert-large-ls960-ft*) and wav2vec2.0 (*facebook/wav2vec2-large-960h*). Models are based on the transformer neural network architecture (Vaswani 2017). Wav2vec2.0 is pre-trained with unlabelled audio and fine-tuned on LibriSpeech (Baevski et al. 2020). HuBERT is similarly fine-tuned on LibriSpeech (Hsu et al. 2021). Whisper (Radford, Kim, et al. 2023) is trained on unspecified transcribed audio from the internet.

3.2.2 Commercial tools

A number of commercial ASR providers are available. We received limited academic access to three providers: Amazon (Amazon, Inc. 2024), Google (Google, Inc. 2024) and Speechmatics (Speechmatics, Ltd. 2024). Each has their own API. There is no public information about model architecture or training data. We used the Speechmatics *enhanced* model and the standard EN-US model for Google and Amazon¹. All three commercial systems output word timings (i.e. a *word alignment*).

As commercial ASR providers are continually updated, there is limited control over which model version is used for transcription. The results are therefore valid for transcriptions obtained in January 2024, when this work was conducted.

3.2.3 Corpora

After choosing ASR systems, we analyse their performance on corpora from the literature. We select corpora from the literature which are provided with a ground-truth transcription. This enables the quality of an ASR transcription to be compared against that of a human transcriber. The corpora are Switchboard (Godfrey et al. 1992) and RoomReader (Reverdy et al. 2022), which are corpora of casual conversational interaction. We also choose LibriSpeech (Panayotov et al. 2015a), a corpus of audiobook speech, as a comparison across conversational and read speech. Table 3.1 provides descriptive statistics, and the corpora

¹Google and Amazon offer an Irish-English model, the predominant dialect type in the RoomReader corpus. However, assessment on the female Irish tutor in RoomReader revealed the WER was higher than the standard EN-US model (12.35% vs 11.6%). We therefore used the EN-US model for all speakers.

Table 3.1 Descriptive statistics of the corpora used to evaluate ASR transcription

Corpus	Domain	Hours	Speakers	Recordings	Mean session length
Switchboard	telephone	260	4,876	2,438	6 min 23 s
RoomReader	videoconference	40	118	30	17 min 30 s
LibriSpeech	audiobook	20	146	11,123	6.8 s

are outlined in further detail in Chapter 2. Important information concerning dialect and available manual transcription is outlined here.

Switchboard Switchboard is a common ASR benchmark consisting of unscripted, two-party, English language telephone conversations about set topics, e.g. recycling. It was manually transcribed. Dialogue partners are native speakers of US English grouped into 7 dialect categories covering the continental US (e.g. New York, Southern, Western, etc.) (Godfrey et al. 1992). Audio is recorded in 16 kHz mono, with one channel per speaker.

The corpus contains lexical transcriptions of NLTs such as ‘*uhm*’ and ‘*ehm*’. Paralinguistic phenomena are inserted into the transcription in various categories such as *laughter* and *vocalised noise*. We remove these from the transcription as they are typically not transcribed by ASR, and we conduct no further processing.

RoomReader RoomReader is a corpus of small-group, English language conversational interactions recorded over Zoom. A tutor guides 3-4 students through a game, outlined in Chapter 2. The corpus is provided with a ground-truth transcription which contains lexical transcriptions of NLTs. No processing of the transcription is required. Audio is recorded in 48 kHz mono, with one channel per speaker.

The predominant dialect of native speakers in the corpus is Irish English. There is a range of English dialects spoken in Ireland, however, no further geographic information is available (Reverdy et al. 2022). The remaining speakers are fluent in English and have a variety of native languages (Greek, French, etc.). The mix of native and non-native speakers in this corpus makes it a valuable benchmark as ASR performance is degraded in non-native accented speech (Alharbi et al. 2021).

Librispeech Librispeech (Panayotov et al. 2015b) is a corpus of English language audiobook recordings. Each recording is 16 kHz mono and assigned either a *clean* or an *other* label, where *other* means that it is more challenging for an ASR system to transcribe (Panayotov et al. 2015a). There three partitions of LibriSpeech: *test*, *dev* and *train* (Panayotov et al. 2015a). We select all the recordings in the *dev* and *test* partitions and evaluate on these recordings. This is because, by convention, ASR researchers and developers do not train

ASR models on the *test* and *dev* splits. As LibriSpeech is literary prose, there are no NLTs, overlapping speech, etc.

Running the ASR models We run each ASR tool separately on each channel of audio on each corpus, as there is a single speaker in each channel of audio. The output files are JSON files, which we standardise into the Praat TextGrid format for consistency (Boersma and Van Heuven 2001). All transcriptions were made available to the wider research community through our publication (O'Connor Russell et al. 2024).

3.3 Evaluating an ASR transcription

After running the ASR models on the corpora, we compare their output transcriptions with the manual transcriptions provided with the corpora. In this section, we outline the evaluation protocol, prior to exploring the performance of the models in Section 3.4. Rather than solely comparing ASR systems' accuracy in terms of the WER, we assess their ability to fulfil typical transcription requirements, which we have highlighted in Section 3.1.

3.3.1 Insertions, deletions, substitutions and the error rate

The first requirement for an ASR transcription we identified in Section 3.1 is to determine if an ASR accurately captures what was said. For this evaluation, we use the WER. We use the *jiwer* Python package² to compute the number of matches as well as deletion, substitution and insertion errors with reference to a human transcribed ground-truth transcript.

The WER formula is given by (Jurafsky and Martin 2000):

$$\text{WER}(\%) = 100 \times \frac{\# \text{ insertions} + \# \text{ deletions} + \# \text{ substitutions}}{\# \text{ words in ground truth transcript}} \quad (3.1)$$

We compute the WER at corpus level. It rewards correct spelling and placement of words. As a rule of thumb, Microsoft advise a 5-10% WER means an ASR is “ready to use”; 20%-30% indicates optimisation is necessary; and >30% indicates poor quality (Microsoft, Inc. 2024). Prior to running *jiwer*, we removed all non-alphanumeric characters except apostrophes and performed normalisation³. This ensures the WER is not penalised by differences between UK and US spelling (rumour vs. rumor), numbers (100 vs. one hundred), contractions (you'll vs. you will), etc. We then conduct a more detailed error analysis of the

²<https://pypi.org/project/jiwer/>

³We used the OpenAI normalisation algorithm, with a slight modification to ensure that NLTs were preserved in the output <https://github.com/openai/whisper/blob/main/whisper/normalizers/english.py>

insertions, deletions and substitutions obtained from `jiwer`. This enables a more detailed consideration of transcription performance on NLT, fulfilling our second evaluation criterion from Section 3.1. We then isolate performance on NLTs by enumerating a list of NLTs in the transcriptions, reporting the error rate, insertions, deletions and substitutions. This enables transcription performance on NLTs to be assessed in detail.

3.3.2 Word timing and overlap

Our third evaluation criterion is for the ASR to accurately record when words were spoken, i.e. the accuracy of the word alignment. The `jiwer` package provides an alignment enabling words in the human transcription to be queried for corresponding words in the ASR transcription. To assess the accuracy of ASR word timing, we choose the words correctly transcribed by the ASR. Then using the alignment provided with `jiwer`, we compute the absolute error in the start and end times of the word in ASR and human transcripts.

Finally, we consider the accuracy of overlapping speech transcription using the word timings, fulfilling our final evaluation criterion. Note that we do not conduct speaker diarisation, as we run ASR separately for each speaker (channel of audio), as discussed further on in Section 3.5.

3.4 Conversational speech transcription with ASR

3.4.1 Transcribing what was said

A key attribute of a good transcription that we have identified is to record exactly what was said (see Section 3.3). We show sample ASR transcriptions in Table 3.2, highlighting ASR transcription errors. In the example, we note a substitution error where the word “*gym*” is replaced with “*jam*”. We also note the deletion of the word “*uh*”, an example of a non-lexical token (NLT) in speech.

The WERs for all ASR systems are reported in Table 3.3 and in bar chart format in Figure 3.1. A lower WER means that the ASR transcript is closer to the human ground-truth. All ASR platforms have a lower WER on read speech (LibriSpeech) than on conversational speech (Switchboard and RoomReader). Taking the example of Amazon, the WER on the ‘clean’ LibriSpeech partition is 2.33%, whereas the WER on Switchboard is 7.9%. The higher WER on conversational speech reflects the aspects which make conversational speech more challenging to transcribe, such as the presence of NLTs and disfluencies (Stolcke and Droppo 2017).

Commercial and open-source ASR perform similarly on read speech, but a clear performance gap appears on conversational speech. On Switchboard, the WERs for the commercial ASR systems (7.9-10.2%) are lower than the three open-source systems (19.4-34.1%). On

Table 3.2 Sample transcription of LibriSpeech (read) and Switchboard (conversational) with errors underlined.

Corpus	Source	Transcription
LibriSpeech	<i>ground truth</i>	one minute a voice said and the time buzzer sounded
	<i>wav2vec2.0</i>	one minute a voice said and the <u>timebuzz</u> sounded
Switchboard	<i>ground truth</i>	do you exercise uh during the day time or go to a gym
	<i>wav2vec2.0</i>	do you exercise during the daytime or go to a <u>jam</u>

Table 3.3 Word-error rates (WER %) across six speech corpora, computed at corpus-level. RoomReader (RR) is split into student (s) and tutor (t) speakers; LibriSpeech (LS) into *clean* and *other* partitions. The best WER is highlighted in bold for each corpus.

ASR System	Switchboard	RoomReader	RR (s)	RR (t)	LibriSpeech (clean)	LibriSpeech (other)
Commercial						
<i>amazon</i>	7.9	16.1	22.8	11.4	2.33	5.16
<i>google</i>	10.2	46.2	50.3	43.6	5.58	10.51
<i>speechmatics</i>	9.4	15.7	21.5	11.8	3.16	6.18
Open-source						
<i>hubert</i>	22.6	39.9	53.6	30.3	1.96	4.05
<i>wav2vec2.0</i>	34.1	51.7	61.3	45.2	3.21	8.6
<i>whisper</i>	19.4	26.7	41.4	17.7	3.12	5.75

RoomReader, the WERs are much higher than on Switchboard. Google is an outlier as it performs poorly on RoomReader, but its performance on Switchboard is comparable to the other ASR systems. In RoomReader, the tutors guide the students through the interaction, whereas students often make small contributions, encouraged by the tutor (Reverdy et al. 2022). We therefore computed the WER separately for students and tutors. This reveals that the tutor WER is closer to the Switchboard WER, whereas the student WER is higher than the WER on Switchboard.

3.4.2 Error analysis of ASR

Having now established that ASR makes more mistakes on conversational speech than on read speech, we now focus more closely on conversational speech transcription on the Switchboard and RoomReader corpora. We begin by exploring whether ASR tends to omit certain words more so than others.

We first count the occurrences of each word in the human and ASR transcripts, excluding stopwords⁴ (I, it, etc.). We then compare word counts in human and ASR transcriptions using the log-likelihood ratio (LLR), which compares relative word frequencies (Rayson and Garside 2000). The higher the LLR, the less likely this word is to occur in an ASR transcription in

⁴We used the standard NLTK English stopword list <https://www.nltk.org/>

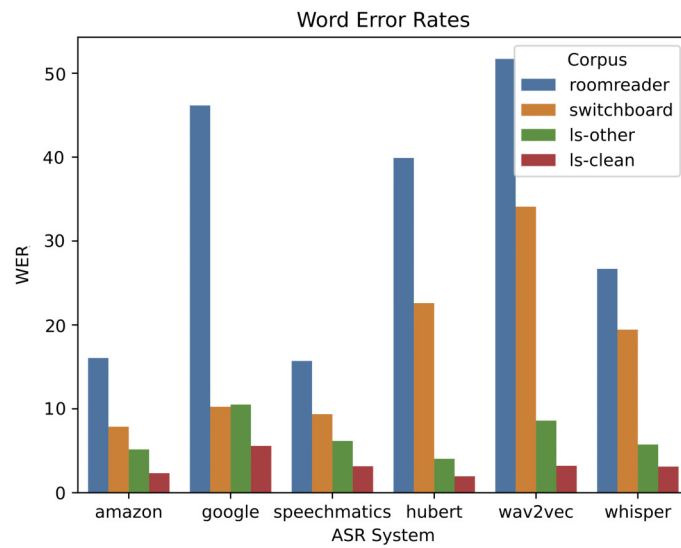


Figure 3.1 Word-error rates (WER, %) for six ASR systems on RoomReader, Switchboard (conversational speech) and LibriSpeech (ls, read speech, clean and other partitions).

Table 3.4 Top five words (excluding stopwords) ranked by log-likelihood ratio (LLR) of word counts in human and ASR transcripts of Switchboard and RoomReader. These words occur more frequently in human transcripts than in ASR transcripts. NLTs are highlighted in bold.

ASR system	Top 5 words
amazon	uhhuh ; umhum ; okay; huh ; aha
google	umhum ; uhhuh ; uh ; alright; huh
speechmatics	umhum ; uhhuh ; yeah; uh ; hm
hubert	uh ; yeah; ye; uhhuh ; umhum
wav2vec2.0	uh ; yeah; uhhuh ; umhum ; ye
whisper	uh ; umhum ; um ; ha ; hum

comparison with a human transcript. The top five words for each ASR system ranked by LLR are shown in Table 3.4. NLTs such as *uhhuh* and *huh* are less likely to occur in an ASR-generated transcript than in a human transcription. Our findings echo those of Liesenfeld et al. (2023) and Lopez et al. (2022) who also found fewer NLTs in ASR transcripts.

Next, we count the number of times each word in the human transcript is matched, deleted, inserted or substituted in the ASR transcripts of Switchboard and RoomReader. We show the top five most common matches and errors in Table 3.5. The results clearly reveal that NLTs comprise most of the top five substitution and deletion errors in ASR transcripts.

Table 3.5 The top five words most commonly transcribed correctly by ASR (matches), as well as the top five deletion, substitution, and insertion errors.

	ASR system	Top 5 words
Matches:	<i>amazon</i>	uh ; know; yeah; like; well
	<i>google</i>	know; yeah; uh ; like; well
	<i>speechmatics</i>	know; uh ; yeah; like; well
	<i>hubert</i>	know; like; think; well; really
	<i>wav2vec2.0</i>	know; like; think; well; really
	<i>whisper</i>	yeah; know; like; well; think
Deletions:	<i>amazon</i>	uh ; yeah; well; umhum ; uhhuh
	<i>google</i>	uh ; umhum ; uhhuh ; yeah; um
	<i>speechmatics</i>	uh ; yeah; umhum ; uhhuh ; right
	<i>hubert</i>	uh ; know; yeah; umhum ; well
	<i>wav2vec2.0</i>	uh ; yeah; know; umhum ; uhhuh
	<i>whisper</i>	uh ; um ; know; yeah; umhum
Substitutions:	<i>amazon</i>	uhhuh ; umhum ; uh ; okay; um
	<i>google</i>	uh ; umhum ; uhhuh ; yeah; well
	<i>speechmatics</i>	uh ; umhum ; uhhuh ; yeah; oh
	<i>hubert</i>	uh ; yeah; um ; uhhuh ; umhum
	<i>wav2vec2.0</i>	uh ; yeah; um ; uhhuh ; know
	<i>whisper</i>	umhum ; uh ; uhhuh ; yeah; um
Insertions:	<i>amazon</i>	uh ; huh ; hmm ; yeah; bye
	<i>google</i>	uh ; yeah; know; bye; oh
	<i>speechmatics</i>	uh ; yeah; know; bye; well
	<i>hubert</i>	ha; e; ye; would; h
	<i>wav2vec2.0</i>	ha; e; would; ye; may
	<i>whisper</i>	yeah; mmhmm ; know; uh ; okay

3.4.3 Uhm or ehm? ASR transcription of non-lexical tokens

Having established that NLTs are less frequent in ASR than manual transcription, and that NLTs comprise the bulk of ASR transcription errors, we explore NLT transcription in further detail. This is because NLTs carry meaning and should hence be transcribed. For instance, K. Latif et al. (2023) note in clinical settings, patients often reply with NLTs like *mh-hum* to questions like “Are you allergic to any antibiotics?”. The performance of NLT transcription is therefore a critical requirement when assessing the suitability of ASR for transcription in many applications.

Deriving a list of NLTs In order to consider NLT transcription in further detail, we derive a list of NLTs from available documentation shown in Table 3.6. The Speechmatics “disfluency” tag contains 41 NLTs. Switchboard transcripts have 14 “disfluency” tags and RoomReader transcripts have 10 NLTs tagged as fillers and disfluencies. Using this list of NLTs, we then compute deletion, substitution, and insertion errors involving NLTs, e.g. ignoring *boat*→*coat*

Table 3.6 Non-lexical tokens (NLTs) alongside the source from which they were identified

Source	List of NLTs
<i>Switchboard annotations</i>	un, um, huh, ha, eh, ooh, oh, er, duh, hm, ah, uh, ugh, hum
<i>Speechmatics disfluency tagging</i>	huh, aha, ah, aw, eh, err, hmm, mm, um, uh, uhoh, uhhuh, uhuh, mhm, aha, aah, aahh, aaw, ahha, ahaa, ahh, ahha, aww, eeh, erm, hhm, hhmm, hm, huhuh, mhm, uggh, ugh, ughh, uhh, uhhm, uhm, uhmm, umm, uuh, uuhh, uum
<i>RoomReader annotations</i>	uhn, huh, ah, um, eh, oh, hum, hm, uh, uhm

but counting $uhm \rightarrow uhh$. We further categorise substitution errors into those where an NLT is swapped for another NLT or for a non-NLT, using the logic that an NLT-to-NLT substitution is less problematic than a deletion or substitution to a non-NLT.

NLTs error analysis Table 3.7 presents the results of this analysis. Open-source ASR transcripts contain 79-86% less NLTs than ground-truth transcriptions (see final two columns). Commercial systems preserve more NLTs, but the overall number is reduced. Further exploring substitutions reveals an important difference between commercial and open-source ASR. Commercial ASR correctly transcribes 37%-69% of NLTs and, importantly, where a substitution occurs, 76-89% of the time the NLT is swapped for another NLT (NLT $\rightarrow$ NLT). Thus, the location, but not necessarily the spelling, of a large majority of NLTs is preserved. However, in open-source ASR, 76-89% of substitutions are to a non-NLT (NLT $\rightarrow$ non-NLT), indicating a large majority of NLTs are removed by the ASR systems.

NLTs performance metrics We report three performance metrics for NLT transcription in Table 3.8. First, the classic WER is computed from all NLT insertion, selection and substitution errors using the same formula as listed in Section 3.3.1. However, insertions, deletions and substitutions are now computed only for the NLTs, which are listed in full in Table 3.6. We note the WER is high for all commercial ASRs at 62-63% on RoomReader, and 31-42% on Switchboard (Table 3.8). Open-source systems are much worse, with the WER ranging from 87-94%, indicating a large majority of NLTs are mistranscribed.

We then calculate an *adjusted* WER (adj-WER). This is similar to the classic WER, however we only count those substitutions where an NLT is replaced with a *non*-NLT token. The analysis reveals the adj-WER of commercial ASR is considerably lower than the classic WER (22%-26% on RoomReader and 13-31% on Switchboard) as NLT $\rightarrow$ NLT substitutions are not penalised. Finally, we compute the percentage of NLTs that are correctly transcribed or substituted for another NLT (percentage recall). This is 69-87% for Amazon and Speechmatics and 13-31% for open-source ASR. The performance of Google on Switchboard was

Table 3.7 Error analysis of non-lexical tokens (NLTs) in ASR transcriptions of RoomReader (RR) and Switchboard (SB). The number of correctly matched NLTs as well as deletions and substitutions are shown and expressed as a percentage of the number of ground truth (GT) NLTs. The number of times one NLT is swapped for another NLT (NLT→NLT) or non-NLT (NLT→non-NLT) are shown and expressed as a percentage of the total number of substitutions. The number of NLTs in the output is shown, and the percentage reduction relative to the ground truth number of NLTs.

ASR	Corpus	NLT Deletions	NLT Substitutions	NLT Matches	NLT → NLT	NLT → non-NLT	GT NLTs	ASR NLTs
amazon	RR	156 (7%)	1254 (56%)	840 (37%)	924 (74%)	330 (26%)	2250	1764 (↓22%)
	SB	9594 (7%)	31290 (24%)	89104 (69%)	23976 (77%)	7314 (23%)	129988	113080 (13%)
google	RR	1619 (72%)	537 (24%)	94 (4%)	2 (0%)	535 (100%)	2250	96 (↓96%)
	SB	28805 (22%)	26181 (20%)	75002 (58%)	16124 (62%)	10057 (38%)	129988	91126 (30%)
speechmatics	RR	272 (12%)	1125 (50%)	853 (38%)	814 (72%)	311 (28%)	2250	1667 (↓26%)
	SB	32616 (25%)	20889 (16%)	76447 (59%)	13346 (64%)	7543 (36%)	129952	89793 (31%)
hubert	RR	492 (22%)	1627 (72%)	131 (6%)	188 (12%)	1439 (88%)	2250	319 (↓86%)
	SB	26003 (20%)	93963 (73%)	9286 (7%)	11588 (12%)	82375 (88%)	129252	20874 (84%)
wav2vec2.0	RR	1040 (51%)	720 (35%)	282 (14%)	107 (15%)	613 (85%)	2042	389 (↓81%)
	SB	90908 (70%)	14986 (12%)	24083 (19%)	3544 (24%)	11442 (76%)	129977	27627 (79%)
whisper	RR	207 (9%)	1846 (82%)	197 (9%)	199 (11%)	1647 (89%)	2250	396 (↓82%)
	SB	25019 (19%)	89918 (70%)	14315 (11%)	12504 (14%)	77414 (86%)	129252	26819 (↓79%)

Table 3.8 Metrics for non-lexical tokens (NLTs). The classic WER penalises misspellings of NLTs, whereas the adjusted WER (adj-WER) does not. The % recall is the combined percentage of NLTs that are correctly identified or misspelt.

ASR/corpus	classic WER ↓		adj-WER ↓		% recall ↑	
	RR	SB	RR	SB	RR	SB
amazon	63	31	22	13	78	87
google	97	42	99	30	4	70
speechmatics	62	41	26	31	74	69
wav2vec2.0	94	93	86	84	14	16
whisper	87	82	82	79	19	21
hubert	91	89	82	79	18	21

comparable to other commercial systems; however, its poor performance on RoomReader is noteworthy.

We then report the percentage recall by computing the percentage of NLTs in the ground-truth transcription that are transcribed as an NLT, ignoring any substitution errors.

3.4.4 Word timing and overlap

As word timings may be required in a transcript for analysis, we analyse the accuracy of ASR word alignments. We use the Switchboard corpus as it contains a ground-truth word alignment. In Figure 3.2, we show how various ASR models transcribe the same utterance, noting the word timings and the NLT *uhm-hum*. We compute the absolute difference between

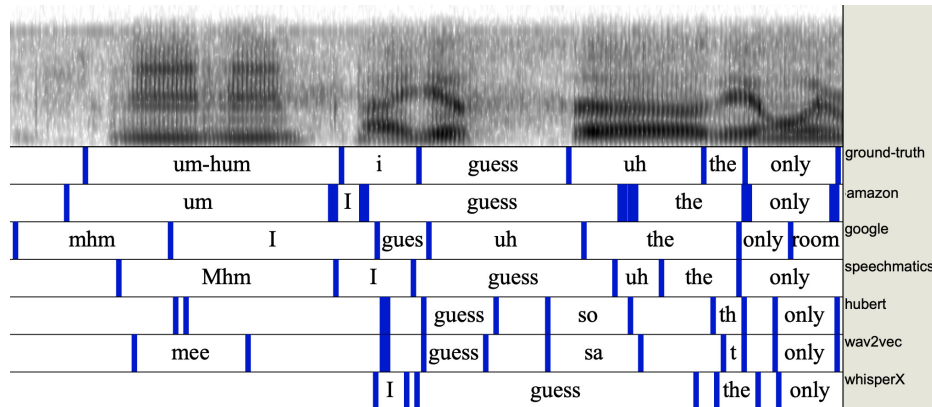


Figure 3.2 Excerpt from session sw2666 of Switchboard shown in Praat. The TextGrid contains the ground-truth transcript and several ASR transcripts. Note the lexical differences in the transcription across the ASR systems, and the differences in word timing.

Table 3.9 Word alignment and absolute onset and offset error in seconds (median per word), and the percentage of speech and overlapping speech (median per session) in Switchboard.

	median (25%,75%) / word		median (25%,75%) / session	
	onset-error (s)	offset-error (s)	% speech	% overlap
<i>ground-truth</i>			82 (78, 90)	15.15 (8.57, 26.1)
<i>amazon</i>	0.50 (0.1, 1.5)	0.53 (0.1, 1.5)	84 (80, 90)	13.31 (7.63, 22.19)
<i>google</i>	0.51 (0.2, 1.5)	0.53 (0.2, 1.5)	89 (85, 90)	15.8 (8.8, 26.9)
<i>speechmatics</i>	0.48 (0.1, 1.4)	0.54 (0.1, 1.5)	74 (74, 90)	5.62 (5.62, 12.12)
<i>hubert</i>	0.54 (0.1, 1.5)	0.45 (0.1, 1.4)	45 (43, 50)	3.98 (1.92, 7.2)
<i>wav2vec2.0</i>	0.54 (0.1, 1.5)	0.45 (0.1, 1.3)	44 (41, 50)	3.87 (1.58, 7.66)
<i>whisperX*</i>	0.56 (0.1, 1.6)	0.54 (0.1, 1.5)	61 (58, 62)	16.08 (9.1, 25.01)

the human and ASR start and end times of each word (the *onset error* and the *offset error*; see method in Section 3.3.2). Table 3.9 shows that the median onset and offset timing errors are approximately 500 ms.

Next, we assess the proportion of the interaction that is marked as containing speech by each ASR model. As can be seen in Table 3.9, Amazon and Google ASR detect 84% and 89%, close to the ground-truth 82%. Other ASR systems detect less speech at 74-44%. This can mean that words are removed from the transcript, or the start and end of words are cut, reducing the overall amount of speech detected. Table 3.9 also shows the percentage of speech which occurs in overlap. Amazon, Google and Whisper record 13.31-16.08% overlapping speech compared with the ground-truth of 15%. This indicates that overlapping is captured well by these systems. Wav2vec2.0, Hubert and Speechmatics record less overlap at 3.8-5.62%. We provide an example of overlapping speech transcription in Figure 3.3.

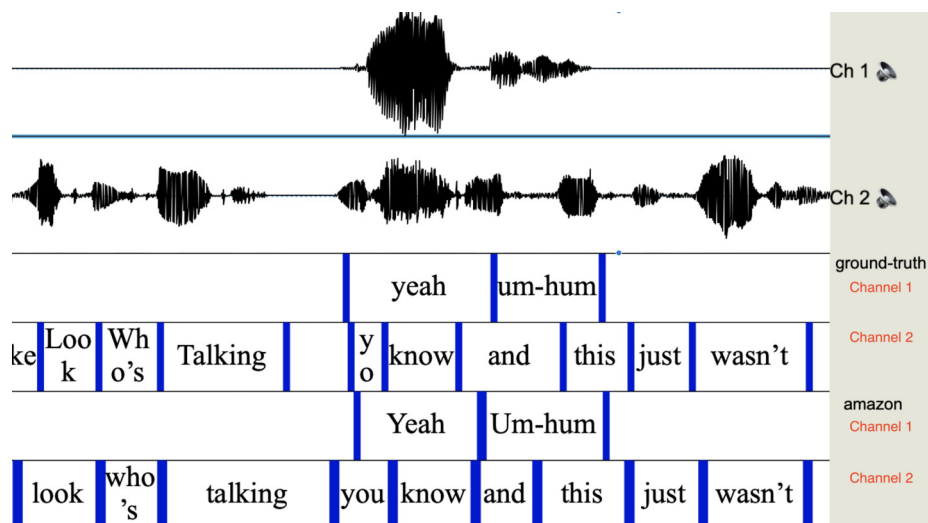


Figure 3.3 Excerpt from session sw2666 of Switchboard shown in Praat. Channels 1 and 2 show the audio of both speakers. The TextGrid contains the ground-truth transcripts and the Amazon transcripts for both channels.

3.5 Discussion of results

Our study provides a comprehensive analysis that uncovers what ASR can and cannot do as a transcription tool. In this section, we discuss the results of our analysis and the extent to which ASR fulfils the four transcription requirements set out in Section 3.1. We then consider the practical implications of our results for the use of ASR in conversational speech transcription in Section 3.6.

3.5.1 Capturing what was said

We evaluated ASR using four general requirements of conversation transcription set out in Section 3.1. Our first requirement of an ASR system was that it should transcribe exactly what was said by the interlocutors. We assessed this capability using the WER, which rewards exact spelling matches, and penalises deletions and substitutions of words.

We found commercial ASR systems transcribe read speech with a 2 to 8% WER (see Table 3.3). For the Switchboard corpus of telephone conversation, the WER scores were higher at 7.9% to 10.2%, yet still within Microsoft's definition of a "ready-to-use" ASR (Microsoft, Inc. 2024). The WER was highest on the RoomReader corpus. Speechmatics and Amazon ASR achieved a 16.1% and 15.7% WER, still indicative of reasonably good performance. Google achieved WER scores similar to other commercial platforms on Switchboard. However, its poor WER on RoomReader underlines the fact that ASR performance can vary hugely across corpora and conversation domains (Morbini et al. 2013). It may also reflect in part the challenging nature of videoconferencing, which disrupts conversational flow (Sellen 1995;

O’Conaill et al. 1993). WERs were higher for students than for tutors in RoomReader, which likely stems from ASR performing poorly on short bursts of speech (Stolcke and Droppo 2017), commonly uttered by students in these tutor-led interactions (Reverdy et al. 2022). Furthermore, tutors are native speakers of English, whereas students are a mix of L1 and L2 speakers (Reverdy et al. 2022). ASR performance is worse for L2 speakers (Graham and Roll 2024).

Open-source ASR is attractive to researchers as it is free-to-use. However, we found that open-source tools had double the WER of commercial ASR systems on conversational speech (see Table 3.3). Taking the example of Whisper, it achieved the lowest WER rate on conversational speech at 19.4% on Switchboard and 26.7% on RoomReader. This compares with 7.9% and 16.1% for Amazon on the same corpora. This indicates the presence of considerably more transcription errors in the open-source ASR transcripts.

3.5.2 Transcribing NLTs: *Uhm* or *uh-uhm*?

Our second requirement was that NLTs should be accurately transcribed. We showed that up to 78% of NLTs, short words such as *uh-hum*, were preserved by commercial ASR (see Table 3.7). However, commercial ASR often substitutes one NLT for another (e.g. *uhm*→*ehm*). These findings are in line with prior work demonstrating that ASR struggles with spelling NLTs (Stolcke and Droppo 2017). However, our results differ from Lopez et al. (2022) as we show that commercial ASR systems do not delete NLTs. Rather, we found that the correct location of the majority of NLTs is preserved, but with a high rate of NLT→NLT substitution. On the other hand, we find that open-source ASRs remove a large majority of NLTs from the transcript. They also frequently change NLT spelling into a non-NLT category, making it unidentifiable, further contributing to the gap between open-source and commercial ASR performance.

3.5.3 Capturing overlapping speech and word alignment

Our final evaluation criterion for transcription concerned the accuracy of word timing and the transcription of overlapping speech. We found that ASR generally handled overlapping speech well (see Table 3.9). Our findings differ from those of Liesenfeld et al. (2023) who found that ASR removes *all* overlapping speech. Speaker diarisation could be a reason behind this dramatic removal of overlap, as diarisation algorithms are known to remove overlap (Park et al. 2022). Diarisation can be avoided if the audio is well separated with one channel per speaker, as done here. If possible, channel separation can be ensured at data collection stage through the use of directional or head-mounted microphones. This may not be possible in settings where ambient microphones are required.

For word alignment, we found a 500 ms error in word timing, which largely affected all systems (see Table 3.9).

3.6 How can ASR support transcription?

In this section, we discuss the implications of our findings for the use of ASR in transcription. We highlight a number of considerations that researchers should make when evaluating the suitability of ASR for conversational speech transcription. Furthermore, we outline limitations and identify directions for future work.

3.6.1 Caution on the use of the WER

The results of our analysis have shown that all ASR systems perform well on read speech, but on conversational speech a performance gap emerged between commercial and open-source tools. The most frequent errors involved non-lexical tokens (NLTs) such as ‘*uhm*’ and ‘*uh-hum*’. Commercial ASR preserves the location, but not necessarily the spelling of most NLTs, whereas open-source systems frequently removed NLTs from the transcription altogether.

Our NLT analysis therefore highlights why the WER alone should not be used to assess transcript suitability. The WER rewards the correct spelling and placement of all words equally, irrespective of their role or importance. Thus, a few keywords in the conversation may be consistently mistranscribed, despite the WER indicating good performance. This can occur in a variety of settings, such as when uncommon or specialist medical or engineering terms are present. As these infrequent but important terms are often mistranscribed by ASR (Morbini et al. 2013; Devine et al. 2000), this may have little impact on the overall WER but it makes an ASR transcription unsuitable without manual correction. This can mean that the ASR has a competitive WER, but the high rate of NLT mistranscription makes it nevertheless unsuitable in many transcription tasks. K. Latif et al. (2023) give the example of the *uh-hum* response to a doctor’s question in place of the word *yes*. This must be recorded accurately if ASR is to support medical note-taking, irrespective of the overall WER.

3.6.2 ASR as an initial transcription aid

In certain applications, researchers may be able to directly use the output of an ASR system (Gorisch and Schmidt 2024). However, our findings show a large number of mistranscribed NLTs. This means the ‘raw’ ASR transcription will require manual correction. We therefore recommend that researchers use ASR to obtain an initial transcript, which can then be corrected.

As manual transcription from scratch can take many times longer than the length of the audio itself, a quality ASR transcript can save a researcher a significant amount of time (Umair et al. 2022; García Molina et al. 2024). Gaur et al. (2016) found that manually correcting an ASR transcription was up to 30% faster than transcribing from scratch, provided the WER was less than 30%. Ma et al. (2024) found that although various transcribers transcribe at different speeds, it was always faster to correct a transcription than create one from scratch. Thus, a high-quality ASR transcript is a “*good enough*” starting point, significantly speeding up the overall transcription process (Umair et al. 2022; Bokhove and Downey 2018).

Thus, our work shows that ASR can be of much benefit in the types of transcription highlighted in Chapter 2 that are conducted by hand in research at present. For example, transcribing interview recordings for qualitative research. As researchers in these areas expect that transcription is a manual and necessary part of the work, ASR can be a valuable tool which provides an initial transcription which can be manually inspected and corrected. On the other hand, there are domains where practitioners would presumably prefer an ASR to provide a transcript which is usable without hand-correction, such as transcribing a patient-doctor consultation or medical dictation. Our work highlights that ASR cannot replace humans in these settings in its present state, due to transcription errors such as the frequent misspellings of NLTs.

3.6.3 How much time can be saved by using ASR?

The time saved by the initial ASR transcript depends on the amount and type of errors in the transcript. Bokhove and Downey (2018) demonstrated that the time saved by having the transcription is reduced as the WER increases (i.e. as the number of errors increases). Papadopoulou et al. (2021) found that deletion and substitution errors slow down correction less than misspellings. A WER of 30% is considered the upper limit for a transcription that is to be corrected by hand, as it is faster to transcribe from scratch than to correct an ASR transcript with a WER that exceeds 30% Gaur et al. (2016).

However, our findings show that an upper limit on the WER is far more useful if considered in conjunction with the NLT error rate of an ASR system. This is because spelling serves as an indicator of the pragmatic function of a word. For example, filled pauses typically written as *uhm* or *ehm* are common in spontaneous speech (Benus 2009; H. H. Clark and Fox Tree 2002). When produced before a turn, they signal an intent to speak (H. H. Clark and Fox Tree 2002; Benus 2009). On the other hand, backchannels serve a different function. They are typically written as *uh-hum* (note the presence of two syllables). They signal understanding without intent to take a turn or an explicit agreement (Goodwin 1986). As there is no one-to-one mapping between spelling and pragmatic function, coding is a process where an annotator identifies and labels pragmatic function. Coding would likely be

complicated by the high level of NLT misspellings and deletions found in this work, slowing down the hand-correction process.

The 500 ms error in word timings indicates the transcription is not accurate enough for phonetic research, which typically requires an accuracy of 20-50 ms (Goldman 2011; McAuliffe et al. 2017). However, in a wide array of tasks where an aligned transcription is required, an approximate location and timing of utterances may be entirely sufficient, for instance, live captioning. The ASR transcript could also speed up manual phonetic alignment, a particularly laborious task (Goldman 2011).

3.6.4 On replacing human transcribers

Whether an ASR could ever truly replace a human annotator is an interesting question to consider. We focussed on speech-to-text transcription, which uses audio only. When interlocutors can see one another, like in a videoconference call or in-person, nonverbal cues become utterances in and of themselves, for example, nodding to signal agreement (Knapp et al. 1978; Isaacs and J. C. Tang 1993). This may need to be included in the transcription.

As ASR is inherently limited to audio, naturally, it cannot transcribe these types of utterances. Consequently, many utterances delivered exclusively nonverbally are likely missed in the ASR transcription of RoomReader in this work, e.g. hand raising. These will always have to be manually added to an ASR transcript. An ASR with perfect accuracy theoretically could replace a human when transcribing the audio from a telephone conversation, where visual cues are not present.

3.6.5 Coding of pragmatic and paralinguistic phenomena

The speech-to-text transcription considered in this work is also just one form of transcription. Transcription may also encompass the subsequent coding of backchannels, fillers, yes-no responses, etc. (Calhoun et al. 2010; Mateer and Taylor 1995). This involves assigning words such as *yes*, *uh-hum* and *ehm* to a series of categories such as *agreement*, *backchannel*, *filled pause* and so on. At the time of writing, ASR simply does not support this form of transcription. Additionally, various paralinguistic phenomena may need to be included, such as laughter, which is included in Switchboard (Godfrey et al. 1992). However, to our knowledge, mainstream ASR systems do not support this annotation.

3.6.6 Additional factors to consider in an ASR deployment

There are a number of additional factors that a researcher should consider when using ASR, which we highlight here.

Performance limitations of ASR Researchers should be aware of the performance limitations ASR. Particularly close attention should be paid to speakers for whom ASR is known to display a negative performance bias, e.g. those with regional dialects, non-native speakers, and older speakers (Feng et al. 2024; Sanabria et al. 2023; Vippera et al. 2008). Researchers should also be aware that poor recording quality can negatively impact ASR transcription accuracy: noise significantly degrades performance even in state-of-the-art ASR tools (Agrawal and Ganapathy 2019; Alharbi et al. 2021). If possible, sources of noise including environmental noise (e.g. a fan in the room) and reverberation (echoes in the room) (Agrawal and Ganapathy 2019) should be identified and reduced. Another step which ensures optimal recording quality is the choice of hardware: poor-quality microphones can degrade ASR performance (Alharbi et al. 2021). Microphone bleed can be avoided with a good choice of head or lapel-mounted microphones. Where this is not possible, speaker diarisation may be used. However, researchers should have a reasonable expectation of performance, since the transcript will likely contain diarisation errors (Park et al. 2022).

Commercial ASR cost considerations Researchers should also consider the cost of commercial ASR. As of January 2024 when the ASR models in this study were run, these are 0.024-0.030 USD per minute for the standard EN-US models used in this study. These costs add up for larger corpora. For example, it would currently cost 2,448 USD to transcribe the 850-hour Candor corpus (Reece et al. 2023) if each channel is billed separately. Models are available for specialist tasks such as medical transcription, which can recognise specific terms. The cost is higher for these models, for example, 0.075 USD per minute for Amazon Medical Transcribe at the time of writing.

Pros and cons of commercial and open-source ASR For some, the cost and privacy benefits of open-source ASR may outweigh performance issues. Open-source ASR systems confer a significant advantage to researchers as they are free to run, although there is the initial cost of a GPU-enabled workstation to consider. Furthermore, not all researchers wishing to use ASR will have the necessary working knowledge of GPU libraries. In contrast, commercial systems have well-documented APIs requiring minimal coding knowledge.

Open-source ASR is more configurable as the source code can be easily accessed. This enables models to be fine-tuned on extra data. For example, fine-tuning on conversations which have specific medical or technical terms may avoid the increased cost of specialist commercial models. However, the downside is that the training requires labelled data, time, and knowledge. The ability to run locally is a key privacy advantage of open-source ASR. All of the audio recordings and output transcripts remain in the researcher's machine or network, whereas commercial ASR requires a remote server, possibly in a different jurisdiction. This

is an important advantage as in a research setting, the conditions of ethical approval may preclude the transfer of data to a third party or different jurisdiction.

3.6.7 Limitations

Our study of conversational speech transcription with ASR presents some inevitable limitations, highlighted in this section.

ASR models are continually updated Commercial ASR providers regularly publish updated models and open-source models will also progress. Thus, the results of this study are a snapshot of performance in Q1 of 2024. Incremental version updates may alter the exact details of performance reported, but the central message of our work remains constant: researchers wishing to incorporate ASR systems into their transcription methodology need to take a close look at their performance. They should not only consider the WER, but also assess the suitability of the transcription for their specific purpose, particularly when conversational speech is involved.

Fine-tuning and training data Another limitation is that we did not conduct any fine-tuning, i.e. optimising the ASR with a small amount of labelled data. This can considerably improve performance (Baevski et al. 2020). We chose to evaluate ASR *off-the-shelf* reflecting common use, e.g. (K. Latif et al. 2023; Loakes 2024).

Some ASR models in our study were trained on LibriSpeech (e.g. wav2vec 2.0). No public information regarding the training data or procedure is available for commercial ASR systems. However, LibriSpeech is a common benchmark within the ASR community, which is why commercial models are likely optimised for it. We used the *test* and *dev* portions of LibriSpeech, which are not used for model training by convention. As RoomReader is only available for non-commercial use, it was not used to train any of the ASR models. Key results are consistent across all models on RoomReader and Switchboard (with the notable exception of Google on RoomReader). This makes it unlikely that the choice of datasets impacted the key findings of this work.

The conversational medium The datasets also differ in their medium (telephone speech vs. videoconferencing speech). A limitation of our work is that we did not include recordings from in-person interaction, due to the lack of suitable corpora of in-person interaction with high-quality transcripts. In-person recordings also introduce uncontrolled sources of noise such as echo and reverb from the room and microphone bleed, all of which degrade ASR performance, complicating analysis (Alharbi et al. 2021; Yu and Deng 2015; Park et al. 2022). Furthermore, NLT annotation is often not conducted or documented in the corpus

release. Thus, we chose RoomReader and Switchboard as they contain clean recordings of spontaneous speech with NLTs captured and recorded in the data (Calhoun et al. 2010; Reverdy et al. 2022).

The list of NLTs may not be exhaustive The NLT list in Table 3.6 may not be exhaustive. Nevertheless, our list is sufficiently comprehensive to appreciate the transcription discrepancies between open-source and commercial ASR. Furthermore, as Ward (2006) highlights, a finite NLT list likely does not exist. It is also possible that certain ASR systems may be "editing out" NLTs, as is the case for Whisper (Radford, Kim, et al. 2023). However, there are no easily-accessible options within any of the models to control for this and source code modifications are beyond the scope of this work. We did not consider lexical interjections such as *gosh* as we found NLTs are amongst the most common ASR errors (see Table 3.5). We imagine that lexical interjections are better transcribed by ASR, but leave this to future work. We also could not consider technical differences between the ASR models, e.g. language models in the decoder, due to a lack of technical detail for commercial ASR.

The use of English, a high-resource language Finally, our work only considered transcriptions in English, a high-resource language. Thus, performance here represents a best-case scenario and will likely be worse for other languages (Radford, Kim, et al. 2023). For example, Gorisch and Schmidt (2024) found that ASR did not significantly speed up transcription time using Whisper to transcribe German. As German typically has a higher WER than English (Radford, Kim, et al. 2023), it would therefore require more manual correction post-ASR.

3.6.8 Future research directions in conversational speech transcription

There are a number of directions for future work. The NLT substitution errors should be further examined. For example, is *uh-hum* typically substituted for *eh-hem*, preserving the syllable structure, or *uhm*, preserving the vowel quality? This could be reported as a WER-based metric for each of the individual NLTs and it could be compared across the different ASR models.

Future work should explore the relationship between ASR transcription quality and the time taken to correct a transcript versus transcribing from scratch. This should be further studied for languages other than English. We did not consider punctuation as we removed all punctuation prior to computing the WER and conducting our analyses, as is commonplace in the ASR literature, e.g. (Radford, Kim, et al. 2023). There is evidence that punctuation errors are particularly laborious to correct (Papadopoulou et al. 2021), so future work should consider the type and amount of punctuation errors produced by various ASR tools.

Other directions for future work include the dialect-specific differences in the pronunciation of NLTs (Wong and Peters 2007). ASR transcription across speakers and accents should

also be considered (Graham and Roll 2024). Child-directed speech should be explored, as it has unique acoustic properties (Klis et al. 2023), as should speech in neurodiverse adults. Finally, the real-time performance of ASR is important as this is required for human-robot interaction (Baumann et al. 2017).

In many commercial ASR platforms, in order to mitigate common errors lists of commonly-mistranscribed words can be easily added by the researcher to improve their transcription. For example, adding specialised medical or engineering terms using the *custom vocabularies* feature in Amazon Transcribe⁵. However, it is likely not as straightforward as using these features in order to improve NLT transcription, given the rich phonetic variation and dialect specific differences in NLT pronunciation (Wong and Peters 2007), all of which are highly suitable for future work. Furthermore, these customisation features are not present in open-source ASR tools, where researchers have far less control of the ASR model output.

3.7 General conclusion

Manual transcription is costly and laborious, yet a transcript is a necessary prerequisite in any work which analyses conversation. In this chapter, we found that commercial ASR can produce a high quality transcript. We found that open-source ASR has substantially more errors than commercial ASR. We found that a notable difference in commercial and open-source systems is that commercial ASR captures more NLTs such as *uh-hum*, though misspellings are common. We discussed how cost and privacy advantages of open-source ASR may outweigh performance issues.

We therefore conclude that manual correction of an ASR transcript remains necessary, although modern commercial ASR systems produce transcriptions of sufficiently high quality for manual correction to be feasible.

3.7.1 Turn-taking model development with ASR transcription

By presenting a thorough analysis of ASR, our work has broad application in a number of fields. The findings of this chapter are also essential in supporting the work of the following chapters. We have ensured to use Switchboard and RoomReader, conversational speech corpora with quality transcriptions, enabling a benchmark of ASR on conversational speech. This enables us to use ASR to transcribe corpora for which there is no high-quality transcription available, such as the Candor corpus. This is especially important for multimodal turn-taking models, where the lack of manually transcribed datasets presents a major challenge for model development and analysis. In later chapters, we can refer back to our benchmarking

⁵<https://docs.aws.amazon.com/transcribe/latest/dg/custom-vocabulary.html>

exercise in this chapter to gain a clear understanding of transcription accuracy and the limitations of automating transcription with ASR.

Chapter 4

Predictive turn-taking with audio

For the remainder of this thesis, we direct our focus to predictive turn-taking models (PTTMs): neural networks which predict upcoming transitions between speakers in a conversational interaction. As we outlined in Chapter 2, most PTTM development to date has focussed on speech, which contains rich prosodic, lexical and syntactic cues linked to turn-taking. In this chapter, we therefore focus on models which predict turn-taking with audio. We establish the performance of a state-of-the-art audio-only turn-taking model from the literature, the voice activity projection (VAP) model, which we fully re-train.

This chapter first addresses **Research Question 2**, which concerns the use of automatic speech recognition (ASR) for PTTM training and development. It is unknown whether ASR is sufficiently accurate to train a PTTM. We therefore train the VAP model on the Switchboard corpus of telephone speech (Godfrey et al. 1992), with the provided alignment and an alignment from ASR transcription. We show the VAP model has no significant difference in performance when trained on manual and ASR alignment. This work builds on Chapter 3, where we highlight the potential of ASR to vastly expand the available data for machine learning, including turn-taking models. It also supports much of the work in the following chapters, where we use ASR for PTTM training.

Next, we address **Research Question 4**, which deals with whether audio can support accurate turn-taking prediction in interactions where interlocutors can see as well as hear one another. The VAP model is at an inherent disadvantage in such interactions, as interlocutors have access to non-verbal information, but the model does not. Given the focus on telephone speech in the literature, it is unknown whether an audio-only PTTM works on such interactions. We address this by training the VAP model on the audio from the Candor corpus of videoconferencing interaction (Reece et al. 2023). We find the audio-only VAP model achieves high accuracy, demonstrating that audio alone can support turn-taking where visual cues are available to the interlocutors. However, we show that a model trained on Switchboard cannot accurately predict turn-taking in a model trained on Candor, and vice versa. We discuss possible reasons for this discrepancy, which include the impact that visual

cues in the Candor corpus might influence the way in which interlocutors use acoustic and prosodic cues to manage turn-taking.

The work in this chapter lays the groundwork for Chapter 5, where we further probe the role of prosodic and lexical cues in turn-taking prediction with audio and large language models. By focussing solely on speech, the comprehensive analysis in these chapters also motivates Chapter 6, where we introduce the visual modality into turn-taking prediction.

Chapter structure

The remainder of this chapter is laid out as follows. In Sections 4.1 and 4.2, we set out the training and evaluation procedure for the model. We outline a comprehensive evaluation protocol, which ensures the model is validated on a more expansive set of turn-taking events than is typically considered in the literature.

In Section 4.3, we present the results of the VAP model when trained on the Switchboard corpus, with both ground-truth alignment and the ASR alignments. Section 4.4 presents the findings of the VAP model trained on the Candor videoconferencing corpus. Finally, Section 4.5 presents the findings of the cross-corpus study, where we train the VAP model on Switchboard or Candor, and test the performance on the other corpus.

Research questions addressed

The work in this chapter addresses the following research questions:

- **Research Question 2** *Can state-of-the-art PTTMs from the literature be trained with automatically-transcribed data?*
- **Research Question 4** *Are visual cues necessary for accurate turn-taking prediction for interactions where participants can see as well as hear each other?*

Publications

The work in this chapter has led to the following publications:

- The analysis of the VAP model on Switchboard and Candor corpora is presented in our ACL Findings publication *Visual Cues Enhance Predictive Turn-taking for Two-party Human interaction*, [O'Connor Russell](#) and Harte (2025).
- The cross-corpus study is presented in our MMSYM publication *Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction* [O'Connor Russell](#) and Harte (2024).

Table 4.1 Voice activity projection (VAP) model architectural details, FFNN=feed-forward neural network

Model Block	Model parameter	Value
Audio encoder	encoder type	CPC (Riviere et al. 2020)
	audio sample rate	16000 Hz
	prediction sample rate	50 Hz
	sequence length	20 seconds
	audio encoder frozen	True
	audio encoder pre-trained	True
Self-attention block	FFNN dimensionality	768
	model dimensionality	256
	number of attention heads	4
	number of layers	1
Cross attention block	model dimensionality	256
	number of attention heads	4
	number of layers	3

4.1 Training an audio-only turn-taking model

To address the research objectives of this chapter, we use the VAP model, a state-of-the-art audio-only turn-taking model from the literature outlined in Section 2.6.5. We train the model using either the Switchboard corpus of telephone speech or the Candor corpus of videoconferencing speech. The corpora are described in detail in Section 2.3. The first part of this chapter outlines the training and evaluation procedures for the VAP model.

4.1.1 Model technical specification

We re-implement the VAP model with reference to the original authors' publicly available code, to allow for maximum control and of model training and to support our analysis. The VAP model architecture is outlined in Section 2.6.5 of Chapter 2. We train the model from scratch, developing our own data loading, preprocessing, and training components independently. All model implementation is in PyTorch (Paszke et al. 2017). The model uses a pre-trained audio encoder, the contrastive predictive coding (CPC) model from Riviere et al. (2020). This encoder is loaded using the publicly available weights, which are then frozen during training. All other layers of the model are randomly initialised. A technical specification of the model is provided in Table 4.1 below.

4.1.2 Audio and label preparation

We resample audio to 16 kHz, the required sample rate of the CPC model (Riviere et al. 2020). The audio from Candor is downsampled to 16 kHz from 48 kHz. The audio from Switchboard requires no resampling as it is already at 16 kHz. Loading and resampling of

Table 4.2 Overview of the three VAP models trained in this chapter. The corpora and training labels for each model are listed.

Training labels derived from	VAP model trained on	Labels derived from	Preprocessing
ground-truth alignment	Switchboard	Alignment provided with corpus	Removal of paralinguistic annotations (e.g., laughter)
ASR alignment	Switchboard	Speechmatics ASR transcript	—
ASR alignment	Candor	Speechmatics ASR transcript	—

audio is conducted in Torchaudio (Hwang et al. 2023). We conduct amplitude normalisation at the speaker level. Both audio tracks are in stereo, with one channel per speaker.

We utilise the word alignment from transcriptions to generate training labels for the model. The labels describe speaking activity in the next 2000 ms. The labels and extraction process are outlined in full in Section 2.6.5. A manual transcription is provided with Switchboard, which we pre-process by stripping paralinguistic labels such as *laughter*¹. For ASR transcriptions of both Switchboard and Candor, we use the Speechmatics² transcription service previously evaluated on conversational speech in Chapter 3. No pre-processing is required as the transcription does not contain any paralinguistic labels. Note that the transcriptions are used for timing information only; no lexical information is used (Section 2.6.5).

This results in three sets of training labels for the two corpora, listed in Table 4.2. We train a VAP model on each of the three corpora and label combinations.

4.1.3 Cross validation

For training each model, we conduct a 5-fold cross-validation. We withheld 122 sessions from Switchboard and 70 sessions from Candor at random as a test set. The remaining sessions are used in the 5-fold cross-validation. The number of sessions and mean duration in hours of each fold is shown in Table 4.3. Note that the total number of sessions in Candor is 1,658, but we train on a subset of 1,415 sessions. This is for consistency with Chapter 6, as explained further on in Section 6.1. We randomly sample batches across all of the training sessions within a cross-validation fold. Due to GPU memory constraints, each item in a batch consists of a 20-second chunk of audio. We train each fold for 10 epochs with AdamW optimiser (Loshchilov and Hutter 2017) in PyTorch using the default parameters³. The learning rate and batch size are set by grid search (Section 4.1.4). The loss function is described in Section 2.6.5. We report the average training and validation set losses for each training epoch.

¹This is for consistency across models trained with manual and ASR alignments, as the ASR tools we use do not transcribe these phenomena (Chapter 3).

²We use Speechmatics as we received academic access to the platform, which enabled the complete transcription of the 850 hr Candor corpus.

³`betas=(0.9, 0.999), eps=1e-08, weight_decay=0.01`

Table 4.3 Number of sessions and duration (mean duration in hours per fold and for train/val and total duration in hours for test) of the cross-validation splits for Switchboard and Candor corpora.

Split	Switchboard		Candor	
	# Sessions	Hours	# Sessions	Hours
Train (per fold)	1,855	197	1,077	604
Val (per fold)	461	49	269	151
Test	122	13	70	40
Total	2,438	260	1,416	710

4.1.4 Hyperparameter selection

We set the learning rate and batch size (the hyperparameters) by a grid search. This involves training the model on a number of different learning rates and batch sizes. We choose 5 learning rates and 5 batch sizes. We consider learning rates 1×10^{-6} , 1×10^{-4} , 3.63×10^{-4} and 1×10^{-3} , where 3.63×10^{-4} was the learning rate used by Ekstedt and Skantze (2022b). The batch sizes are 4, 8, 16, 32 and 64 where 64. To select optimal hyperparameters, we train a model with each learning rate and batch size pair. The Switchboard corpus is used, with labels from the ground-truth alignment. The first fold of cross-validation is used. All training is conducted on an NVIDIA RTX 6000 GPU.

The minimum loss on the validation set for each pair of hyperparameters is in Figure 4.1. The loss is achieved with a learning rate of 1×10^{-4} and batch size 32. As the loss is the same for a batch size of 16, we choose the larger batch size of 32 for faster training.

4.1.5 Model training procedure

Having selected optimal hyperparameters, we train the VAP model twice on Switchboard (ASR and ground-truth alignment) and once on Candor (Table 4.2). The train and validation losses are plotted for each epoch in Figure 4.2. The training loss shows all models achieve good convergence as it steadily decreases with each epoch. The model trained with the Switchboard ASR labels has a higher loss than the model trained on the ground-truth alignment. Note that these losses are in agreement with previously-reported values (Inoue, Bing'er Jiang, et al. 2024).

The loss on the validation set is shown in Figure 4.2b. The average loss decreases in all models as the number of epochs increases. This shows that there is no overtraining and the learning rate is appropriate. Multiple informal attempts were made to adjust the learning rate, but no improvement was noted. The epoch number with the lowest validation loss is tabulated in Table 4.4. Weights from these epochs are used in all subsequent analyses.

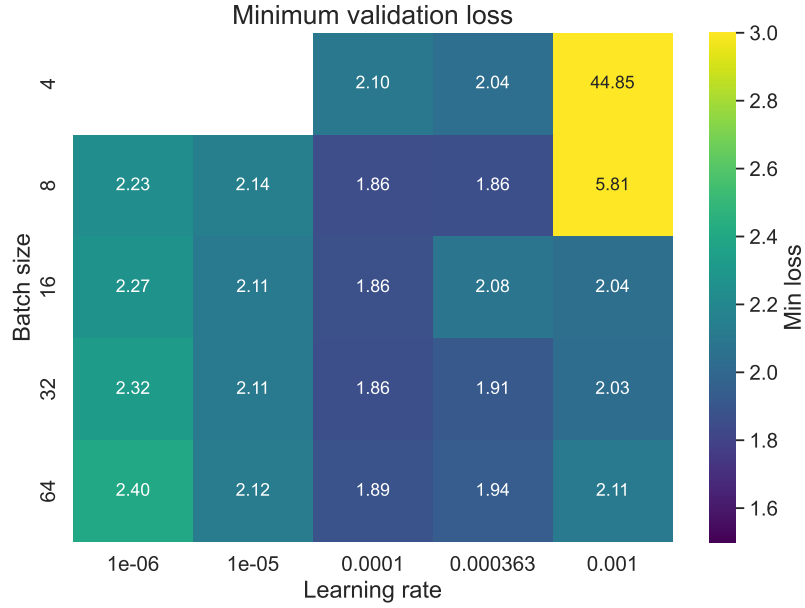


Figure 4.1 Validation set loss for each pair of learning rate and batch size hyperparameters. The blank squares indicate the model did not converge for this pair of hyperparameters.

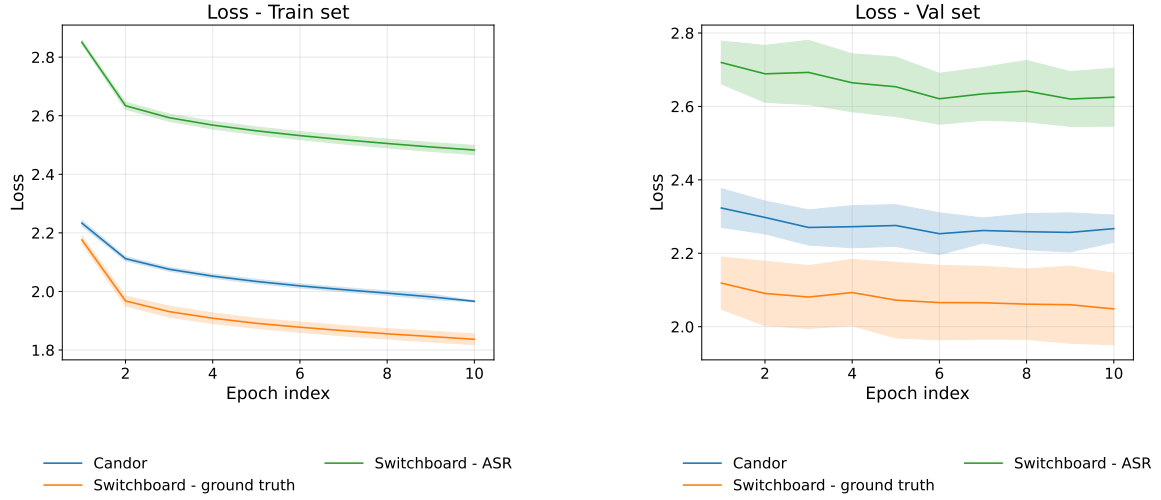
Table 4.4 Epochs corresponding to the lowest validation loss for each fold across different corpora.

Lowest Validation Loss Epoch					
Corpus	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
SWB (ground-truth)	5	8	8	1	9
SWB (ASR)	9	8	8	1	2
CND	6	2	8	6	5

4.1.6 Running the trained models and extracting p_{future}

We run trained models by windowing the validation and test set audio into 20-second chunks, with an overlap of 10 seconds. We discard predictions from the first 10 sections of each window and merge the last 10 seconds of each window. We repeat the procedure for each fold of cross-validation.

This gives a 256 dimensional VAP prediction label at 50 Hz for each training and validation set. This is the raw output of the model, which describes future speaking activity in the next 2000 ms, discussed in Section 2.6.5. For evaluation, we transform this into a compact scalar probability, p_{future} , which represents the scalar probability of speech from either interlocutor in the next 800-2000 ms. The procedure for label extraction and for obtaining p_{future} from the model are outlined in full in Section 2.6.5.



(a) Training set loss

(b) Validation set loss

Figure 4.2 Training and validation losses (mean per batch). The average across 5 folds is plotted for the 10 training epochs, with ± 1 standard deviation.

4.2 Metrics for turn-taking model evaluation

Next, we describe the evaluation procedure for the trained models. First, the time of holds (no speaker change) and shifts (speaker changes) are extracted using the procedure outlined in Section 2.6.5. We then extract 7 turn-taking metrics, which evaluate the performance of the trained VAP models, as is standard in the turn-taking literature (Ekstedt and Skantze 2022b; Roddy, Skantze, et al. 2018a).

We categorise the metrics into *hold/shift prediction* and *shift prediction*, which are described in Sections 4.2.2 and 4.2.3. We show how this evaluation protocol considerably expands the number of evaluation points when compared to prior work in the turn-taking literature. In Section 4.2.4, we outline how we obtain balanced accuracy and $F1$ scores for each metric.

4.2.1 Identifying shift and hold times

We extract holds and shifts using the procedure outlined in Section 2.6.5 of Chapter 2. Table 4.5 shows the number of shifts and holds grouped by the minimum floor-transfer offset (FTO). This is the period of silence ($FTO > 0$ ms) or overlap ($FTO < 0$ ms) between adjacent speaking turns. Note that the ASR and ground-truth alignments result in different numbers of holds and shifts in Switchboard. We also note from the table that holds are inherently more frequent than shifts in the corpora, with the notable exception of the ASR. This reflects the differences in the start and end times of words across different ASR tools.

Table 4.5 Statistics for shifts and holds under varying minimum FTO thresholds. The FTO is measured as the time between the end of one speaker’s turn and the beginning of the next turn (negative indicates overlap). The overall number of shifts and holds is shown, as well as the proportion of events that are shifts and the number of shifts and holds per minute.

Corpus / FTO Threshold	Min. FTO [ms]	# Shifts	# Holds	Shift Proportion	Shifts / min	Holds / min
<i>Switchboard</i> (ground-truth alignment) 260 hours	-250	6599	9419	0.70	0.42	0.61
	0	26883	150407	0.18	1.73	9.67
	250	16267	74510	0.22	1.05	4.79
	500	9909	43793	0.23	0.64	2.82
	750	6087	27348	0.22	0.39	1.76
	1000	3893	14180	0.27	0.25	0.91
	1250	2407	6958	0.35	0.15	0.45
	1500	1385	3348	0.41	0.09	0.22
<i>Switchboard</i> (ASR alignment) 260 hours	-250	10595	7761	1.37	0.68	0.50
	0	34886	210386	0.17	2.24	13.53
	250	20097	143182	0.14	1.29	9.21
	500	10302	74666	0.14	0.66	4.80
	750	5269	33299	0.16	0.34	2.14
	1000	2702	14952	0.18	0.17	0.96
	1250	1383	6599	0.21	0.09	0.42
	1500	663	2786	0.24	0.04	0.18
<i>Candor</i> (ASR) 710 hours	-250	23515	62483	0.38	0.42	1.11
	0	115975	414666	0.28	2.05	7.34
	250	83158	206830	0.40	1.47	3.66
	500	53360	115475	0.46	0.94	2.04
	750	32831	65804	0.50	0.58	1.16
	1000	19699	35697	0.55	0.35	0.63
	1250	11438	19831	0.58	0.20	0.35
	1500	6148	10250	0.60	0.11	0.18

4.2.2 Hold/shift prediction metrics

The first set of metrics is hold/shift prediction, which evaluates the ability of the trained model to distinguish holds (end of turn where there is a speaker change) from shifts (end of turn where there is no speaker change). Diagrams of shifts and holds are provided in Section 2.4.6. This metric was first considered by Skantze (2017). We evaluate hold/shift prediction during mutual silence between turns, before the end of a turn, and before overlapping speech. This is a more comprehensive evaluation than prior work, which only considered hold/shift prediction during mutual silence (Skantze 2017; Ekstedt and Skantze 2022b).

During mutual silence We first consider hold/shift prediction following the original VAP paper by Ekstedt and Skantze (2022b). Hold/shift prediction was evaluated during the mutual silence between turns, after a turn had ended. As minimum FTO duration was not specified in their work, we choose a 250 ms as this is slightly above the minimum duration of a noticeable silence to a human listener (200 ms, Walker and Trimboli (1982)). After each hold and shift, we sum p_{future} probabilities of the next speaker in a 200 ms window. The summed

probabilities are then used to obtain a prediction of a hold or shift from the model in each window.

Before the end of a turn We then assess model performance at predicting an upcoming speaker change before the end of a turn, while speech is ongoing. This was not considered in the VAP paper (Ekstedt and Skantze 2022b). We include it as the model needs to be able to distinguish upcoming shifts from holds whilst speech is ongoing, akin to human turn-taking (Levinson and Torreira 2015; Sacks et al. 1974). Before each hold and shift, we sum p_{future} prediction scores of the next speaker in a 200 ms window. As before, summed probabilities are used to obtain a prediction of a hold or shift in each window.

We repeat the process for both $FTO > +250$ ms and $FTO > +0$ ms. This removes the requirement of a minimum period of silence between turns enforced in earlier work (Ekstedt and Skantze 2022b). The number of holds and shifts available for evaluation increases. In Candor, the total number of holds used in the evaluation goes from 16,267 to 26,883 and shifts from 74,510 to 15,0407 (Table 4.5): equivalent to a 65% and 102% increase, respectively. It also more closely reflects real-world conditions where turns often occur with very little silence between adjacent turns (Sacks et al. 1974).

Before overlapping speech Although earlier work on other models did consider how models handle overlapping speech (Roddy, Skantze, et al. 2018a), overlap was not considered in the evaluation of the VAP model (Ekstedt and Skantze 2022b; Inoue, Bing'er Jiang, et al. 2024). This is surprising as speaker transitions with overlap are common in conversational speech (Sacks et al. 1974; Heldner and Edlund 2010), e.g. in 23,515 turns in Candor and 10,595 turns in Switchboard (Table 4.5). A large number of speaker transitions would therefore be omitted if overlapping speech were ignored. We consider hold/shift prediction performance before overlapping speech by summing p_{future} prediction scores for the current speaker immediately before the overlap, obtaining a hold/shift prediction in each window. A diagram of a speaker transition with a short period of overlapping speech is provided in Section 2.4.6.

4.2.3 Shift prediction metrics

A good turn-taking model should be able to distinguish shifts from mid-turn speech. This ensures that the model does not falsely predict a transition in the middle of a turn. This capability is measured by the shift-prediction metric, also introduced by Ekstedt and Skantze (2022a). To compute the metric we follow the same process as before, summing p_{future} of the most recently active speaker in a 200 ms window before a shift. We then sample an equal number of 200 ms windows of mid-turn speech at random. These are located at least

Table 4.6 Turn-taking metrics, evaluation points and minimum floor-transfer offset (FTO) value between adjacent speaking turns. The probability p_{future} of either the speaker active after the event (hold/shift prediction) or the current speaker (shift prediction) is summed over a 200 ms window located at the evaluation point.

Metric category	Distinguishes	Evaluation point	p_{future} from speaker	Criteria
Hold/shift prediction	shifts from holds	during mutual silence before end of turn before end of turn before overlap	next active speaker	FTO > +250 ms FTO > +250 ms FTO > 0 ms FTO < -250 ms
Shift prediction	shifts from mid-turn speech	before end of turn	current active speaker	FTO > +250 ms FTO > 0 ms

1500 ms from the start or end of a turn. We sum p_{future} scores for the current speaker within a 200 ms window during the mid-turn speech. The summed probabilities are then to obtain a prediction of a hold (mid-turn speech) or shift from the model.

4.2.4 Summary of key turn-taking metrics

The 7 metrics and evaluation points are summarised in Table 4.6. By removing the requirement for silence between speaking turns and including overlapping speech, our evaluation protocol expands the number of evaluation points, making the evaluation better representative of real conversational speech.

As we noted, the summed probabilities are used to extract hold and shift predictions. We convert the summed probabilities to a binary hold/shift prediction using a threshold, reporting performance in terms of the F_1 and balanced accuracy scores. We choose a threshold to convert summed p_{future} into binary predictions which maximise the F_1 and balanced accuracy on the validation set for each metric. Finally, we report test set accuracy for each fold, reporting the average performance.

4.2.5 Baseline model

We compare all models against a dummy baseline model, which never predicts a speaker change (always hold). The derivation is derived and reported in Appendix A.

4.3 Audio-only turn-taking on the Switchboard telephone speech corpus

We begin by assessing the VAP model trained on the Switchboard corpus. We report the 5-fold average performances in Tables 4.7 and 4.8. Receiver operating characteristic (ROC) and precision-recall (PR) curves are reported in Figure 4.3. In Section 4.3.6, we show an

Table 4.7 Average hold/shift prediction across on the Switchboard (SWB) corpus. The 5-fold average test set balanced accuracy and F_1 scores are shown.

Hold/shift prediction metrics, Switchboard ground-truth and ASR alignment					
Evaluation point	Corpus	F_1 (Weighted)	F_1 (Hold)	F_1 (Shift)	Accuracy (Balanced %)
during mutual silence ($FTO > +250$ ms)	SWB (ASR)	0.81	0.90	0.45	65
	SWB (ground-truth)	0.82	0.90	0.47	66
before end of turn ($FTO > +250$ ms)	SWB (ASR)	0.79	0.89	0.40	63
	SWB (ground-truth)	0.80	0.89	0.41	63
before end of turn ($FTO > 0$ ms)	SWB (ASR)	0.83	0.91	0.42	65
	SWB (ground-truth)	0.84	0.91	0.43	65
before overlap ($FTO < -250$ ms)	SWB (ASR)	0.62	0.56	0.66	61
	SWB (ground-truth)	0.64	0.58	0.67	63

example of the trained model predicting a transition between speakers and discuss qualitative performance.

4.3.1 Performance on hold/shift prediction metrics

First, we consider the performance of the VAP model using the hold/shift prediction metrics. We begin with the model that was trained on the Switchboard corpus with the ground-truth alignment. Evaluation of the model during the mutual silence between speaking turns archives a 66% balanced accuracy score, and weighted $F_1 = 0.89$ and shift $F_1 = 0.40$ (*during mutual silence ($FTO > +250$ ms)*, Table 4.7). Next, we consider how performance changes when we move the evaluation window to before the end of the turn. Balanced accuracy falls slightly to 63%. Inspection of the F_1 shift scores reveals a slight performance decrease (0.47 to 0.41, *before end of turn ($FTO > +250$ ms)*, Table 4.7). This shows that predicting turn-taking while speech is ongoing is harder than deciding who speaks next after the turn has ended.

We then consider the performance when relaxing the requirement of a minimum FTO duration between speaking turns (*before end of turn ($FTO > 0$ ms)*, Table 4.7). This metric has not been included in prior work on the VAP model. It results in more turns being used in the evaluation of the model (Table 4.5). Balanced accuracy increases slightly (63% vs 65%), showing robust performance is maintained as faster speaker transitions are included.

Finally, we consider performance on overlapping speech, which is common in human speech (Sacks et al. 1974), yet not included in the VAP model paper (Ekstedt and Skantze 2022b). We find a balanced accuracy of 63%. Interestingly, the F_1 shift score is higher than the F_1 hold score for overlapping speech. The model performs better in predicting shifts than holds prior to moments of overlap, which is the opposite of the other metrics. The decreased overall accuracy shows that predicting turn-taking before overlapping speech is a challenging task for a predictive turn-taking model.

Table 4.8 Average shift prediction across on the Switchboard (SWB) corpus. The 5-fold average test set balanced accuracy and F_1 scores are shown.

Shift prediction metrics, Switchboard ground-truth and ASR alignment					
Evaluation point	Corpus	F_1 (Weighted)	F_1 (Hold)	F_1 (Shift)	Accuracy (Balanced %)
<i>before end of turn (FTO > +250 ms)</i>	SWB (ASR)	0.73	0.76	0.70	73
	SWB (ground-truth)	0.74	0.76	0.71	73
<i>before end of turn (FTO > 0 ms)</i>	SWB (ASR)	0.75	0.77	0.74	76
	SWB (ground-truth)	0.75	0.77	0.74	76

4.3.2 Performance on shift prediction metrics

Next, we consider performance with the shift prediction metric. Results are listed in Table 4.8. The balanced accuracy of the model trained on Switchboard with the ground-truth alignment is 63%, with a minimum period of silence between turns (before end of turn, $FTO > +250$ ms). Inspection of the F_1 scores shows the model performs better at identifying holds (mid-turn speech) than shifts. We find that the shift prediction score increases slightly as the requirement for a minimum period of silence relaxes (compare *before end of turn, $FTO > +250$ ms* with *$FT > 0$ ms*, Table 4.8). Comparing the F_1 hold and shift scores for this metric reveals that the performance is more balanced in this metric than in the previous hold/shift prediction metric (compare F_1 (hold) and F_1 (shift) columns in Tables 4.8 and 4.7).

4.3.3 Comparing models trained on manual and ASR alignments

Next, we compare models trained with manual and ASR word alignments. Two-way paired t-tests compare the average performance of models trained on both ASR and ground-truth alignments on the common test set. The full list of paired t-tests is presented in Section B.2 in Appendix B. Paired t-tests reveal there is largely no difference in the performance of the hold/shift prediction scores between Switchboard ground-truth and ASR trained models. Where there is a significant difference, the difference is negligible, e.g. weighted F_1 0.89 vs 0.90, $p = 0.02$. For the shift prediction metric, we find no significant differences in the vast majority of metrics comparing the ground-truth and ASR models.

The lack of any meaningful difference in performance of ASR and ground-truth alignment therefore validates the use of ASR in PTTM training and development. As comparable performance can be achieved with ASR and ground-truth alignments, this opens the door for development of PTTMs on large corpora which are not provided with a manual transcription; for instance, the Candor corpus used in Section 4.4.

4.3.4 Performance relative to baseline model and the literature

As a sanity check of our models, we assess performance relative to a baseline model which always predicts a hold (A). One-sample t-tests compare the average weighted F_1 and

balanced accuracy to the baseline model. They show both Switchboard ground-truth and ASR outperform⁴ the baseline. The t-tests are listed in Tables B.1 and B.2 of Appendix B.

In comparison with reported performances from the literature, our results are broadly in line with Ekstedt and Skantze (2022b), who used an earlier version of the VAP model (hold/shift, during mutual silence weighted $F_1 = 0.89$, shift $F_1 = 0.5$, and hold $F_1 = 0.73$). There is a slight discrepancy between our hold/shift performance and theirs. However, it is unclear what the minimum period of silence between speaking turns was in their analysis, as it was not reported. Furthermore, the full cross-validation split used in the paper has not been published.

4.3.5 Analysing performance with ROC and PR curves

Previously, we converted p_{future} to binary hold/shift prediction scores by choosing a threshold which maximised model performance on a validation set (Section 4.2.4). This is the typical procedure in the literature. Receiver operating characteristic (ROC) and precision-recall (PR) curves enable a much richer analysis of model performance by considering performance across all thresholds. To our knowledge, they have not been applied to PTM analysis.

Consider the example of the trained VAP model deployed in a human-robot interaction. Ideally, the robot should always predict a shift where a shift should actually occur, i.e. a true positive rate of 100% and a false positive rate of 0%. Now consider a dummy robot which randomly chooses between shifts and holds. This dummy robot model would have a false positive rate which linearly increases with the true positive rate. A real-world model ideally allows the robot to take the turn when it should (high true positives), yet it does not interrupt when the speaker wishes to continue speaking (low false positives). By examining ROC and PR curves, we can adjust the model's decision threshold to achieve a balanced performance.

We plot ROC and PR curves of VAP models for each fold of cross-validation in Figure 4.3. Focussing on the hold/shift prediction task in Figure 4.3a, we show that performance is above the random baseline indicated by the diagonal line, which indicates the performance of a random model. The curves show a true positive rate of ≈ 0.90 can be achieved with a false positive rate of ≈ 0.6 . This means that the robot correctly takes the turn in the overwhelming majority of instances, at the expense of interrupting relatively frequently. On the other hand, a more aggressive optimisation would reduce the false positive rate to 0.2. However, this corresponds with a reduction in the true positive rate to ≈ 0.75 . This would mean the robot fails to take turns slightly more often, but it interrupts far less often, highlighting the trade-off that must be made when deploying a trained model. Comparing the ROC curves for the VAP model trained on ground-truth and ASR alignments in Figures 4.3a and 4.3c, we note that

⁴Note that for the ASR-trained model, there is no significant difference between weighted F_1 and F_1 hold scores of the baseline and trained model. However F_1 shift and balanced accuracy are significantly above baseline performance.

the lines become slightly closer to the diagonal in the ASR trained model. The ROC curves for the shift prediction task are shown in Figures 4.3a and 4.3a. Performance is again above baseline in both models, with little difference in ASR and ground-truth aligned models.

Next, we consider the PR curves. Precision (proportion of shifts predicted that are truly shifts) is plotted against recall (proportion of all shifts that are correctly predicted as shifts) at each threshold. An ideal classifier should have a high precision and a high recall. Focussing on the hold/shift prediction in Figure 4.3a, the PR curve indicates a good trade-off between precision and recall. A recall of ≈ 0.75 can be achieved with a precision of ≈ 0.4 in Figure 4.3a. Some of this performance is lost when moving to the ASR alignment, indicating a slightly increased error (Figure 4.3c, right-hand side).

This analysis shows that the model is well-trained as it can be optimised to have a high true positive rate and a low false positive rate. The curves are tightly together in each fold, indicating little performance variation across cross validation folds. Thus, the VAP model has a good trade-off between predicting transitions between speakers correctly, whilst minimising the number of false positive predictions. This corresponds with a model which takes the turn when it should, whilst minimising interruption.

4.3.6 Performance on an example from the corpus

We show an example of the VAP model’s ability to predict speaker transitions on an exchange between speakers from a test set session in Figure 4.4. It depicts a transition from one speaker depicted in grey who asks “*uh what kind of books do you enjoy reading*”? to another speaker depicted in red, who responds “*uh having have uh [sic] i have a wide variety ...*”. The waveform is shown on top, with the first speaker depicted in red and the second speaker depicted in grey. Note the model does not use the tokens from the transcript, but the ground-truth transcription is superimposed for visualisation. Notice the period of silence between the speaking turns.

The probability that the speaker in red will speak in the next 800-2000 ms is plotted in the lower half of the figure. We show the output of both ground-truth (GT, dashed line) and ASR (solid line) trained models. Where p_{future} exceeds 0.5, this is highlighted in red⁵. The plot shows that both models predict the transition to the speaker in grey confidently and ahead of time. The p_{future} of both models rapidly shifts from 0.2 to 0.8 in the final hundred milliseconds of the turn of the speaker in grey. Before the end of the grey speaker’s turn, both models are confidently predicting that the next speaker, depicted in red, will speak in the future. There is minimal variation in the output of both models. An interesting observation about this example is the disfluent speech at the start of the speaker in grey’s turn “*uh having have uh*”. There is a large pause in the speaker’s turn, which is approximately equal to the

⁵note this threshold is not the same as the threshold previously optimised on the validation set. We choose 0.5 for visualisation purposes as both speakers are equally probable at this point

gap between the speaking turns. During this pause, p_{future} correctly remains high in both models.

4.3.7 Discussion

The results in this section demonstrate that the VAP model can distinguish between holds and shifts in the Switchboard corpus. The model performs above baseline, and our reported performances are in line with the literature and above a baseline model. We expanded the analysis to include more metrics and turn-taking events than is typically considered, and found above baseline performance on all metrics. The ROC curves show that the model has a decent trade-off between minimising false positive shift predictions, whilst correctly identifying all shifts in the corpus. We did find that performance on holds was higher compared to shifts, also noted in prior work (Ekstedt and Skantze 2022b). Table 4.7 shows that the $F1$ score for holds is much higher than for shifts. This reflects the fact that turn-taking is a highly challenging problem, given the fast-paced nature of speaker transitions (Skantze 2021).

An important finding in this section is that we found no meaningful differences in the performance of the VAP model when trained and tested on word alignments derived from ASR and ground-truth transcription. This supports Research Question 2 of this thesis. This finding, in conjunction with our analysis of ASR in Chapter 3, enables us to utilise the Candor videoconferencing corpus in the following sections of this chapter.

4.4 Audio-only turn-taking on the Candor videoconferencing corpus

In this section, we continue by exploring the VAP model performance when trained and tested on the Candor videoconferencing corpus. The corpus is of videoconferencing interactions, but the VAP model only uses the audio from the interactions. As ASR transcription provided with the corpus is unreliable and there is no ground-truth transcription, we use Speechmatics ASR for transcription and obtaining the necessary word timestamps for training. An advantage of running ASR on the corpus is that we use an ASR system which we thoroughly evaluated (including on videoconferencing speech) in Chapter 3.

4.4.1 Performance on hold/shift prediction metrics

Hold/shift prediction metrics for the Candor corpus are presented in Table 4.9. We include performance on Switchboard for reference. Evaluation of the model during mutual silence achieves a balanced accuracy of 79% (*during mutual silence*, $FTO > +250$ ms). Next, we consider evaluation before the end of a turn, noting a balanced accuracy of 77% (Table 4.9).

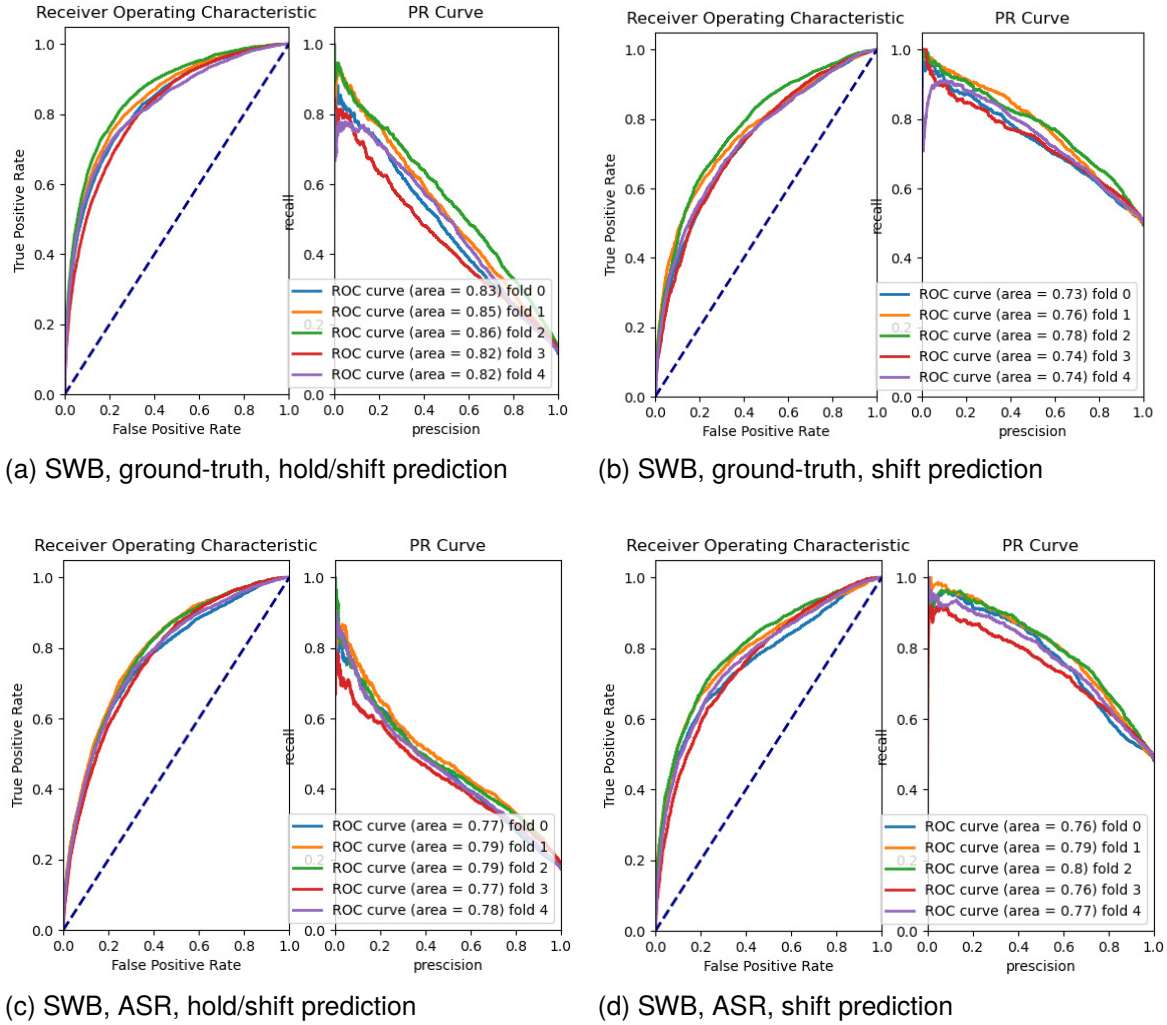


Figure 4.3 ROC curves for the voice activity projection (VAP) turn-taking model trained on the Switchboard (SWB) corpus (ground-truth and Speechmatics ASR alignments). Performance on hold/shift and shift metrics is illustrated (FTO > +250 ms).

This indicates that predicting a speaker transition while speech is ongoing is more difficult than predicting a transition during the silence between speaking turns.

We then remove the requirement for a minimum period of silence between speaking turns. As is the case in Switchboard, this leads to the inclusion of more holds and shifts for evaluation in the corpus (Table 4.5). We find that this has a negligible impact on the balanced accuracy in Candor 78% vs 79%. Finally, we consider the model performance in distinguishing holds and shifts before a period of overlapping speech. We find a balanced accuracy of 70%, indicating reduced performance.

Exploring the F_1 shift and F_1 hold score reveals that, compared with Switchboard, on Candor there is more balanced performance between hold and shift classes. For example, on

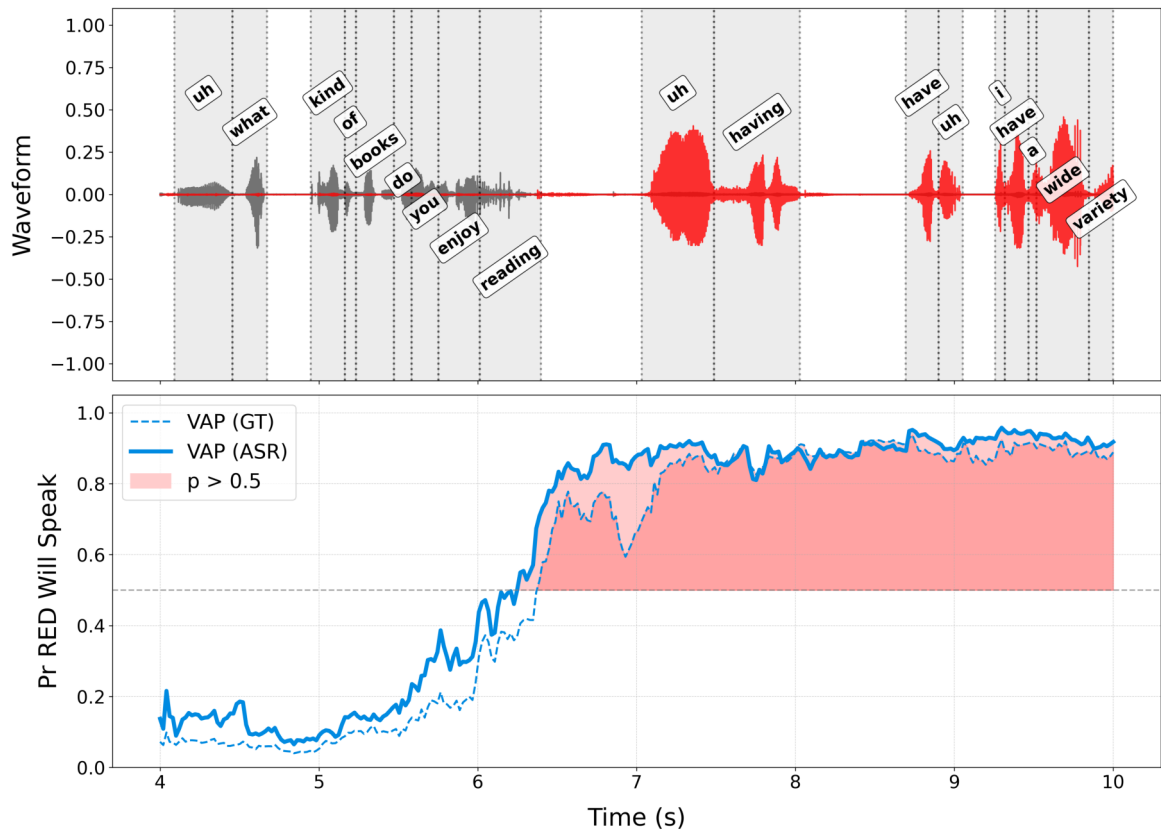


Figure 4.4 Example of model output on session ID 2024 of the Switchboard corpus (test set). The example depicts a transition from one speaker (grey) to another (red). The trace of the p_{future} probability of the speaker depicted in red from VAP models trained with the ground-truth (GT) and ASR alignments is shown.

Table 4.9 Average hold/shift prediction across 5 folds on the Candor (CND) and SWB (ASR) corpora.

Hold/shift metrics, Switchboard and Candor corpora					
Evaluation point	Corpus	F_1 (Weighted)	F_1 (Hold)	F_1 (Shift)	Accuracy (Balanced %)
during mutual silence ($FTO > +250$ ms)	CND	0.83	0.88	0.70	79
	SWB (ASR)	0.81	0.90	0.45	65
before end of turn ($FTO > +250$ ms)	CND	0.81	0.87	0.67	77
	SWB (ASR)	0.79	0.89	0.40	63
before end of turn ($FTO > 0$ ms)	CND	0.86	0.91	0.66	78
	SWB (ASR)	0.83	0.91	0.42	65
before overlap ($FTO < -250$ ms)	CND	0.78	0.85	0.57	70
	SWB (ASR)	0.62	0.56	0.66	61

Candor, F_1 hold is 0.88 and F_1 shift is 0.70 for hold/shift prediction during mutual silence. On the other hand, on Switchboard F_1 hold is 0.90, whereas F_1 shift is 0.45. This general trend is repeated in the other metrics apart from overlapping speech, where the trend is reversed

Table 4.10 Results for shift prediction evaluation metrics across on Candor (CND) and Switchboard (SWB) on the test set

Shift prediction metrics, <i>Switchboard and Candor corpora</i>					
Evaluation point	Corpus	F ₁ (Weighted)	F ₁ (Hold)	F ₁ (Shift)	Accuracy (Balanced %)
<i>before end of turn (FTO > +250 ms)</i>	CND	0.86	0.84	0.87	85
	SWB (ASR)	0.73	0.76	0.70	73
<i>before end of turn (FTO > 0 ms)</i>	CND	0.86	0.84	0.87	86
	SWB (ASR)	0.75	0.77	0.74	76

(0.57 vs 0.66 F_1 hold and shift). This means that the VAP model is better at detecting shifts in the Candor corpus than in the Switchboard corpus.

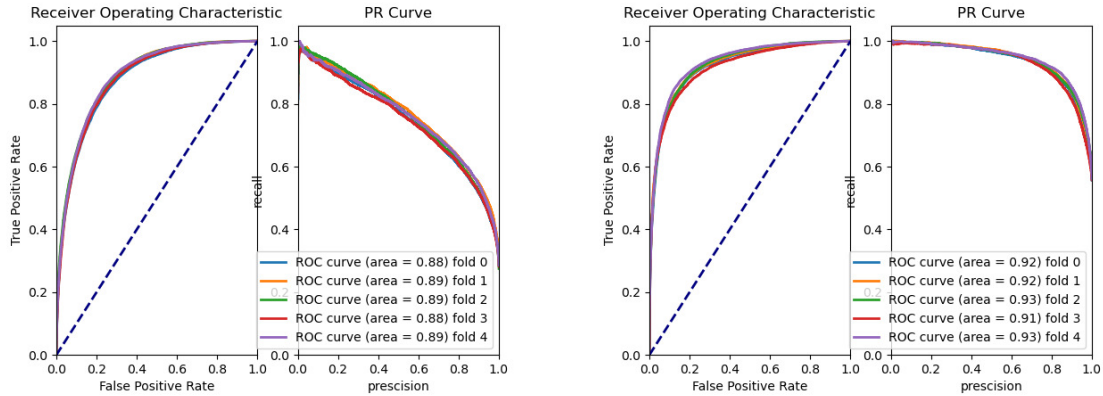
4.4.2 Performance on shift prediction metrics

Next we consider the shift prediction metric which considers the ability of the model to distinguish shifts with random mid-turn speech. Performance on the test set is shown in Table 4.10. Performance on Candor is 85% before the end of the turn, with $FTO > +250$ ms. Performance remains high when the requirement of a minimum FTO is removed ($FTO > 0$ ms). This contrasts with Switchboard (ASR), where performance increased slightly. The F_1 hold and shift scores reveal balanced performance across shift and mid-turn speech (hold) on the Candor corpus.

4.4.3 Performance relative to baseline model and the literature

We compare performance with a baseline model which always predicts a hold. The baseline model performance is derived in Appendix A.2. One sample t-tests reveal the Candor trained model outperforms the baseline for both hold/shift and shift prediction metrics. The t-tests are listed in full in Table B.3 in Appendix B.

To our knowledge, the VAP model has not been tested on the Candor corpus. There are therefore no reported performances to compare against; however our findings show that the model performs well on a comprehensive set of turn-taking metrics. This is significant as the Candor corpus is a corpus of videoconferencing interactions; interlocutors can see as well as hear one another. The role of visual cues in turn-taking is well-established (Chapter 2), yet we find good performance using audio alone. This suggests that the use of audio alone is sufficient to achieve turn-taking prediction with high accuracy in videoconferencing interactions. Nevertheless, our literature review in Chapter 2 highlights the supportive role of non-verbal communication in turn-taking. The inclusion of non-verbal cues therefore has the potential to improve this performance, a question which we return to in Chapter 6.



(a) CND, ground-truth, hold/shift prediction

(b) CND, ground-truth, shift prediction

Figure 4.5 ROC curves for the voice activity projection (VAP) turn-taking model, trained on on the Candor (CND) corpus (Speechmatics ASR alignment). Performance on the hold/shift and shift metrics are illustrated (FTO > + 250 ms). The metric shown is hold/shift prediction during mutual silence and shift prediction before the end of a turn.

4.4.4 Analysing performance with ROC and PR curves

ROC and PR curves are plotted in Figure 4.5. For hold/shift prediction, the ROC curve is far above the baseline model indicated by the diagonal line (Figure 4.5a). This indicates the model can be tuned to have a high true positive rate low false positive rate, e.g. a ≈ 0.8 true positive rate is achievable with a low ≈ 0.2 false positive rate. The PR curve reveals a high line, similarly showing a balanced precision and recall. Next, we consider shift prediction, where the ROC curve similarly indicates that the model is well trained (Figure 4.5b). The PR curve indicates that this model can achieve a high precision and recall. The curves are tightly bunched, which indicates little variance in test set performance across the 5 folds of cross validation.

The ROC and PR curves for the model trained on the Candor corpus show that a higher true positive rate is achievable for the same false positive rate in comparison with the model trained on Switchboard (Figure 4.3). This increased performance is reflected in their increased $F1$ and accuracy scores we previously noted in Tables 4.9 and 4.10, where we compare Candor and Switchboard (ASR) models.

4.4.5 Performance on an example from the corpus

Next, we show an example of the Candor trained VAP model predicting a speaker transition in Figure 4.6. The example is sourced from a session in the test set of the corpus. The example depicts a transition from one speaker (grey) to another (red). The waveform is depicted on top with the words and word boundaries for illustration. The speaker in red asks “*what are*

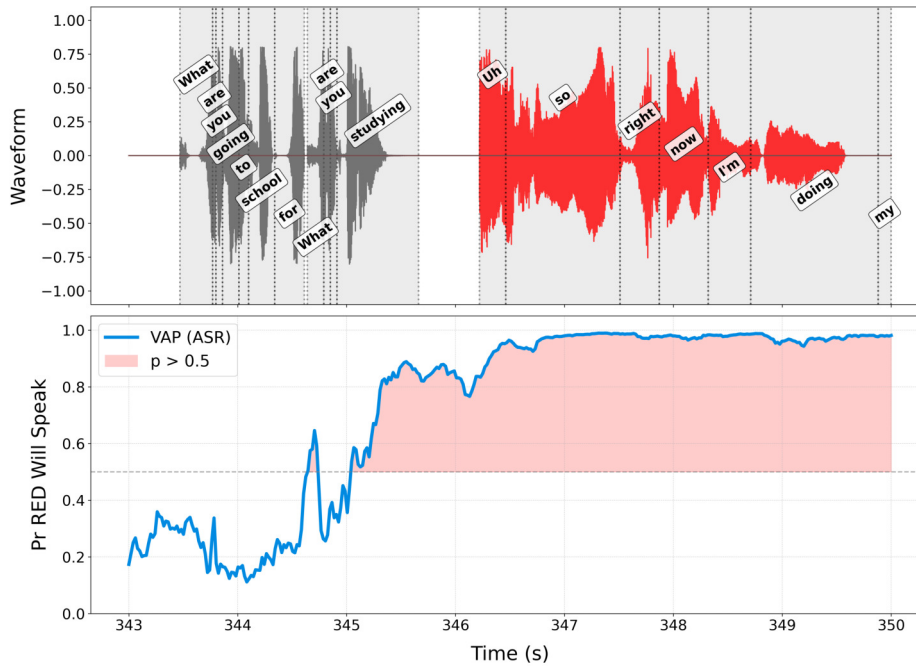


Figure 4.6 Example of model output on session ID 67a84a1f-e628-40d8-a000-baac829fd25f of the Candor corpus (test set). The example depicts a transition from one speaker (grey) to another (red). The trace of the p_{future} probability of the speaker depicted in red of VAP model trained on Candor is shown.

you going to school for what are you studying”? to which the speaker in grey responds “*yeah so right now im doing my ...*”. There is a period of silence between the speaking turns.

The probability of speech from the speaker depicted in red in the next 800-2000 ms, p_{future} , is plotted in the bottom half of the figure. The example shows that during the last word of the turn of the speaker in grey, p_{future} rapidly rises to > 0.8 , indicating that a transition from one speaker to another is predicted. During the silence between turns, p_{future} remains elevated, favouring the speaker depicted in red, before the onset of their speech. The model therefore correctly predicts the transition to the speaker depicted in red in advance, whilst speech from the speaker in grey is ongoing. This example is interesting as there is a region where $p > 0.5$ “*what you going to school for what*”. This may stem from the partial utterance “*what you going to school*” being a complete question. The probability then dips back below 0.5.

In summary, the VAP model performs well on the Candor corpus of videoconferencing speech. This is despite the model not availing of the visual cues that are afforded to the interlocutors by the videoconferencing medium.

4.4.6 Discussion

In this section, we expanded the analysis of the VAP model, previously evaluated on telephone speech, to a new domain: videoconferencing. In these interactions, the interlocutors have access to both verbal and non-verbal cues, both of which support turn-taking (Kendon 1967; Levinson and Torreira 2015). Despite having no access to non-verbal communication modality, we show that the VAP model performs well. There are no reported values in the literature with which to compare performances here, but the balanced accuracy and $F1$ scores are well above baseline values. Similarly, the ROC and PR curves demonstrate that the model is well-trained and distinguishes shifts from holds and mid-turn speech. This addresses Research Question 4, as we show an audio-only model can be trained to predict turn-taking without visual cues, even when interlocutors can see one another.

Interestingly, we found the performance of the VAP model on Candor was better than on Switchboard. A possible explanation could be the different statistics in the two corpora. In Candor, there are more shifts relative to holds than in Switchboard. For instance, the proportion of events extracted that are shifts is 0.18 in Switchboard (ASR), compared with 0.28 for Candor (FTO $> +250$ ms, Table 4.5). This perhaps leading to more balanced performance on the Candor corpus. Another explanation for the discrepancy may stem from the nature of the interactions in the two corpora. As discussed in Section 2.3 of the literature review, in Switchboard, interlocutors are given prescribed topics and the interactions are approximately 5 mns in length. On the other hand, in Candor, no set topic is provided and interactions last longer at 30 mns. Furthermore, the videoconferencing medium is also known to make interaction less spontaneous and fast-paced (Sellen 1995; O’Conaill et al. 1993).

These factors likely influence the ease with which turns can be predicted in the Candor corpus. It may therefore be possible that turn-taking is simply more predictable from speech alone in the Candor videoconferencing interactions than in the Switchboard telephone interactions.

4.5 A cross-corpus analysis of audio-only turn-taking

Predictive turn-taking models (PTTMs) are developed to learn human-like turn-taking capabilities, ultimately with the view to a real-world deployment (Skantze 2021). As we have discussed, PTTM development has overwhelmingly focussed on telephone speech. An issue arises here as there is an implicit assumption that, once deployed, a telephone speech trained model will perform well on other types of interaction, such as human-robot interaction. The models must therefore be able to generalise to new corpora and domains. However, it is currently unknown whether a telephone speech trained model will perform well in other interactions; in particular, those where non-verbal cues are available to interlocutors.

Table 4.11 Cross-corpus hold/shift prediction results (test set performance, averaged across 5 folds). Models were trained on one corpus and evaluated on the other. SWB=Switchboard (ASR alignment) and CND=Candor. Percentage change relative to within-corpus performance (Tables 4.7 and 4.9) is shown in brackets.

Hold/shift prediction metrics, cross-corpus deployment					
Evaluation point	Train → Test	F ₁ (Weighted)	F ₁ (Hold)	F ₁ (Shift)	Accuracy (Balanced %)
during mutual silence (FTO > +250 ms)	CND → SWB	0.76 (↓6%)	0.88 (↓2%)	0.27 (↓40%)	57 (↓13%)
	SWB → CND	0.71 (↓15%)	0.81 (↓9%)	0.46 (↓34%)	63 (↓20%)
before end of turn (FTO > +250 ms)	CND → SWB	0.75 (↓5%)	0.88 (–)	0.23 (↓41%)	55 (↓11%)
	SWB → CND	0.71 (↓13%)	0.80 (↓8%)	0.45 (↓32%)	62 (↓19%)
before end of turn (FTO > 0 ms)	CND → SWB	0.79 (↓4%)	0.90 (↓1%)	0.23 (↓46%)	56 (↓14%)
	SWB → CND	0.77 (↓11%)	0.87 (↓5%)	0.40 (↓39%)	62 (↓20%)
before overlap (FTO < -250 ms)	CND → SWB	0.53 (↓15%)	0.45 (↓19%)	0.58 (↓12%)	53 (↓14%)
	SWB → CND	0.68 (↓13%)	0.81 (↓6%)	0.33 (↓41%)	57 (↓19%)

In this section, we therefore conduct a cross-corpus analysis, training the VAP model on Switchboard telephone speech and deploying it on Candor videoconferencing speech, and vice-versa. The purpose of this analysis is to establish whether a telephone speech trained model is valid for deployment in an interaction where interlocutors can see one another. The latter scenario more closely resembles a deployment where non-verbal communication plays a role, for example embodied interaction. We take the models trained on Switchboard (ASR alignment) in Section 4.3 and Candor in Section 4.4. We then deploy the model to the other corpus, e.g. training on Switchboard and deploying on Candor, denoted SWB→CND.

4.5.1 Performance on turn-taking metrics

Performance on the hold/shift prediction metrics is reported in Table 4.11. Performance on the shift prediction metric is shown in Table 4.12. We report the percentage change in each metric relative to the performance when trained and tested on the same corpus.

The results show a drop in the balanced accuracy of all turn-taking metrics considered. The drop is larger for Candor at an $\approx 20\%$ decrease in balanced accuracy than for Switchboard, which has an 11-13% decrease in balanced accuracy. In both corpora, holds have less of a performance decrease than shifts. With the exception of *hold/shift prediction before overlap*, all F_1 shift scores show a drop in performance, which is slightly greater in Switchboard (40-46%) than Candor (32-39%). We note a similar drop in performance on the shift prediction metric is shown in Table 4.12. Paired t-tests reveal that the drop in performance is significant (Tables B.5 and B.6 in Appendix B).

4.5.2 Performance analysis with ROC and PR curves

We also conduct an analysis with ROC and PR curves in Figure 4.7. Hold/shift and shift prediction metrics are shown, where $FTO > +250$ ms. ROC curves for the hold/shift prediction metric on Switchboard show performance is close to the diagonal baseline in Figure 4.7a in

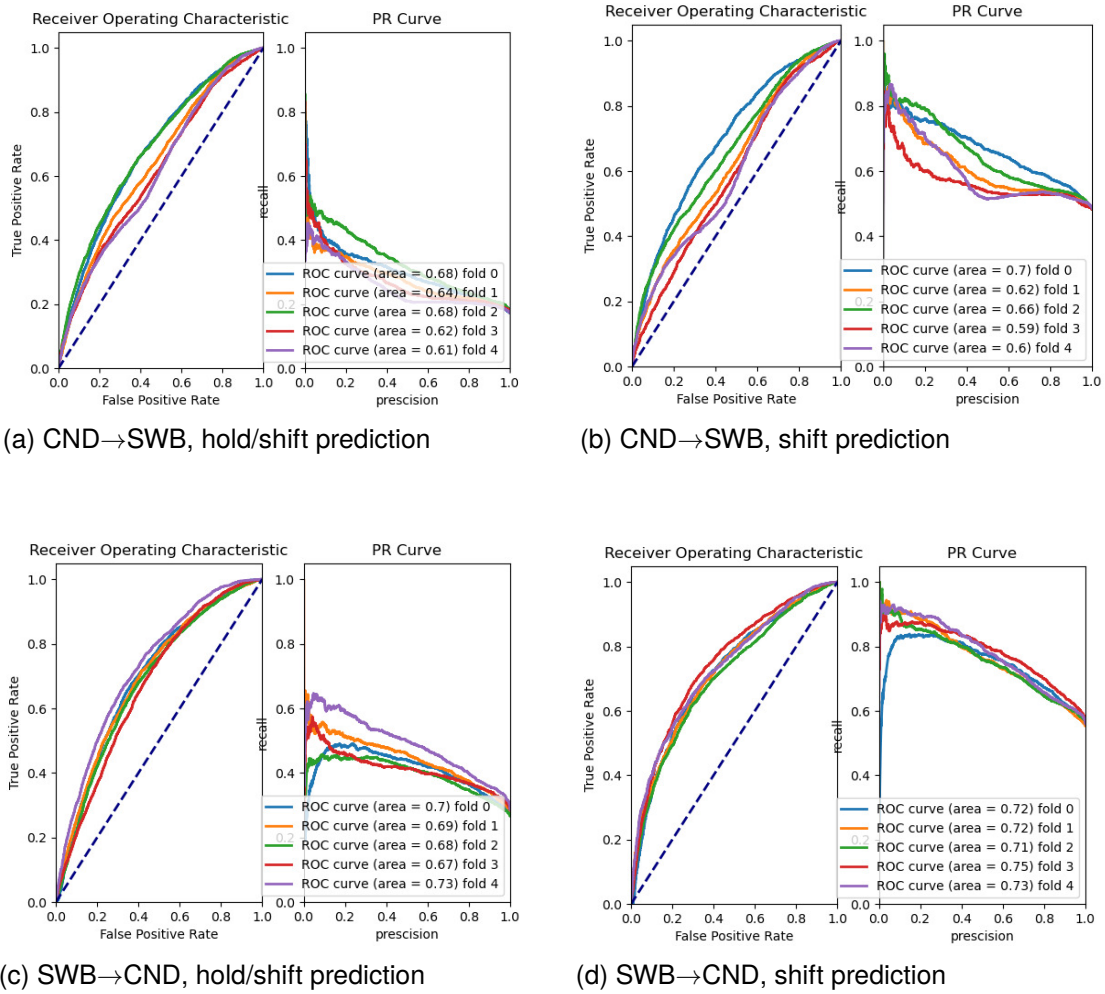
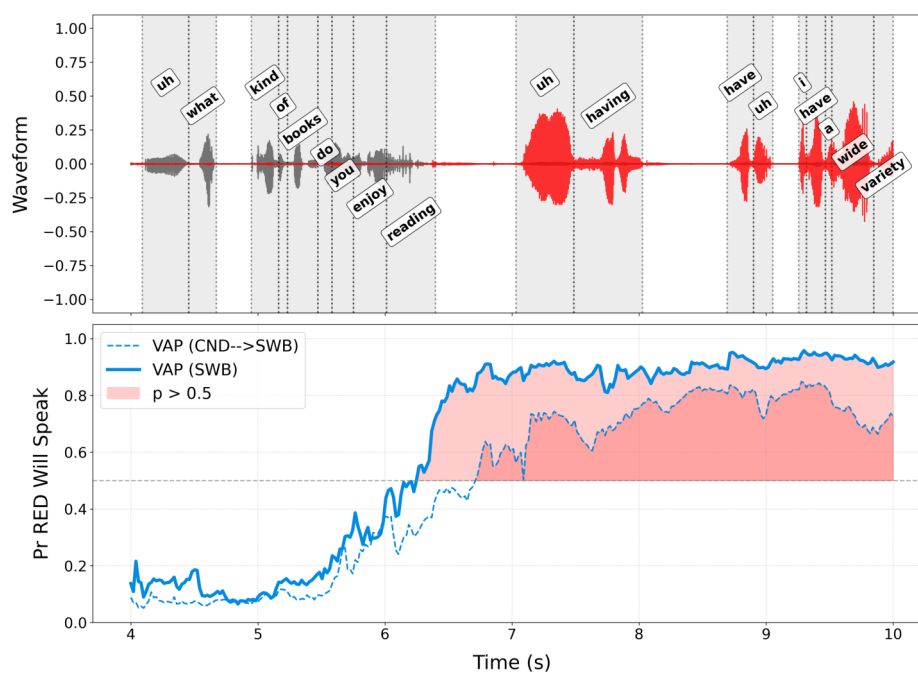
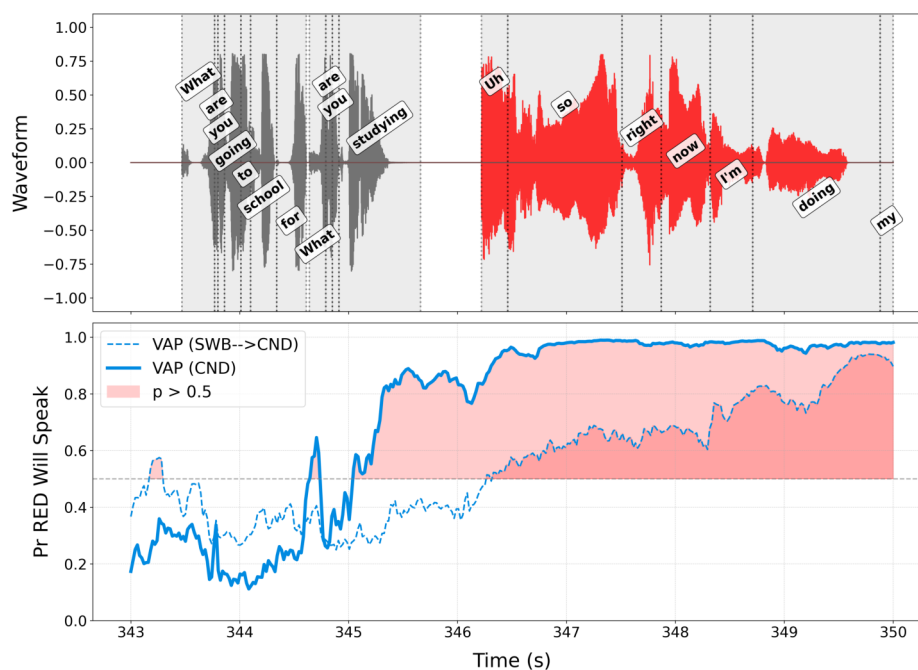


Figure 4.7 ROC curves for the voice activity projection (VAP) turn-taking model, trained on either Switchboard (SWB) and deployed Candor (CND), or vice-versa. Denoted SWB→CND and CND→SWB respectively. Performance on the hold/shift and shift metrics are illustrated (FTO > + 250 ms). The metric shown is hold/shift prediction during mutual silence and shift prediction before the end of a turn.



(a) Trained on Candor, deployed on Switchboard



(b) Trained on Switchboard, deployed on Candor

Figure 4.8 Cross-corpus analysis examples from two corpora: (a) Trained on Candor, deployed on Switchboard and (b) Trained on Switchboard, deployed on Candor.

Table 4.12 Cross-corpus generalization performance, with the relative % change from within-corpus performance (Tables 4.8 and 4.10) shown in brackets.

Shift prediction metrics, cross-corpus deployment					
Evaluation point	(trained → test)	F ₁ (Weighted)	F ₁ (Hold)	F ₁ (Shift)	Accuracy (Balanced %)
before end of turn (FTO > +250 ms)	CND → SWB	0.61 (↓17%)	0.67 (↓12%)	0.54 (↓23%)	61 (↓16%)
	SWB → CND	0.70 (↓18%)	0.68 (↓19%)	0.71 (↓18%)	70 (↓18%)
before end of turn (FTO > 0 ms)	CND → SWB	0.61 (↓18%)	0.61 (↓20%)	0.62 (↓16%)	62 (↓18%)
	SWB → CND	0.70 (↓18%)	0.69 (↓19%)	0.72 (↓18%)	71 (↓18%)

comparison with the previous result (in Figure 4.3c). A similarly low performance is observed for the shift prediction metric on Switchboard in Figure 4.7b, in comparison with the previous result (Figure 4.3d). Thus, training on Candor and deploying on Switchboard results in a sharp drop from the previous analysis in Figure 4.3, where the model was trained and tested on Switchboard.

Next, we consider the ROC curves when the model is trained on Switchboard and deployed on Candor. The hold/shift prediction (Figure 4.7a) and shift prediction metrics (Figure 4.7b) show above baseline performance. There is a sharp drop in performance compared with previous performance (Figure 4.5), but performance remains above baseline.

4.5.3 Example of the cross-corpus model predicting a speaker transition

We show an example of the cross-corpus performance of the models. The first example is from Switchboard in Figure 4.8a. We show p_{future} for the model trained on Switchboard (solid blue) and also for the model trained on Candor (dashed blue). The example is the same as previously discussed in 4.3. The Candor-trained model fails to predict the transition to the speaker in red whilst the turn of the speaker in grey is ongoing. In Figure 4.8b we see an example from Candor, the same as previously discussed in Section 4.4. We see p_{future} for both Candor (solid blue) and Switchboard (dashed blue) trained models. We see that p_{future} remains below 0.5 until the onset of speech from the speaker in red, indicating a failure to predict the upcoming speaker transition.

4.5.4 Discussion

The cross-corpus analysis has shown a sharp drop in performance when the VAP model is trained on telephone speech and deployed on videoconferencing speech, and vice-versa. As we outlined in Section 4.4.6, there are many differences between the two corpora in terms of their communication medium (telephone versus videoconferencing), whether a discussion topic is prescribed or not, and the duration of each interaction. These likely hamper cross-corpus generalisation, contributing to this performance discrepancy.

However, another contributory factor is the presence of non-verbal communication in the videoconferencing interactions. Oliveira and Freitas (2008) compared in-person and telephone speech, and found that speakers exaggerated their prosody at speaker changes in telephone speech more so than in-person. The absence of non-verbal turn-taking cues may therefore lead interlocutors to exaggerate their prosody. As prosody is linked with turn-taking (Chapter 2), the prosodic cues learned by the telephone speech model and the videoconferencing speech model may therefore be quite different, impacting generalisation.

A limitation of our method is that we assess the generalisation of the VAP model trained on the Switchboard corpus to the Candor corpus using ASR for training data generation. However, as there is no ground-truth transcription available for the Candor corpus, we must use ASR. The use of ASR means it not possible exclude the effect of differences in ASR alignment performance across the Switchboard and Candor corpora on our findings. However, we ensure that the same ASR tool, Speechmatics, was used for training both models.

This analysis highlights the wider concern of wider problem of generalisation of turn-taking models. Our results suggest a poor generalisation of models across different corpora. This is an important finding as PTTMs are trained on large corpora of human-human interaction, with the view to downstream deployment on a range of tasks (Skantze 2021) such as human-robot interaction (Skantze and Irfan 2025).

4.6 General conclusion

This chapter has explored audio-only turn-taking in telephone and videoconferencing speech corpora, with the VAP model. Section 4.2 outlined our analysis methodology, which is more comprehensive than prior work. In Section 4.3, we compared the model when trained with timing information from ASR with ground-truth transcript. We found no significant difference in performance of VAP models trained with ground-truth and ASR alignments, addressing our Research Question 2.

In Section 4.4 we trained VAP the model on the Candor videoconferencing corpus. Despite the interlocutors in the corpus being able to see as well as hear one another, we found good performance in turn-taking prediction. This analysis therefore addresses our Research Question 4. The accuracy exceeded performance of the Switchboard model, and discussed the nature of turn-taking in the two corpora, which might make turn-taking more easily predictable in Candor. Despite the good performance on Candor from the audio-only VAP model, we noted the role of visual cues in turn-taking is well-established (Chapter 2). The inclusion of non-verbal cues could therefore further improve this performance, and this forms the basis of the work in Chapter 6.

Finally, in Section 4.5, we found a poor ability of the VAP model to generalise across corpora. This is an important finding, as much of the turn-taking literature is motivated by the

need to design turn-taking models which will work in diverse real-world interactions (Skantze 2021). We discussed how non-verbal communication is present in Candor but absent in Switchboard. This might have an impact on an interlocutor's use of verbal cues in two corpora, hindering generalisation.

Chapter 5

Prosodic and lexical cues in predictive turn-taking

As we outlined in our literature review in Chapter 2, humans utilise a wide range of cues in the speech signal to achieve fluid turn-taking. In Chapter 4, we showed how the voice activity projection (VAP) model can predict upcoming holds and shifts in two-party interactions. In this chapter, we delve into uncovering which aspects of the speech signal support turn-taking prediction. Specifically, we explore the contributions of prosody and lexical cues to turn-taking prediction using the VAP model and large language models (LLMs).

We begin the chapter by conducting an analysis of audio in the Candor videoconferencing corpus. Our findings show acoustic prosodic features of speech, namely fundamental frequency, intensity, jitter and shimmer, exhibit distinct patterns during mid-turn speech and before holds and shifts; cues which could be exploited for turn-taking prediction. Next, we explore which aspects of the audio signal are exploited by the VAP model, addressing **Research Question 3**. However, the model relies on self-supervised speech representations (S3Rs) features which are known to pose interpretability challenges. This means determining which aspects of the audio signal they rely on for prediction is not straightforward (Gimeno-Gómez et al. 2025). We therefore propose a vocoder-based methodology to manipulate speech, which enables us to explore the contributions of prosodic and lexical cues to turn-taking prediction with the model.

We find that both prosodic and lexical cues support prediction. Notably, when either prosodic or lexical information is disrupted, the model exploits intact cues without further training. This indicates that S3Rs encode prosodic and lexical cues with limited interdependence, as one can be disrupted, whilst the model exploits the other.

A limitation of our vocoder methodology is that temporal information, notably voicing and syllable duration cues are preserved. As these are related to turn-taking (Duncan 1972), they could be used by the VAP model. On the other hand, LLMs-based turn-taking models are

purely lexical. We show that two LLM-based turn-taking models from the literature cannot resolve instances of *syntactic ambiguity*, giving rise to a high number of false end-of-turn predictions. We demonstrate that the VAP model is not susceptible to these errors. This demonstrates the use of multimodal cues present in speech, such as prosody, voicing, and so on to resolve the ambiguity in the lexical channel, addressing **Research Question 7a**. In our analysis of LLMs, we note that the high balanced accuracy of models reported in earlier work (Ekstedt and Skantze 2020; Leishman et al. 2024) masks the high number of false end-of-turn predictions made by these models. We propose architectural modifications and evaluate their performance.

Chapter structure

The remainder of this chapter is laid out as follows. In Section 5.1, we analyse prosody and voice quality in the Candor videoconferencing corpus. In Section 5.2, we propose a method based on the WORLD vocoder (Morise et al. 2016) to isolate prosodic and lexical cues from the speech signal. Then in Section 5.3 we assess the performance of the VAP model from Chapter 4 on the vocoded speech, addressing Research Question 3.

In Section 5.4 analyse LLM-based turn-taking, showing the models have a high number of false positive turn shifts. We show that by using audio, the VAP model resolves this ambiguity, addressing Research Question 7a. Finally, in Section 5.4.7 we propose modifications to a state-of-the-art LLM-based turn-taking model and evaluate their impact.

Research questions addressed

The work in this chapter addresses the following research questions:

- **Research Question 3** *To what extent can prosodic and lexical cues support accurate and robust turn-taking with state-of-the-art predictive turn-taking models (PTTMs) from the literature?*
- **Research Question 7a** *Can PTTM utilise multimodal cues in the same way as humans?; For instance, by using prosodic cues to resolve lexical ambiguity*

Publications

The work in Sections 5.2 and 5.3 appear in our ICASSP 2026 paper:

- O'Connor Russell, Sam, Charuau, Delphine and Naomi Harte “*The Role of Prosodic and Lexical Cues in Turn-Taking with Self-Supervised Speech Representations*” ICASSP 2026, Barcelona, Spain

5.1 Prosodic cues of holds, shifts, and mid-turn speech

Many studies demonstrate the role of prosody in turn-taking, as highlighted in Chapter 2. We therefore conduct an exploratory analysis of prosody in the Candor corpus. This motivates our work in the following sections, where we investigate the role of prosody in the VAP turn-taking model.

5.1.1 Extracting acoustic features

We conduct an analysis comparing prosody and voice quality features of speech immediately before holds or shifts with mid-turn speech. We extract the prosodic features of fundamental frequency¹ ($F0$) and intensity from the Candor corpus using the `parselmouth` Python library (Jadoul et al. 2018). For $F0$ extraction, we use a `time_step` of 10 ms, a `pitch_floor` of 75 Hz and a `pitch_ceiling` of 600 Hz. For intensity, we also use `parselmouth` and a $F0$ floor of 75 Hz. We also extract jitter and shimmer, which are voice quality features derived from $F0$ and intensity (Chapter 2).

We compute the mean value of these features within windows that are between 100 - 1000 ms long. Windows are located before the shift and hold timings from the corpus listed in Table 4.5 (minimum floor-transfer offset (FTO) of +250 ms, timing from the Speechmatics ASR alignment). For the purposes of this analysis, we refer to these timings as *turn_taking_events*, where each *turn_taking_event* can be a *hold*, a *shift* or *mid-turn speech* sampled at random from within a speaking turn. Each *turn_taking_event* is associated with a *speaker* which uniquely identifies the interlocutor, is also assigned a unique identifier which we refer to as an *event*. We extract prosody and voice quality features within the windows, following Gravano and Hirschberg (2009). The process is depicted in Figure 5.1. We compute the mean feature value within each window, ignoring windows where there less than 10 voiced frames. To ensure there are sufficient samples for statistical analysis, we skip sessions where there are less than 10 shifts.

5.1.2 Statistical analysis

To analyse the extracted prosody and voice quality features, we fit separate linear mixed-effects models (Bolker 2015). Each model predicts $F0$, intensity, jitter and shimmer. All models are defined by the R following code:

```
window_length * turn_taking_event + (1 | speaker / event)
```

¹We use the term fundamental frequency or $F0$ in this chapter as it is measurable, whereas *pitch* refers to human perception of $F0$ (Oxenham 2012)

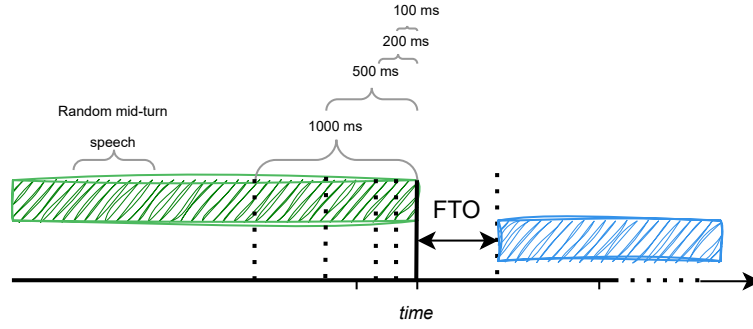


Figure 5.1 Illustration of analysis windows used to compute prosodic and voice quality features. Transition from one speaker in green to another in blue is shown. The mean value of the features is computed within windows located at the end of speaking turns, ranging in width from 100 to 1000 ms. Windows of random mid-turn speech are also considered, with their lengths matching the analysis window duration.

Modelling is conducted with the `lmer` function from the `lme4` package in R (Kuznetsova et al. 2017). The fixed effects are *window_length* (the window length in ms) and *turn_taking_event* (hold, shift or mid-turn speech). A random effect for *speaker* accounts for individual speaker differences in prosody and voice quality. A random effect uniquely identifying each *turn_taking_event*, *event*, is nested within *speaker* as there are multiple windows overlapping windows in each turn (Figure 5.1). We set the reference window to 1000 ms and the reference event type to mid-turn speech.

To visually compare differences in prosody and voice quality across speakers, we follow Gravano and Hirschberg (2009). We normalise $F0$, intensity, jitter and shimmer by their mean and standard deviation values during mid-turn speech at the speaker level. Let w denote a window of length $|w| \in \{100, 200, 300, 500, 1000\}$ ms located before a shift or hold. The mean $F0$ is:

$$F0(w) = \frac{1}{|w|} \sum_{i \in w} F0_i \quad (5.1)$$

Then, separately for each speaker s , compute the mean and standard deviation of $F0$ during mid-turn speech:

$$\begin{aligned} \mu_s &= \text{mean} \{ F0(w) : w \in \mathcal{W}_{\text{mid-turn speech}}(s) \}, \\ \sigma_s &= \text{std} \{ F0(w) : w \in \mathcal{W}_{\text{mid-turn speech}}(s) \}. \end{aligned} \quad (5.2)$$

We then z-normalise each $F0$ using the statistics of mid-turn $F0$ of the relevant speaker:

$$z(w) = \frac{F0(w) - \mu_s}{\sigma_s} \quad (5.3)$$

Table 5.1 Linear mixed-effects model for mean $F0$ in Hz. ($F0 \sim \text{window_length} * \text{turn_taking_event} + (1 \mid \text{speaker} / \text{event})$). ‘—’ = repetition of previous value. SE=standard error, df=degrees of freedom.

Window length	Event type	Estimate [Hz]	SE	df	t	p
(Intercept) ^a	mid-turn speech	180.16	0.72	3,889	248.72	< .001
100 ms	—	-1.06	0.30	951,307	-3.50	< .001
200 ms	—	-0.82	0.29	949,057	-2.82	.005
300 ms	—	-0.55	0.29	948,177	-1.90	.057
500 ms	—	-0.50	0.28	947,212	-1.74	.082
Ref. (1000 ms)	Hold	-3.21	0.24	1,018,506	-13.31	< .001
Ref. (1000 ms)	Shift	+1.94	0.27	959,572	7.24	< .001
100 ms	Hold × Window length	+17.36	0.41	979,891	42.28	< .001
200 ms	—	+11.32	0.37	972,476	31.02	< .001
300 ms	—	+7.83	0.35	965,516	22.64	< .001
500 ms	—	+2.09	0.33	950,426	6.40	< .001
100 ms	Shift × Window length	+41.91	0.57	966,741	73.41	< .001
200 ms	—	+30.09	0.47	961,783	64.20	< .001
300 ms	—	+17.69	0.41	955,296	43.13	< .001
500 ms	—	+5.66	0.38	947,835	14.97	< .001

^aBaseline: mid-turn speech. Random effects: SD(Speaker)=39.3, SD(Event:Speaker)=40.1, SD(Residual)=49.9.

repeating the process for intensity, jitter and shimmer. The normalisation removes individual speaker differences, establishing whether a feature shows a distinct pattern in each feature before a shift or hold relative to a mid-turn speech baseline. A separate normalisation is conducted for each window length.

5.1.3 Fundamental frequency $F0$ and turn-taking

Our analysis starts with an examination of $F0$ and turn-taking. The results of the linear mixed-effects model are summarised in Table 5.1. The fixed effect of *window_length* is significant for 100 and 200 ms ($p < 0.05$). However, the corresponding difference in $F0$ is relatively small (-1.06 Hz and -0.82 Hz). Thus *window_length* has little impact on the $F0$ of mid-turn speech. For the other fixed effect *event_type*, there is a significant difference in $F0$ for both hold and shift and mid-turn speech ($p < 0.001$). This shows that in the 1000 ms before a turn, there is overall little difference between mid-turn speech and holds (-3.21 Hz) and shifts (+1.94 Hz).

The interactions between *event_type* and *window_size* are highly significant. As *window_length* decreases (i.e. the end of turn approaches), $F0$ becomes increasingly elevated relative to mid-turn speech for both holds and shifts (e.g. +18 Hz → 42 Hz, shift × window). In other words, the turn boundaries are associated with a local increase in $F0$. The magnitude of the effect is greater for shifts than for holds. The effect emerges in the 500 ms before a turn

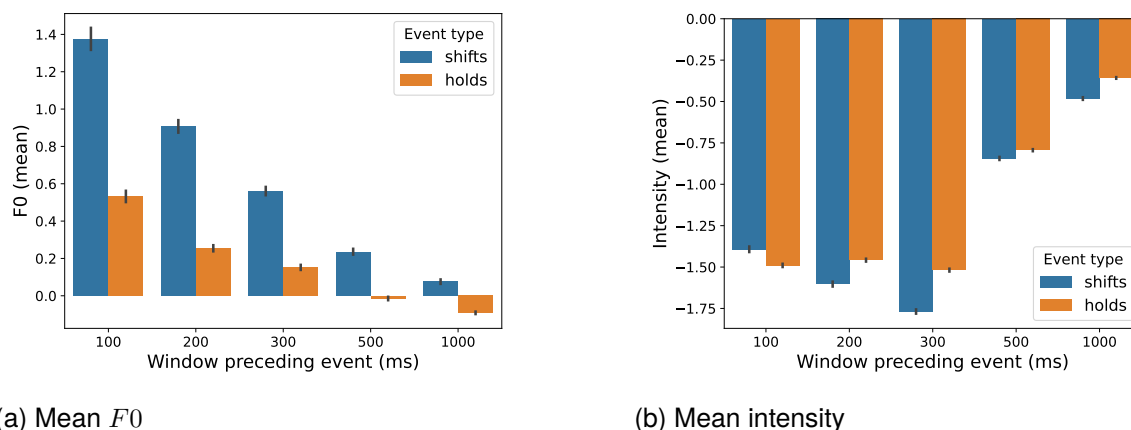


Figure 5.2 Mean F_0 and intensity (Z-normalised per speaker) for holds and shifts in the Candor corpus during windows preceding each hold and shift.

boundary (+5.66Hz shift vs +2.09 hold), but it is most prominent in the 300-100 ms at the end of a turn (+41.91Hz shift vs +17.36Hz hold). This means that shifts are marked by an elevated F_0 relative to both holds and mid-turn speech in the corpus.

The Z-normalised F_0 values shown in Figure 5.2a clearly echo the trend observed in the mixed effects model, showing a clear distinction in patterns of F_0 before holds and shifts. Mean F_0 rises before holds and shifts relative to baseline mid-turn F_0 . However, we can clearly observe that for shifts, the mean F_0 is elevated more in shifts than in holds, e.g. ≈ 1.3 standard deviations for shifts vs ≈ 0.4 standard deviations for holds.

5.1.4 Intensity and turn-taking

We continue the analysis with intensity. The linear mixed-effects model results are shown in Table 5.2. The fixed effect of *window length* is a significant ($p < 0.01$), but the difference in intensity is relatively small, ranging from -0.24 dB to -0.51 dB. The fixed effect of *event type* is significant ($p < 0.01$), but in practical terms, mean intensity is similar to mid-turn speech in the 1000 ms before a shift or hold (-0.94 dB for holds and -1.54 dB for shifts).

The interactions between *window length* and *event type* are highly significant ($p < 0.001$). As the *window_length* decreases (and the end of the turn approaches), the intensity of both holds and shifts falls relative to mid-turn speech. There is, however, little difference between the intensity of holds and shifts. For example, in the 100 ms before the end of a turn before a hold, the intensity is -10.43 dB less than mid-turn speech, in comparison with -9.04 dB less for shifts.

Z-normalised mean intensity and absolute slope are plotted in Figure 5.2 and reflect the results of the linear mixed-effects regression. The plot shows that the mean intensity gradually falls as the turn approaches its boundary, compared to intensity levels in the middle

Table 5.2 Linear mixed-effects model for mean *intensity* in decibels ($\text{Intensity} \sim \text{window_length} * \text{turn_taking_event} + (1 | \text{speaker} / \text{event})$). ‘—’ = repetition of previous value. SE=standard error, df=degrees of freedom.

Window length	Event type	Estimate [dB]	SE	df	t	p
(Intercept) ^a	Mid-turn speech	72.39	0.03	9,493	2157.15	< .001
100 ms	—	−0.48	0.03	1,176,830	−15.81	< .001
200 ms	—	−0.51	0.03	1,175,968	−16.71	< .001
300 ms	—	−0.44	0.03	1,175,494	−14.60	< .001
500 ms	—	−0.24	0.03	1,174,839	−7.98	< .001
Ref. (1000 ms)	Hold	−0.94	0.03	1,260,147	−37.36	< .001
Ref. (1000 ms)	Shift	−1.54	0.03	1,191,205	−54.23	< .001
100 ms	Hold × Window length	−10.43	0.04	1,200,644	−275.20	< .001
200 ms	—	−8.18	0.04	1,193,571	−224.94	< .001
300 ms	—	−7.09	0.04	1,185,303	−201.29	< .001
500 ms	—	−2.38	0.03	1,174,738	−69.18	< .001
100 ms	Shift × Window length	−9.04	0.05	1,192,942	−188.50	< .001
200 ms	—	−8.56	0.04	1,186,967	−194.07	< .001
300 ms	—	−7.96	0.04	1,179,102	−193.99	< .001
500 ms	—	−2.13	0.04	1,174,461	−53.30	< .001

^aBaseline: mid-turn speech. Random effects: SD(Speaker)=1.36, SD(Event:Speaker)=3.21, SD(Residual)=5.34.

of the turn. This pattern is similar for shifts and holds between 1000 and 500 ms before the end of a turn. However, in the 300 and 200 ms before a turn, the mean intensity of shifts is slightly lower than holds.

5.1.5 Voice quality and turn-taking

Finally, we consider the two voice-quality features: jitter and shimmer.

Jitter The linear mixed-effects model for jitter in Table 5.3 shows that the fixed effect of *window length* is significant. This shows jitter varies with *window length* during mid-turn speech, with less jitter in smaller window lengths (-2.80×10^{-3} vs -0.39×10^{-3} for 100 and 500 ms). The fixed effect of *event type* is significant. Comparing shifts and holds, jitter is more elevated before shifts ($+4.98 \times 10^{-3}$) than before holds ($+1.18 \times 10^{-3}$). The interaction between hold and *window length* is significant for each *window length* except 500 ms. This reveals in the 300-100 ms before a hold, jitter is steadily increasing ($+0.82 \times 10^{-3}$, $+1.50 \times 10^{-3}$ and $+2.53 \times 10^{-3}$). However, for shifts, the only significant interaction between *window length* is at 500 ms. Jitter is therefore constantly elevated ($+4.98 \times 10^{-3}$), but stable during the 1000 before a speaking turn.

We plot the speaker normalised jitter values in Figure 5.3a. The visualisation corroborates the results of the linear-mixed effects model. Jitter is elevated for shifts relative to mid-turn

Table 5.3 Linear mixed-effects model for mean *jitter* ($jitter \sim window_length * turn_taking_event + (1 | speaker / event)$). ‘—’ = repetition of previous value. ‘—’ = repetition of previous value. SE=standard error, df=degrees of freedom.

Window length	Event type	Estimate [Unitless]	SE	df	t	p
(Intercept) ^a	Mid-turn speech	1.70×10^{-2}	1.24×10^{-4}	9,574	136.75	< .001
100 ms	—	-2.80×10^{-3}	1.20×10^{-4}	949,144	-23.27	< .001
200 ms	—	-1.27×10^{-3}	1.16×10^{-4}	944,290	-10.95	< .001
300 ms	—	-0.67×10^{-3}	1.14×10^{-4}	942,402	-5.82	< .001
500 ms	—	-0.38×10^{-3}	1.13×10^{-4}	940,380	-3.39	< .001
Ref. (1000 ms)	Hold	$+1.18 \times 10^{-3}$	0.94×10^{-4}	1,049,529	12.50	< .001
Ref. (1000 ms)	Shift	$+4.98 \times 10^{-3}$	1.06×10^{-4}	962,179	47.02	< .001
100 ms	Hold x Window length	$+2.53 \times 10^{-3}$	1.62×10^{-4}	993,353	15.64	< .001
200 ms	Hold x —	$+1.50 \times 10^{-3}$	1.44×10^{-4}	977,584	10.40	< .001
300 ms	Hold x —	$+0.82 \times 10^{-3}$	1.37×10^{-4}	965,138	5.99	< .001
500 ms	Hold x —	$+0.10 \times 10^{-3}$	1.29×10^{-4}	943,344	0.81	.419
100 ms	Shift x Window length	-0.44×10^{-3}	2.25×10^{-4}	980,228	-1.96	.050
200 ms	Shift x —	-0.33×10^{-3}	1.85×10^{-4}	968,703	-1.77	.077
300 ms	Shift x —	-0.03×10^{-3}	1.62×10^{-4}	954,829	-0.19	.846
500 ms	Shift x —	$+0.82 \times 10^{-3}$	1.50×10^{-4}	940,911	5.47	< .001

^aBaseline: mid-turn speech. Random effects: SD(Speaker)=0.0051, SD(Event:Speaker)=0.0105, SD(Residual)=0.0198.

speech, with an elevation of between 0.6-0.8 standard deviations. For holds, we see a rise in the jitter from 0.2 to 0.6 in the 300 ms before the hold.

Shimmer Next, we consider shimmer. The linear mixed-effects model for shimmer is shown in Table 5.4. The fixed effect of *window length* is significant ($p < 0.001$), but there is no clear trend e.g. $+3.79 \times 10^{-3}$ at 500 ms, $+7.14 \times 10^{-3}$ at 300 ms $+1.06 \times 10^{-3}$ at 200 ms. The other main effect of *event type* is significant, showing elevated shimmer for both holds and shifts in the 1000 ms before the end of a speaking turn relative to random speech ($+6.72 \times 10^{-3}$ for hold and $+8.83 \times 10^{-3}$ for shift, $p < 0.001$). The interaction between hold and *event type* is significant, showing an elevated shimmer for each window length with no clear trend.

The speaker normalised shimmer values in Figure 5.3b reflect this trend. For holds, shimmer is constant at a slight elevation 300 ms before the end of a turn. However, for shifts, shimmer is decreased, and then rises sharply at 500 ms, and drops approaching the end of the turn. The Z-normalised shimmer values therefore show distinct temporal patterns in shimmer for shifts and holds relative to mid-turn speech.

5.1.6 Discussion

Prosody and voice quality cues have both been linked with turn-taking (Chapter 2). We explored *F0* and intensity in the Candor corpus, finding significant differences in prosody and voice quality between holds, shifts and mid-turn speech.

Table 5.4 Linear mixed-effects model for mean *shimmer* ($\text{shimmer} \sim \text{window_length} * \text{turn_taking_event} + (1 \mid \text{speaker} / \text{event})$). ‘—’ = repetition of previous value. SE=standard error, df=degrees of freedom.

Window length	Event type	Estimate [Unitless]	SE	df	t	p
(Intercept) ^a	Mid-turn speech	7.24×10^{-2}	2.29×10^{-4}	13,806	315.87	< .001
100 ms	—	$+1.86 \times 10^{-2}$	2.38×10^{-4}	1,142,886	78.10	< .001
200 ms	—	$+1.06 \times 10^{-2}$	2.35×10^{-4}	1,140,608	45.09	< .001
300 ms	—	$+7.14 \times 10^{-3}$	2.34×10^{-4}	1,139,450	30.52	< .001
500 ms	—	$+3.79 \times 10^{-3}$	2.32×10^{-4}	1,137,943	16.32	< .001
Ref. (1000 ms)	Hold	-6.72×10^{-3}	1.93×10^{-4}	1,253,893	-34.87	< .001
Ref. (1000 ms)	Shift	$+8.83 \times 10^{-4}$	2.19×10^{-4}	1,163,746	4.02	< .001
100 ms	Hold × Window length	$+2.93 \times 10^{-2}$	2.98×10^{-4}	1,177,941	98.26	< .001
200 ms	Hold × —	$+2.12 \times 10^{-2}$	2.83×10^{-4}	1,164,202	75.17	< .001
300 ms	Hold × —	$+1.96 \times 10^{-2}$	2.74×10^{-4}	1,153,107	71.40	< .001
500 ms	Hold × —	$+2.62 \times 10^{-3}$	2.66×10^{-4}	1,137,642	9.87	< .001
100 ms	Shift × Window length	$+2.42 \times 10^{-2}$	3.81×10^{-4}	1,174,786	63.48	< .001
200 ms	Shift × —	$+2.83 \times 10^{-2}$	3.47×10^{-4}	1,162,421	81.63	< .001
300 ms	Shift × —	$+3.12 \times 10^{-2}$	3.21×10^{-4}	1,147,840	97.29	< .001
500 ms	Shift × —	-9.14×10^{-4}	3.08×10^{-4}	1,137,063	-2.96	.003

^aBaseline: mid-turn speech. Random effects: SD(Speaker)=0.0085, SD(Event:Speaker)=0.0157, SD(Residual)=0.0412.

Our analysis of the Candor corpus has found a significant elevation in F_0 before the end of turns. This echoes the findings of Gravano and Hirschberg (2009) who also noted turn boundaries are marked with F_0 elevation. However, we found the elevation in F_0 was significantly more intense before shifts than holds. Schaffer (1983) similarly found F_0 elevation to be a vital indicator of turn-taking, and Gravano and Hirschberg (2009) also noted this elevation. This suggests that F_0 elevation not only marks turn boundaries, but it also plays a role in distinguishing shifts from holds.

Our analysis of intensity revealed a significant decrease in intensity approaching turn boundaries. This was also noted in prior work (Gravano and Hirschberg 2009; Duncan 1972; Hjalmarsson 2011). We found the distinction in intensity between shifts and holds was less clear than that of F_0 , although there was some evidence that intensity decreased slightly more in shifts than in holds 200-300 ms before a turn boundary. The connection between voice quality and turn-taking is less studied, but Cutler and Pearson (2018) note voice quality is likely to play a key role in turn-taking, given its close relationship to prosody. We found a significant relationship between both jitter and shimmer and turn-taking: an elevated, stable jitter in the 1000 ms was associated with yielding, whereas a rise in jitter was associated with a hold.

Behavioural studies from the literature have connected some of these statistical relationships observed here directly with turn-taking. For example, (Bögels and Torreira 2015) showed that flattening F_0 disrupts humans' ability to predict the ends of turns, suggesting the elevation in F_0 is exploited by humans as a turn-taking signal. There is however, limited

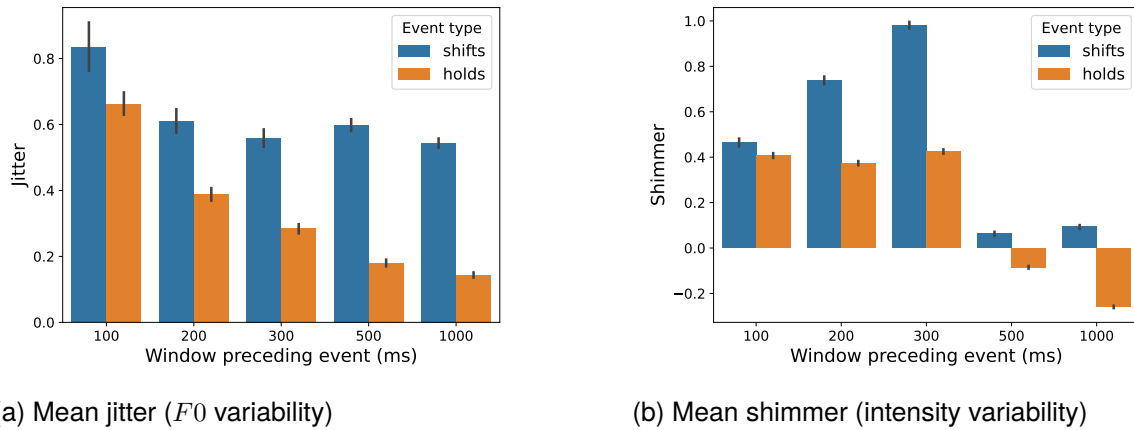


Figure 5.3 Mean jitter and shimmer (Z-normalised per speaker) for holds and shifts in the Candor corpus during windows preceding each hold and shift.

analysis of how these cues support predictive turn-taking models. The VAP model considered in Chapter 4 was previously found to be sensitive to the flattening of F_0 or intensity, suggesting these cues play a role in prediction (Ekstedt and Skantze 2022b). In Section 5.3, we explore the model sensitivity to F_0 and intensity cues explored here.

Our study is limited by the accuracy of the word alignment from the Speechmatics automatic speech recognition (ASR), the accuracy of which was considered in Chapter 3. However, we found errors in the turn-boundaries appears to largely originate from the addition of trailing silences, and should therefore not adversely impact results and end-of-turn speech is not cut. A lack of phonetic alignment means we cannot explore durational cues such as final syllable lengthening, a notable turn-taking cue (Duncan 1972). Furthermore, prosody varies across languages (Muradás-Taylor 2022) and dialect (Fletcher and Harrington 2001). It is also dependent on the interaction type, for instance, whether it is task-based or casual conversation (Lai 2014), or whether the interlocutors are familiar with one another (Wynn et al. 2022). The results in this section are therefore likely to be highly corpus-specific.

Furthermore, prosodic patterns in speech are highly variable across speakers (Cole 2015). To account for this variation, our linear mixed effects models contain a random effect for speaker, which is assigned a unique identifier. The z score normalisation process also takes into account each speaker's mean F_0 and variance in F_0 , computed at speaker. However, for example although we noted an elevated F_0 at turn boundaries which was observed at corpus level (Figure 5.2a), this effect may not be present in all speakers. An examination of individual speaker-level prosodic patterns at shifts and holds would be necessary to explore this, however this analysis is beyond the scope of this work.

5.2 Isolating prosodic and lexical cues in the speech signal

In Chapter 4, we found that the VAP model can distinguish holds and shifts in the Candor corpus with high accuracy. Having established a relationship between prosody and turn-taking in the Candor corpus, we now explore which cues in the speech signal the VAP model exploits to make turn-taking predictions. This enables a deeper understanding of how the model utilises known turn-taking cues from human behavioural studies.

5.2.1 Using a vocoder to control for prosodic and lexical cues in speech

A challenge in the analysis of the VAP model is that it is a self-supervised speech representation (S3R) front-end (Riviere et al. 2020). As we highlighted in Chapter 2, S3Rs have replaced engineered features and revolutionised speech processing; yet they pose a number of interpretability challenges. There is, therefore, a limited understanding of which aspects of the speech signal the VAP exploits in order to achieve turn-taking prediction. Prior work has shown that the accuracy of the VAP model falls when either $F0$ or intensity are flattened (Ekstedt and Skantze 2022b). This suggests these cues are exploited by the model, akin to human turn-taking. However, performance remains robust. Lexical information may therefore support prediction. Alternatively, as only one of $F0$ or intensity was flattened, the model may rely on the unmodified cue.

We therefore conduct a study to uncover the role of prosody and lexical cues in the VAP model. Given the interpretability challenges posed by the black-box nature of the model, we propose to use a vocoder to selectively control prosodic² information in the speech signal. Our analysis methodology involves the selective removal of $F0$, intensity and lexical cues in speech. We then measure the impact that this has on turn-taking prediction to establish the model's sensitivity to the removed turn-taking cue.

We use vocoders, as neuroscience research routinely uses them to cleanly isolate prosody in speech by its removing lexical cues³, or vice versa (E. C. Brown et al. 2014; Mushtaq et al. 2019). This is because vocoders decompose speech into spectral, $F0$, and aperiodicity (voicing) components, enabling the isolation of prosodic and lexical cues (E. C. Brown et al. 2014). To our knowledge, vocoders have not been applied to turn-taking, yet they have advantages over prior work involving direct manipulation of the speech signal (Schaffer 1983; J.-P. De Ruiter et al. 2006; Parsons et al. 2025; Ekstedt and Skantze 2022a). This is because methods such as low-pass filtering directly impact both prosody and intelligibility, and fail to fully isolate prosody.

²For this analysis, we define *prosody* as the $F0$ and intensity contours

³We define the lexical cues as the phonetic information in speech which supports word recognition. This is relevant to turn-taking, as word identity gives rise to higher-level grammar, syntax, and pragmatic cues which in turn support turn-taking (Bögels and Torreira 2015; J.-P. De Ruiter et al. 2006).

Table 5.5 Overview of prosody-matched noise and prosodic manipulations derived from the original speech signal (✓=preserved, x=flattened)

Type of speech	Features present in speech			Manipulation of vocoder components		
	Lexical	F0	Intensity	Spectral envelope from	F0	Spectral envelope intensity
Original speech	✓	✓	✓	unchanged	unchanged	unchanged
prosody-matched noise	x	x	✓	average spectrogram (utterance-level)	flatten to utterance mean	unchanged
	x	✓	x	average spectrogram (utterance-level)	unchanged	flatten to utterance mean
	x	✓	✓	average spectrogram (utterance-level)	unchanged	unchanged
Prosodic manipulation	✓	x	✓	original utterance	flatten to utterance mean	unchanged
	✓	✓	x	original utterance	unchanged	flatten to utterance mean
	✓	x	x	original utterance	flatten to utterance mean	flatten to utterance mean

5.2.2 Altering the speech signal with WORLD synthesis

We use the Candor corpus and isolate $F0$, intensity and lexical cues with the WORLD vocoder. The vocoder decomposes speech into spectral, $F0$ and aperiodicity components (Morise et al. 2016). From these components, the speech can be re-synthesised (synthesis). Manipulation before synthesis enables us to alter the $F0$, intensity and lexical information present in the speech signal. Table 5.5 summarises the manipulations and the resulting effects on the synthesised speech.

Running the vocoder To conduct WORLD synthesis and analysis, we use PyWorld (2025). We chunk the speech into utterances for ease of processing, using utterance boundaries from the Speechmatics transcription. We additionally compute the root-mean square (RMS) energy of the signal. We run the WORLD analysis step with 512-point FFT and a frame period 10 ms. All processing is conducted separately for each speaker (channel of stereo audio).

Intensity flattening Intensity flattening manipulates each frame in the speech signal to have the same total energy across all frequency bands in each frame of utterance, thereby removing the intensity contour from speech.

Let $s_{t,i}$ represent the spectral envelope of an utterance at frame t in the frequency bin i as extracted from WORLD synthesis. We compute the log-energy of each frame t by summing the spectral magnitudes across all frequency bins i :

$$E_t = \log \left(\sum_i s_{t,i} \right) \quad (5.4)$$

Next, we compute the mean log-energy of the entire utterance:

$$\bar{E} = \frac{1}{T} \sum_{t=1}^T E_t. \quad (5.5)$$

We then flatten the intensity ensuring that each frame in the utterance has the same overall energy by scaling the original spectral envelope $s_{t,i}$:

$$\tilde{s}_{t,i} = s_{t,i} \cdot \exp(\bar{E} - E_t) \quad (5.6)$$

As the intensity flattening procedure involves the multiplication of the spectral envelope by a scalar ($\exp(\bar{E} - E_t)$), the relative distributions of energy in each frequency bin and the temporal structure of the spectral envelope is preserved. Vowel and consonant identity are therefore maintained (Jongman et al. 2000; Stevens and Blumstein 1978; Peterson and Barney 1952). The operation, therefore, has no impact on the lexical information present in speech, which we define as the phonetic information in speech that supports word identity. This is later confirmed by analysis with ASR, where we show that the transformation has a minimal impact on the word-error rate (Table 5.6).

F0 flattening The second operation is $F0$ flattening. This involves manipulation of the $F0$ prior to WORLD synthesis. For this operation, we simply replace the extracted $F0$ signal from WORLD analysis with the mean $F0$ of the utterance. This operation removes the $F0$ contour from speech.

Removal of lexical information We remove lexical information from the speech signal by replacing the spectral envelope of each utterance with its average spectrogram. Let $\bar{s}_{j,k}$ be the average spectral envelope of the j^{th} utterance in the k^{th} frequency bin, computed as follows:

$$\bar{s}_{j,k} = \frac{1}{T} \sum_{i=0}^T s_{i,f} \quad (5.7)$$

Where T is the number of frames in the utterance. This operation therefore removes lexical information from speech by replacing the spectral envelope with an utterance level average, whilst preserving $F0$. We then scale the signal using the amplitude envelope a_t of the original speech to preserve the intensity contour:

$$s'_t = s_t * a_t \quad (5.8)$$

RMS energy matching The last operation we apply is RMS energy matching, to ensure the *total* energy in the utterance is preserved. Let an utterance from the original speech signal be denoted by u , and the same utterance transformed by WORLD synthesis be denoted $\hat{u}$. We then scale the synthesised utterance $\hat{u}$ to have the same energy as the original utterance u .

$$\hat{u}' = \hat{u} \times \frac{\text{RMS}(\hat{u})}{\text{RMS}(u)} \quad (5.9)$$

This results in a noise signal which matches the prosody of the speech signal, but contains no useful lexical information. We later confirm the removal of useful lexical information from the speech signal in the following section using ASR.

5.2.3 Selective removal of prosodic and lexical cues from speech

These manipulations enable us to generate two types of signals from the original speech: prosody-matched noise and prosodic manipulation.

Prosody-matched noise Prosody-matched noise removes intelligibility by replacing the spectral part of the signal with an average spectrogram, whilst preserving prosody. The noise preserves either the $F0$ contour, the intensity contour or both.

A block diagram of the generation process is shown in Figure 5.4 in pink. We first replace the spectral envelope with an utterance-level average, removing formant information whilst preserving $F0$ and amplitude. The $F0$, aperiodicity of the original signal and the average spectrogram are fed directly into the WORLD synthesiser. We generate three types of prosody-matched noise: the first preserves the original $F0$ and intensity; the second flattens the $F0$ to the utterance mean, whilst preserving intensity; the third noise preserves the $F0$, whilst flattening intensity. The three types of prosody-matched noise are listed in Table 5.5 (pg. 120).

Prosodic manipulation Prosodic manipulation preserves the lexical information of the original speech signal by maintaining the original spectral component. However, the prosody is altered by flattening the $F0$, intensity or both.

We preserve the spectral envelope of the original speech signal. A block diagram is shown in blue in Figure 5.4. To flatten both $F0$ and intensity, $F0$ and the spectral envelope are flattened using the previous procedure prior to synthesis. We generate two further manipulations by only flattening either $F0$ or intensity. As before, the aperiodicity is preserved. The three types of prosodic manipulation are listed in Table 5.5.

Analysis of vocoded speech To quantify the method results in the preservation or removal of lexical information in each manipulation, we use ASR, commonly used as proxy for intelligibility in the literature e.g. (Choi et al. 2023). We transcribe the test set on clean speech and speech with prosodic manipulation and prosodic noise applied with Whisper (Radford, Kim, et al. 2023), an open-source model explored in Chapter 3. We compute the word error rate (WER) of the test set relative to clean speech, which we also transcribe with Whisper.

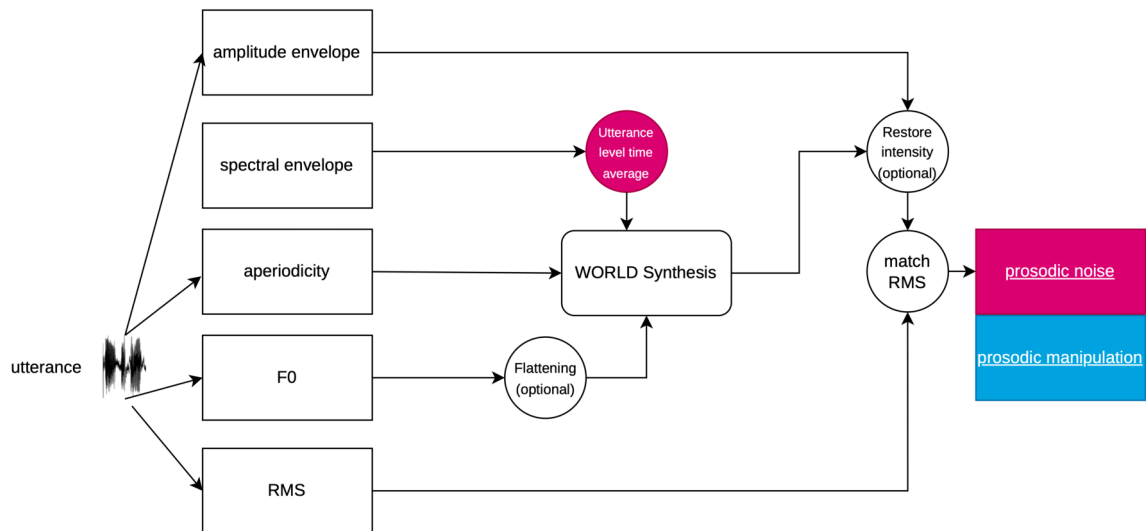


Figure 5.4 Generation of prosody-matched noise (pink) and prosodic manipulation (blue). Prosody-matched noise replaces the spectral envelope with an utterance-level average spectrogram. Prosodic manipulation is shown in blue, where the spectral envelope is preserved. Total energy in the signal is matched by RMS scaling post synthesis.

Table 5.6 Candor corpus test set word error rate (WER, %, mean per session) and error counts for prosodic manipulations and prosody-matched noise conditions (✓=preserved, x=flattened). Errors are computed relative to the original, unmodified speech signal (clean speech)

Type of speech	Lexical	F0	Intensity	WER (%)	Ins.	Del.	Subs.
<i>prosodic noise</i>	x	x	✓	98.54%	2	343,659	5,851
	x	✓	x	99.27%	0	351,752	1,186
	x	✓	✓	99.01%	2	337,323	13,826
<i>prosodic manipulation</i>	✓	x	✓	18.18%	11,411	30,225	18,990
	✓	✓	x	15.75%	11,606	23,662	17,936
	✓	x	x	24.83%	14,213	39,776	29,002

The mean WER of each test set session in Table 5.6 for each manipulation, along with the total number of errors.

The analysis reveals that prosodic manipulation of speech results in a WER which ranges from 18% to 25%. As noted in Chapter 3, a WER < 30% indicates that the transcription contains much useful information, albeit with many errors. Despite the spectral envelope being preserved, the vocoder introduces some distortion, elevating the WER. However, this range of WER indicates that there is much useful phonetic information to support word recognition in the speech signal. Furthermore, the WER is only a rough proxy for intelligibility; informal listening confirms the preservation of much lexical content. On the other hand, we see that for prosodic noise the average WER per session is $\approx 99\%$, indicating a total removal of useful lexical information from the speech signal.

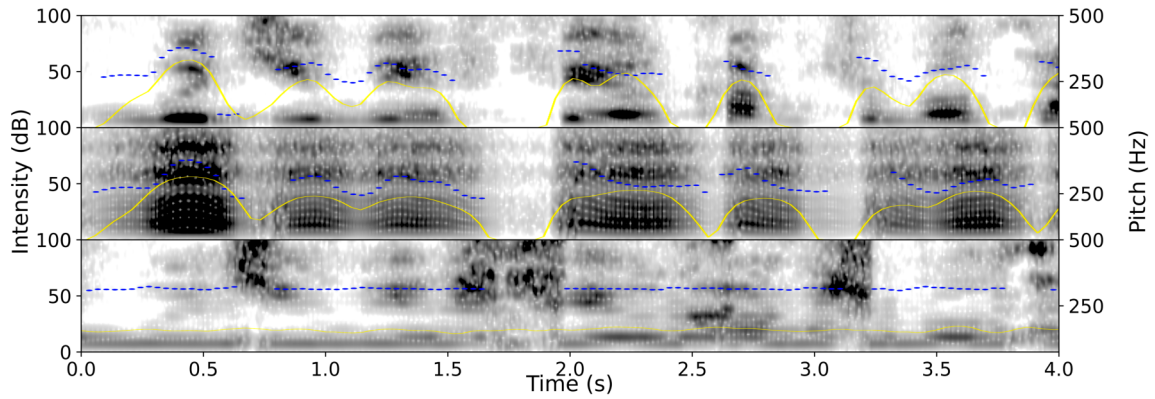


Figure 5.5 Spectrogram of utterance “would you feel comfortable” from the Candor corpus. Top=original, middle=noise+prosody preserved, bottom=intelligible+prosody flat. F_0 =blue, intensity=yellow.

We did not conduct a similar analysis of prosody. A visual inspection of the spectrograms shows that F_0 and intensity contours are preserved. The spectrogram in Figure 5.5 shows the absence (middle) and preservation (bottom) of spectral information. Although spectral information is removed, voicing is implicitly preserved by the aperiodicity component of the vocoder, left unmodified before WORLD synthesis. This means that the locations and durations of voiced and unvoiced segments are preserved.

5.3 Lexical and prosodic cues in audio-only turn-taking

Using the methodology outlined, we test the sensitivity of the VAP model to the removal of prosodic (F_0 and intensity) and lexical information from the speech signal. This enables us to explore their role in turn-taking prediction. We conduct all analyses on the Candor corpus.

5.3.1 Audio-only turn-taking when prosodic and lexical cues are removed

We first consider the VAP model trained on the original, unmodified Candor corpus, which we term “*clean speech*”. This is the model that we previously explored in Section 4.4. We manipulate the test set using our method described in Section 5.2 to selectively remove F_0 , intensity and lexical cues from the speech signal. 7 test sets are generated: 1 clean speech test set, and 6 test sets generated by manipulations of prosody and lexical cues (Table 5.5).

We report the performance averaged over 5 folds of cross-validation in Table 5.7. We note the percentage change in F_1 and balanced accuracy relative to the clean speech performance. The metrics hold/shift prediction (during mutual silence) and shift prediction (before the end of a turn) with a minimum silence of 250 ms between adjacent turns ($FTO > 250\text{ms}$).

Table 5.7 VAP model performance on clean speech and on acoustic manipulations, grouped by their preservation of lexical content and prosody (5-fold average; ✓=preserved, x=flattened. ✓✓✓= clean speech. The FTO between turns is > 250 ms. Percentage change relative to clean speech is shown in parentheses. The baseline always-hold model is included for reference.

Metric	Test set contains			F ₁ (Weighted)	F ₁ (Hold)	F ₁ (Shift)	Bal. Acc. (%)
	Lexical	F0	Intensity				
Hold/shift prediction <i>during mutual silence^a</i>	✓	✓	✓	0.83	0.88	0.70	79
	x	x	✓	0.67 (-19%)	0.81 (- 8%)	0.33 (-53%)	59 (-25%)
	x	✓	x	0.66 (-21%)	0.80 (- 9%)	0.31 (-56%)	57 (-28%)
	x	✓	✓	0.69 (-17%)	0.81 (-8%)	0.37 (-47%)	61 (-23%)
	✓	x	✓	0.76 (- 8 %)	0.84 (- 5%)	0.56 (-20%)	72 (- 9%)
	✓	✓	x	0.72 (-13%)	0.81 (- 8%)	0.46 (-34%)	66 (-16%)
	✓	x	x	0.70 (-16%)	0.80 (- 9%)	0.43 (-39%)	63 (-20%)
Shift prediction <i>before end of turn^b</i>	✓	✓	✓	0.86	0.84	0.87	85
	x	x	✓	0.58 (-33%)	0.54 (-36%)	0.62 (-29%)	58 (-32%)
	x	✓	x	0.58 (-33%)	0.52 (-38%)	0.64 (-26%)	58 (-32%)
	x	✓	✓	0.60 (-30%)	0.56 (-33%)	0.64 (-26%)	60 (-29%)
	✓	x	✓	0.76 (-12%)	0.74 (-12%)	0.80 (- 8%)	76 (-11%)
	✓	✓	x	0.71 (-17%)	0.67 (-20%)	0.74 (-15%)	71 (-16%)
	✓	x	x	0.68 (-21%)	0.63 (-25%)	0.72 (-17%)	68 (-20%)

^aBaseline F₁ (weighted)=0.45, F₁ (hold)=0.75, F₁ (shift)=0, bal. acc.=50

^bBaseline F₁ (weighted)=0.33, F₁ (hold)=0.67, F₁ (shift)=0, bal. acc.=50

Hold/shift prediction The results show hold/shift prediction falls from the baseline 79% balanced accuracy on clean speech to 63% when F0 and intensity are removed. Performance is similar to when lexical cues are removed and prosody is left intact, at 61%. This shows that the model is sensitive to the removal of both prosodic and lexical cues to a similar extent. Concerning the removal of individual prosodic cues, we note the model is more sensitive to the removal of intensity than F0. When lexical cues are removed, removing intensity gives a balanced accuracy of 57%, whereas the corresponding accuracy when F0 is removed is 66%. When lexical cues are intact, removing intensity gives 59% accuracy, whereas removing F0 gives 72% accuracy.

Overall, the model is more robust to the removal of prosody than lexical information, and the removal of intensity results in a larger drop in performance than the removal of F0.

Shift prediction For shift prediction, we notice a similar trend. However, the removal of lexical cues has a greater impact on this metric. When lexical cues are removed and prosodic cues are left intact, performance falls from the baseline 85% balanced accuracy on clean speech to 60%. However, when F0 and intensity cues are removed, and lexical information is preserved, the accuracy is 68%.

We also note that the removal of prosodic cues has a greater impact on shift prediction than the previous hold/shift prediction metric. The fall in accuracy is equivalent to a 29-32% relative decrease for shift prediction, compared with 23-28% decrease in hold/shift prediction.

This implies that prosodic cues play more of a role in the shift prediction task, which involves disambiguating mid-turn speech from speech immediately prior to a shift. Again, we find that the removal of intensity has a greater impact on performance than the removal of $F0$: when lexical cues are preserved, balanced accuracy falls to 71% when intensity is removed compared with 78% when $F0$ is removed.

5.3.2 Quantifying the contribution of lexical and prosodic cues to turn-taking prediction

In the previous analysis, we trained the model on clean speech and tested it on prosody-matched noise and prosodic manipulation. The model was not exposed to examples of manipulated speech during training. The performance may therefore be impacted in part by a lack of robustness to noise and distortions introduced by the vocoder, rather than the removal of prosodic or lexical cues.

We therefore re-train the VAP model 6 times, on each of the 6 types of noise listed in Table 5.5. By comparing performance relative to the clean speech trained and tested model, we quantify the extent to which prosodic and lexical cues in speech support turn-taking prediction. We use a training procedure identical to the one described in Section 4.1 of Chapter 4. We test the models on the same type of speech on which they were trained, isolating the contributions of prosodic and lexical cues to turn-taking prediction with the VAP model.

We report the results of the analysis in Table 5.8. The heat-map in Figure 5.6 displays the percentage of balanced accuracy on clean speech achieved by each model with prosodic and lexical cues removed.

Hold/shift prediction For hold/shift prediction, models trained on prosodic noise achieve 64-69% accuracy compared with 79% accuracy on clean speech. This compares with the 70-76% balanced accuracy achieved by models trained on speech with $F0$ and intensity flattened. The removal of lexical information therefore has a greater impact on performance than the removal of prosodic cues.

Next, we express the performance of each model as a percentage of balanced accuracy relative to the clean speech model in Figure 5.6. This reveals that a model trained on intact $F0$ and intensity cues, but with lexical cues removed, achieves 87% of the performance on clean speech. This is in comparison to when $F0$ and intensity are flattened and lexical information is preserved, we obtain 89% of accuracy on clean speech.

The analysis reveals that both prosodic and lexical cues support turn-taking, as the removal of either cue results in a decrease in performance. However, either can be used in isolation, achieving 87-89% of the accuracy of a model trained and tested on clean speech.

Shift prediction For shift prediction, models trained on prosodic noise achieve a balanced accuracy ranging from 71-78%, compared with 85% on clean speech. This is equivalent to a 9-17% relative decrease in balanced accuracy. Models trained on speech with flat prosody achieve a balanced accuracy of 70-76%, equivalent to a 4-13% decrease in balanced accuracy. In comparison with hold/shift prediction, shift prediction is less sensitive to the removal of lexical information. Expressing performance as a percentage of balanced accuracy of the clean speech model, we note 91% of the accuracy is achieved using prosody alone, compared with 87% for hold/shift prediction.

This analysis therefore also reveals that both prosodic and lexical cues support turn-taking, as the best performance is achieved when both cues are present and unaltered (clean speech). However, prosodic cues alone achieve 91% of the accuracy on clean speech, and lexical cues achieve 86%. This reinforces our prior finding of prosodic cues playing more of a role in the shift prediction task of distinguishing mid-turn speech from speech immediately prior to a shift.

F0 and intensity Next, we consider the effect of removing individual prosodic cues. When lexical information is removed, intensity is the dominant cue used by models as its removal has the greatest impact on performance (-20% and -16% decrease when intensity is removed, vs. -15 and -9% when prosody is removed). Similarly, when lexical information is preserved, intensity is the dominant cue, as the removal of $F0$ has less (-5% and -4% decrease) of an impact on performance than when intensity is removed (-11 and -6% decrease).

This shows that flattening of $F0$ consistently results in less of a decrease in balanced accuracy than the flattening of intensity. Despite being widely considered as an important turn-taking cue (Schaffer 1983; Gravano and Hirschberg 2009), the VAP model is less sensitive to the removal of $F0$ in comparison with intensity.

5.3.3 Sample prediction

We show an example of the model prediction on a shift between two speakers in the Candor corpus in Figure 5.7 from the test set. We compare the predictions of the following three models: the clean speech trained VAP model; the VAP model trained on prosodic noise with lexical information removed and prosody preserved; and the VAP model trained on speech with flat prosody. We plot p_{future} of models, along with a transcription and word boundaries. Note that we run the models using the same audio on which they were trained, but for clarity, the clean speech waveform is displayed.

The example shows that the clean speech model (dashed black) achieves the most robust prediction, with p_{future} rising above 0.5 the earliest of the three models. During the previous turn in grey, the prosodic manipulation (flat prosody) model hovers around 0.5, indicating

Table 5.8 VAP model performance, trained and tested on either clean or manipulated speech (5-fold average; ✓=preserved, x=flattened. ✓✓✓= clean speech. Percentage change relative to clean speech is shown in parentheses. The baseline always-hold model is included for reference.)

Metric	Test + train sets contain			F ₁ (Weighted)	F ₁ (Hold)	F ₁ (Shift)	Bal. Acc. (%)
	Lexical	F0	Intensity				
Hold/shift prediction <i>during mutual silence^a</i>	✓	✓	✓	0.83	0.88	0.70	79
	x	x	✓	0.72 (-13%)	0.82 (-7%)	0.43 (-39%)	68 (-14%)
	x	✓	x	0.70 (-16%)	0.81 (-8%)	0.49 (-30%)	64 (-19%)
	x	✓	✓	0.73 (-12%)	0.82 (-7%)	0.62 (-11%)	69 (-13%)
	✓	x	✓	0.82 (-1%)	0.85 (-3%)	0.55 (-21%)	76 (-4%)
	✓	✓	x	0.75 (-10%)	0.83 (-6%)	0.53 (-24%)	71 (-10%)
	✓	x	x	0.75 (-10%)	0.83 (-6%)	0.52 (-26%)	70 (-11%)
	✓	✓	✓	0.86	0.84	0.87	85
Shift prediction <i>before end of turn^b</i>	x	x	✓	0.77 (-11%)	0.74 (-12%)	0.75 (-14%)	77 (-9%)
	x	✓	x	0.72 (-16%)	0.68 (-19%)	0.81 (-7%)	71 (-17%)
	x	✓	✓	0.78 (-9%)	0.75 (-11%)	0.84 (-3%)	78 (-8%)
	✓	x	✓	0.82 (-5%)	0.80 (-5%)	0.83 (-5%)	82 (-4%)
	✓	✓	x	0.81 (-6%)	0.78 (-7%)	0.83 (-5%)	80 (-6%)
	✓	x	x	0.74 (-14%)	0.71 (-16%)	0.70 (-20%)	74 (-13%)
	✓	✓	✓	0.86	0.84	0.87	85
	✓	✓	✓	0.86	0.84	0.87	85

^aBaseline F₁ (weighted)=0.45, F₁ (hold)=0.75, F₁ (shift)=0, bal. acc.=50

^bBaseline F₁ (weighted)=0.33, F₁ (hold)=0.67, F₁ (shift)=0, bal. acc.=50

uncertainty. This suggests the model is reacting to lexical cues, which are ambiguous turn-taking cues throughout the utterance, as there are multiple points in time where it could be considered complete (“*would you feel comfortable*”, “*would you feel comfortable telling me*” and so on). It does, however, predict the transition to the other speaker ahead of time, rising sharply to >0.5 during the final word of the turn, “*living*”.

On the other hand, the prosodic noise model (no lexical content) shows no such ambiguity, being more confident of no immediate speaker transition during the grey speaker’s turn, before rising sharply to $p_{future} > 0.5$ during the final word of the turn. This suggests that it is reacting to end-of-turn prosody, which we identified as being distinct from mid-turn speech in Section 5.1.

5.3.4 Generalisation to other self-supervised speech representation (S3R)

The VAP model has a pre-trained S3R feature extractor, the contrastive predictive coding (CPC) model by Riviere et al. (2020). To explore whether our findings generalise to another S3R, we train a VAP variant with a wav2vec2.0 (Baevski et al. 2020) which replaces the CPC encoder in the VAP model.

We load the base wav2vec2.0 model from Huggingface (Huggingface 2025). We project the wav2vec2.0 features from 768 to 256 dimensions with a learned linear projection to match the dimensionality of the VAP model and CPC features. We train the model as before on clean speech and test the model on prosodic manipulation and prosodic noise. The performance averaged over 5 folds of cross-validation on the test set is shown in Table 5.9.

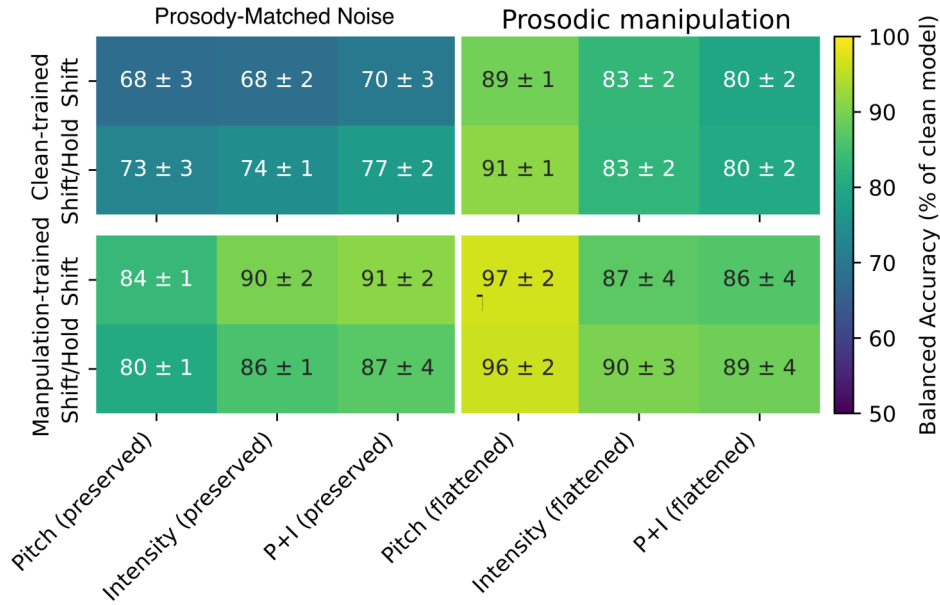


Figure 5.6 Percentage of clean speech model accuracy achieved by when trained on clean speech and tested on prosodic noise and manipulation (top), and when the model is both trained and tested on (bottom) on prosodic noise and manipulation.

Table 5.9 Wav2vec2.0 model performance trained on clean speech and tested on acoustic manipulations (5-fold average; ✓=preserved, x=flattened. ✓✓✓= clean speech. Percentage change relative to clean speech is shown in parentheses. The baseline always-hold model is included for reference.)

Metric	Test sets contain			F ₁ (Weighted)	F ₁ (Hold)	F ₁ (Shift)	Bal. Acc. (%)
	Lexical	F0	Intensity				
Hold/shift prediction <i>during mutual silence</i>	✓	✓	✓	0.94	0.95	0.89	93
	x	✓	✓	0.92 (-2%)	0.94 (-1%)	0.85 (-4%)	92 (-1%)
	✓	x	x	0.92 (-2%)	0.94 (-1%)	0.85 (-4%)	92 (-1%)
Shift prediction <i>before end of turn</i>	✓	✓	✓	0.94	0.93	0.95	94
	x	✓	✓	0.90 (-4%)	0.88 (-5%)	0.91 (-4%)	90 (-4%)
	✓	x	x	0.92 (-2%)	0.94 (-1%)	0.85 (-11%)	92 (-2%)

The results show that when prosodic and lexical cues are removed, performance remains robust. This echoes our prior findings with the CPC encoder in Section 5.3.1. Although we found performance fell to a much greater extent with the CPC model, performance remained above chance. This indicated the model could flexibly utilise the intact prosodic and lexical cues without further training, like the wav2vec2.0 model shown here.

It is important to note that, unlike CPC encoder and human turn-taking, wav2vec2.0 is not causally masked. Each representation contains information from future frames, meaning future context is implicitly used to predict upcoming turn-taking, leading to unrealistically high performance.

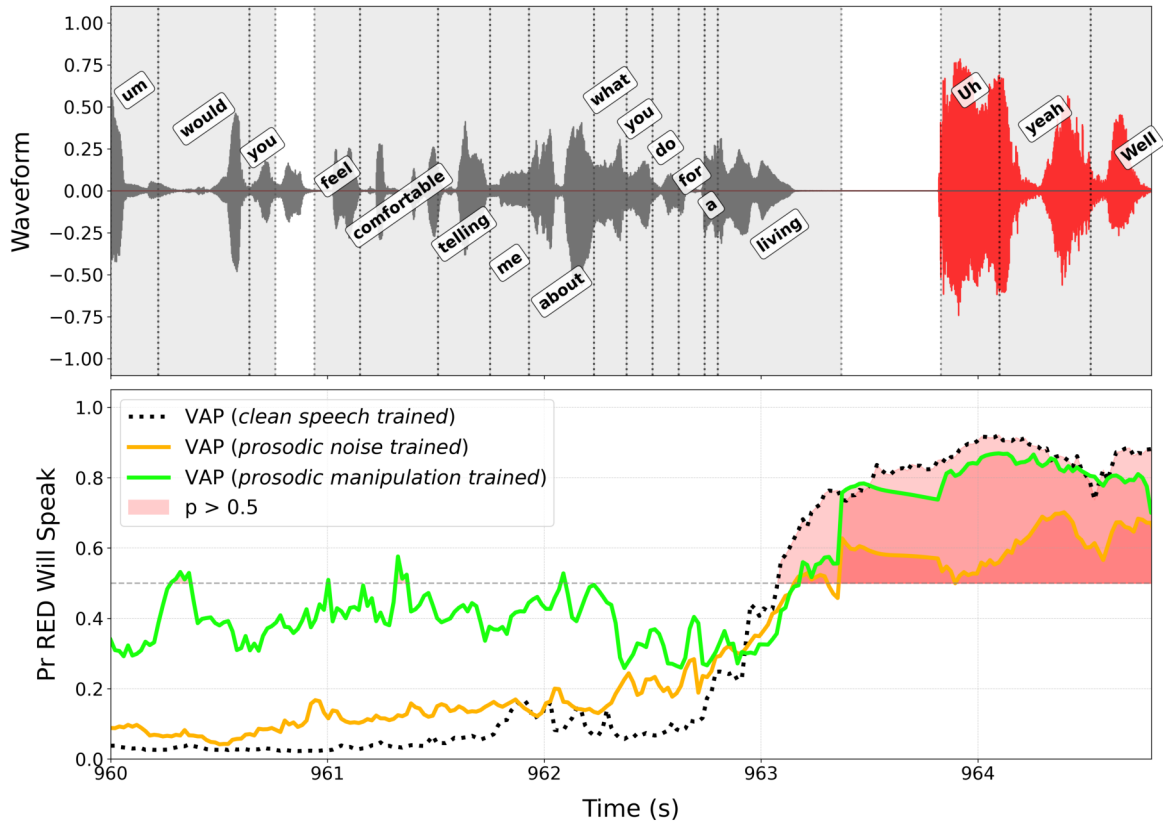


Figure 5.7 Example of model prediction during a shift from session *f7e582c-c6bc-46b6-b5a4-e5d78e8a46ac* of the Candor corpus. Clean speech, prosodic noise and prosodic manipulation trained models are shown. Note that models are tested on the same type of speech on which they were trained, but the waveform on top is of the unmodified speech signal, for clarity.

5.3.5 Discussion

In this section, we explored the role of prosodic and lexical cues in the VAP predictive turn-taking model using the Candor corpus. In Section 5.3.1, we found performance remained above chance when clean speech was swapped for noise matched to the original prosody (e.g. $\approx 61\%$ balanced accuracy, hold/shift prediction). This was remarkably similar to performance on speech with flat prosody, yet remarkably, the noise was completely unintelligible ($\approx 63\%$ balanced accuracy, hold/shift prediction). As no training was conducted here, the results show the VAP model can rely on prosody when lexical information is degraded and vice-versa.

This result raises interesting observations about the CPC front-end used in the model. As we outlined in our literature review in Chapter 2, a key concern with S3Rs is interpretability. Rich phonetic, lexical and prosodic cues are captured by S3Rs (Pasad et al. 2021; Gimeno-Gómez et al. 2025), yet they are collapsed into a single representation, making interpretability challenging (Gimeno-Gómez et al. 2025). In our analysis, we found that prosodic and lexical

cues can be removed from the VAP model, whilst achieving above-chance accuracy without any re-training. This demonstrates that the internal representation of prosodic and lexical cues are not completely interdependent in the CPC model. If the removal of lexical cues caused the internal representation of prosodic cues to be corrupted, the performance of the VAP model would be poor. However, it remained above chance, indicating the model flexibly used intact cues without further training, albeit with a considerable drop in performance. Importantly, we showed this was also true for VAP model with wav2vec2.0. This suggests that prosodic and lexical cues are not interdependently encoded in S3R.

Next, we explored the extent to which prosodic and lexical cues can support turn-taking in isolation. As we found the best performance is achieved when both cues are retained, this shows that both prosody and lexical cues support turn-taking. However, we found that using prosody alone achieved 91% of clean speech accuracy. Similarly, speech with flat prosody achieved 86% of a clean speech model. This suggests that prosody and lexical cues alone can function as turn-taking predictors, with remarkably similar performance to clean speech. In particular, the finding on prosody raises a significant advantage as useful lexical information from the signal is removed, supporting privacy e.g. sensitive topics removed. However, both the spectral envelope and pitch contours display individual speaker differences (Akagi and Ienaga 1997), meaning identifiable information is not fully removed.

A limitation of our method is the potential for residual phonetic information to persist in prosody-matched noise, though this is unlikely, as all spectral information is removed. Furthermore, the WER value confirms no useful information persists. However, the vocoder does not fully isolate the contribution of lexical cues. Notably, voicing is preserved, meaning duration cues such as the final syllable draw, strongly linked to turn-taking (Schaffer 1983; Duncan 1972) may impact the results of the VAP model with flattened prosody.

As noted in Section 5.1.6, prosodic patterns in speech are highly variable across speakers (Cole 2015). We have presented results at corpus level, averaged across speakers. An interesting direction for future work would be to explore the VAP model's performance across individual speakers. This could uncover whether the model performs better or worse on individual speakers, and how this relates to their individual prosodic patterns at holds and shifts.

5.4 Going purely lexical: turn-taking with large language models

In the previous section, we removed the prosodic contour of speech in order to isolate the contribution of lexical cues to prediction. However, speech with flattened prosody maintains temporal aspects of speech linked with turn-taking, such as syllable rate, vowel duration and so on, all of which are known turn-taking cues in human-human interaction (Duncan 1972; Gravano and Hirschberg 2009). We therefore continue by exploring LLMs-based turn-taking

models, which operate purely on text and must therefore rely solely on lexical and syntactic cues to achieve turn-taking.

We fully re-implement two LLM-based turn-taking models from the literature: TurnGPT (Ekstedt and Skantze 2020) and (Leishman et al. 2024). These models are described in Chapter 2. Due to the time at which this work was conducted, the Switchboard corpus (Godfrey et al. 1992) is used. We first describe the necessary pre-processing steps for the data, and subsequently explore model performance.

5.4.1 Pre-processing of text

We use the transcription provided with the Switchboard corpus. We normalise the text using the OpenAI normaliser described in Chapter 3, removing all punctuation and paralinguistic cues (e.g. laughter annotation) and preserving non-lexical tokens including filled pauses. There are 3,628,595 tokens in total in the corpus.

5.4.2 Serialisation of speaking turns (TurnGPT)

A simple way to conduct turn-taking with LLMs is to ignore overlapping speech (Ekstedt and Skantze 2020). The interactions are converted into a single stream of tokens (serialised). A special token, <TS>, is inserted to indicate a speaker change. TurnGPT uses this methodology (Ekstedt and Skantze 2020). The total number of tokens and special <TS> tokens in Table 5.10. An excerpt from the corpus using this method is shown in Table 5.11.

We re-implement the TurnGPT model, which is identical to the GPT-2 LLM (Radford, Wu, et al. 2019). We load the model and the pre-trained weights (Radford, Wu, et al. 2019) sourced from the Huggingface repository (Huggingface 2025). We then apply byte pair encoding to each token in the corpus (Sennrich et al. 2016). The vocabulary size of GPT-2 is 50,256 unique tokens, so we expand the vocabulary size by 1 to allow for the additional <TS> token into the model.

5.4.3 Pairwise alignment of tokens (Pairwise TurnGPT)

A disadvantage of the serialised approach is that much of the overlap and temporal dynamics of the interaction is lost. Leishman et al. (2024) proposed Pairwise TurnGPT to account for this. The model simultaneously predicts two streams of tokens, one from each speaker.

We re-implement their approach, described in detail in their paper (Leishman et al. 2024). Tokens are paired based on the level of overlap between them, using the provided word alignment. The total number of tokens is listed in Table 5.10. The large increase in the token count originates from <EMP> tokens that are inserted to represent silence. Note the discrepancy in the number of lexical tokens between approaches originates in the removal of

Table 5.10 Special token counts and descriptions for TurnGPT (Ekstedt and Skantze 2020) and Pairwise TurnGPT (Leishman et al. 2024) methods, from the ground-truth alignment of the Switchboard corpus.

Method	Token type	Description	Total number	In test set
TurnGPT	All tokens		3,785,762	187,284
	Lexical tokens		3,628,595	179,451
	<TS>	<i>Speaker transition</i>	157,167	7,833
Pairwise TurnGPT	All tokens		7,892,468	388,630
	Lexical tokens		3,748,320	185,601
	<EOT>	<i>Speaker transition with mutual silence</i>	124,463	6,143
	<YIELD>	<i>Speaker transition with overlap</i>	39,301	2,021
	<PAUSE>	<i>Mutual silence without speaker change</i>	206,963	9,670
	<EINT>	<i>Completely overlapping segment e.g. backchannel</i>	34,526	1,796
	<EMP>	<i>No speech / silence from the speaker</i>	3,738,895	183,399

completely overlapping segments in the serialised approach, i.e. backchannels (3,628,595 vs 3,748,320, an increase of 119,725).

In Table 5.12, we show the same excerpt from the Switchboard corpus encoded with the pairwise method. Note the increased richness of the approach as it allows the preservation of simultaneous speech.

We re-implement the Pairwise TurnGPT model with reference to the author’s code (Leishman et al. 2024). The model architecture, which is outlined fully in Section 2.6.7, is also based on GPT-2. We therefore load in the pre-trained weights from Hugging Face. We expand the vocabulary size by 5 to account for the extra tokens (total of 50,231 unique tokens). In the original paper, there was no token for a pause. For consistency with the VAP model evaluation, we insert an additional token, <PAUSE>, where a period of silence of at least 200 ms between speakers occurs.

5.4.4 Training and validating large language models for turn-taking

We utilise the same session-level cross-validation split previously described in Section 4.1.3. Batches consist of 256 tokens sampled at random from the training set. The batch sizes are 64 and 32 for serialised and pairwise models, to account for the increased memory usage of the latter. We train both pairwise and serialised models for a total of 10 epochs with a learning rate of 1×10^{-5} . Convergence is achieved in all cases, with the 10th epoch having the lowest validation loss.

We obtain validation set predictions for each fold. As the models output a probability distribution over all possible tokens, we extract the marginal probabilities of special tokens (<TS> in the serialised model, and <EOT>, <YIELD>, etc. in the pairwise model). We formulate a binary prediction task for each special token: a positive prediction means the relevant special token is predicted, a negative prediction means any other token in the vocabulary is predicted. We then compare with ground-truth labels (transcripts), and apply a threshold

Table 5.11 TurnGPT (Ekstedt and Skantze 2020) method of serialising speaking turns, where the <TS> token marks turn boundaries. ID is the unique identifier of each token.

#	Start (s)	End (s)	ID	Token
1	175.05	175.13	270	it
2	175.13	175.22	338	's
3	175.22	175.48	45910	torn
4	175.48	175.74	4533	ado
5	175.74	176.26	6230	season
6	176.26	176.36	50257	<TS>
7	175.78	175.95	282	al
8	175.95	176.13	1493	ready
9	176.13	176.27	282	al
10	176.27	176.41	1493	ready
11	176.41	176.54	17945	hit
12	176.54	176.86	482	ok
13	176.86	177.18	9802	lahoma
14	177.18	177.46	392	and
15	177.46	177.75	26949	probably
16	177.75	177.94	3506	right
17	177.94	178.18	2902	down
18	178.18	178.42	1456	here
19	178.42	178.75	36194	soon
20	178.75	179.02	568	so
21	179.02	179.12	50257	<TS>
22	177.67	178.04	388	um
23	178.04	178.40	17047	hum
24	178.40	178.55	5832	you
25	178.55	178.67	3826	see
26	178.67	178.87	5661	this
27	178.87	179.12	271	is
28	179.12	179.32	5661	this
29	179.32	179.47	271	is
30	179.47	179.71	45910	torn
31	179.71	179.95	4533	ado
32	179.95	180.29	23563	weather

Table 5.12 Pairwise TurnGPT (Leishman et al. 2024) method, where tokens are aligned based on their overlap. Note the insertion of <EMP> tokens to represent silence. Yellow rows indicate a transition between speakers with a period of overlap, denoted <YIELD>. The unique IDs and Tokens are shown for both speakers, S_0 and S_1 .

#	Start (s)	End (s)	ID S_0	ID S_1	Token S_0	Token S_1
1	175.05	175.13	270	50258	it	<EMP>
2	175.13	175.22	338	50258	's	<EMP>
3	175.22	175.48	45910	50258	torn	<EMP>
4	175.48	175.74	4533	50258	ado	<EMP>
5	175.74	176.26	6230	1493	season	ready
6	175.78	175.95	50260	282	<YIELD>	al
7	176.13	176.27	50258	282	<EMP>	al
8	176.27	176.41	50258	1493	<EMP>	ready
9	176.41	176.54	50258	17945	<EMP>	hit
10	176.54	176.86	50258	482	<EMP>	ok
11	176.86	177.18	50258	9802	<EMP>	lahoma
12	177.18	177.46	50258	392	<EMP>	and
13	177.46	177.75	50258	26949	<EMP>	probably
14	177.67	178.04	388	3506	um	right
15	177.94	178.18	50258	2902	<EMP>	down
16	178.04	178.42	17047	1456	hum	here
17	178.40	178.75	5832	36194	you	soon
18	178.55	178.67	3826	50258	see	<EMP>
19	178.67	179.02	5661	568	this	so
20	178.87	179.12	271	50260	is	<YIELD>
21	179.12	179.32	5661	50258	this	<EMP>
22	179.32	179.47	271	50258	is	<EMP>
23	179.47	179.71	45910	50258	torn	<EMP>
24	179.71	179.95	4533	50258	ado	<EMP>
25	179.95	180.29	23563	50258	weather	<EMP>

to the raw probability scores ($1 > threshold$, otherwise 0). We select the threshold that maximises the balanced accuracy of each special token on the validation set. We apply separate thresholds for each token and repeat the process for maximising the $F1$ score.

This is a more comprehensive evaluation than the original papers (Leishman et al. 2024; Ekstedt and Skantze 2020), where a cross-validation was not conducted. Furthermore

method allows us to tune the model in one of two ways, either by maximising the $F1$ or balanced accuracy.

5.4.5 Predictive turn-taking with large language models

We report the average performance on the test set in Table 5.13. There are two performance metrics for each token, as either the validation set balanced accuracy or the $F1$ score is optimised.

TurnGPT performance The TurnGPT model shows a high balanced accuracy score of 82% when the threshold is optimised for the balanced accuracy. This performance is in line with previous work (Ekstedt and Skantze 2020; Leishman et al. 2024). However, in prior papers, reported metrics were limited to balanced accuracy. In our table, we report a more thorough breakdown of performance reporting the number of true and false positives. The number of false negatives is low, which indicates that most end-of-turn $\langle EOT \rangle$ predictions are correctly predicted as such. This is reflected in the high sensitivity and specificity of the model. Thus, the model correctly identifies the vast majority of end-of-turns.

However, the analysis reveals that the model has a very low precision of 0.17, which originates in the high proportion of false positive end-of-turn $\langle EOT \rangle$ predictions (6,350 true positives and 30,767 false positives). Optimising the model for the $F1$ score results in a slightly more balanced performance. The precision of the model doubles to 0.31; however, this is at the expense of a large drop in the number of true positives, reflected in the fall in sensitivity to 0.46. The high balanced accuracy score reported in prior work therefore masks the high number of false end-of-turn predictions. Although the majority of lexical tokens are correctly identified (high number of true negatives), the extreme class imbalance between end-of-turn tokens (157,167) and all lexical tokens (3.6 million) masks the low precision of the model.

Pairwise TurnGPT We note a similar pattern in the performance in the Pairwise TurnGPT model. Examining the performance on the end-of-turn $\langle EOT \rangle$ token reveals a high balanced accuracy of 84%. If the $\langle EOT \rangle$ token is considered to perform the same function as the $\langle TS \rangle$ token in the TurnGPT model, this is a significant increase in performance ($p < 0.01$).

However, we again note a high number of false positive $\langle EOT \rangle$ predictions relative to the number of true positives. The precision of the Pairwise TurnGPT on the $\langle EOT \rangle$ token is worse than the TurnGPT model, at 0.09 vs 0.17, which originates in the large number of false positives (51,061). Again, optimising for the $F1$ score gives a more balanced performance. However again this is at the expense of the decrease in the number of true positives and a corresponding fall in sensitivity.

Table 5.13 Performance metrics for the special tokens across TurnGPT and Pairwise (p/w) TurnGPT models on the test set, average of 5 folds. TN=number of true negatives, FP=number of false positives. Two results are shown for each token, as the threshold is chosen to either optimise balanced accuracy/F1 (Thresh. Sel. BA, F1).

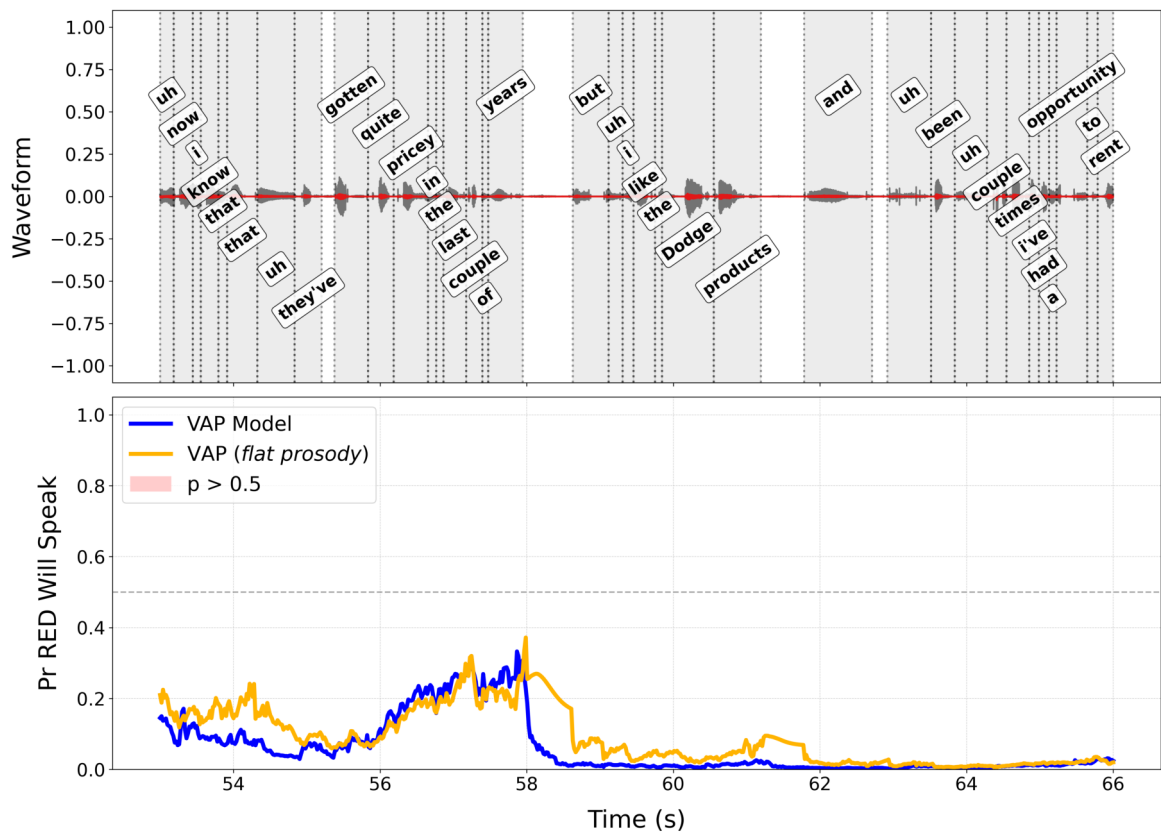
Model+Token	Thresh. Sel.	TN	FP	FN	TP	Sensitivity	Specificity	Precision	F1	Bal. Acc. (%)
<i>TurnGPT</i>										
<TS>	BA	147,028	30,767	1,502	6,350	0.81	0.83	0.17	0.28	81.8
<TS>	F1	170,394	7,401	4,213	3,639	0.46	0.96	0.33	0.38	71.1
<i>P/w TurnGPT</i>										
<EOT>	BA	329,718	51,061	1,146	4,973	0.81	0.87	0.09	0.16	83.9
<EOT>	F1	374,929	5,850	3,862	2,257	0.37	0.98	0.28	0.32	67.7
<YIELD>	BA	371,613	13,268	344	1,673	0.83	0.97	0.11	0.20	89.7
<YIELD>	F1	383,465	1,416	982	1,035	0.51	1.00	0.42	0.46	75.5
<PAUSE>	BA	268,544	108,712	1,121	8,521	0.88	0.71	0.07	0.13	79.8
<PAUSE>	F1	368,410	8,846	6,926	2,716	0.28	0.98	0.24	0.26	62.9
<EINT>	BA	376,582	8,521	250	1,545	0.86	0.98	0.15	0.26	91.9
<EINT>	F1	384,086	1,017	768	1,027	0.57	1.00	0.50	0.54	78.5
<EMP>	BA	192,047	12,060	10,112	172,679	0.94	0.94	0.93	0.94	94.3
<EMP>	F1	192,035	12,072	10,101	172,690	0.94	0.94	0.93	0.94	94.3

We notice a similar pattern of performance in the <YIELD>, <PAUSE> and <EINT> tokens where the number of false positives vastly out numbers the number of true positives. However, as was the case in the TurnGPT model, in all cases the specificity is very high. This means that the turn-taking events in the data are correctly identified in the majority of cases, but the model overall performs poorly as there are a large number of places where they are falsely inserted.

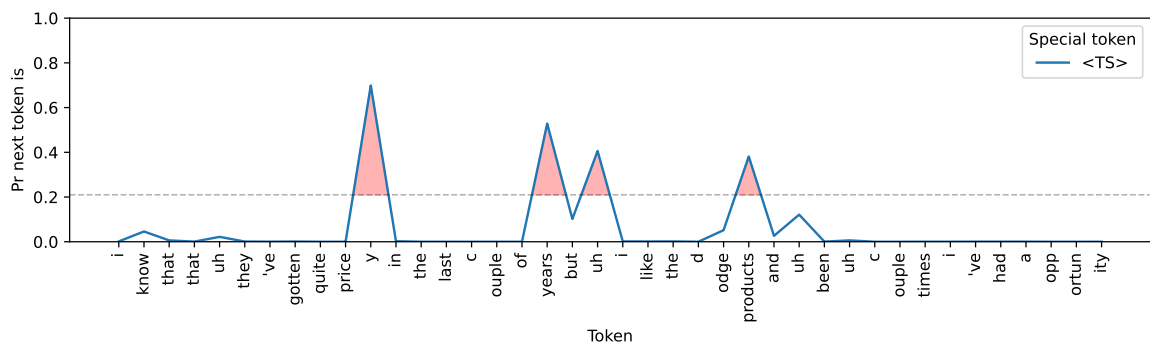
5.4.6 Comparing audio and lexical turn-taking models

The analysis shows that text alone can correctly identify the majority of turn-taking events in the dataset, but with very low precision due to the high number of false positives. The performance of the Pairwise TurnGPT model shows that by modelling both streams of text from both interlocutors, this effect has not been reduced. This suggests that text alone is a poor predictor of turn-taking in a dialogue.

We explore the performance of the TurnGPT model on two examples from the Switchboard corpus, shown in Figures 5.8 and 5.9. In the first example, the start of the utterance is “*I know that they uh they’ve gotten pricey*”. After “*pricey*” the model falsely predicts a turn shift. The utterance continues “*in the last couple of years*”, followed by another false positive turn shift prediction. The TurnGPT model predicts four turn shift tokens <TS>, each of which occur when a sentence is grammatically complete, despite <TS> never appearing in the extract. We also note this behaviour in the second example in Figure 5.9. The <TS> tokens are inserted after grammatically correct and complete sentences “*a keyboard*”, “*a keyboard and you can press*”, “*tell it what you want to do*”, and so on. Note that the actual <TS> is correctly predicted with high confidence.

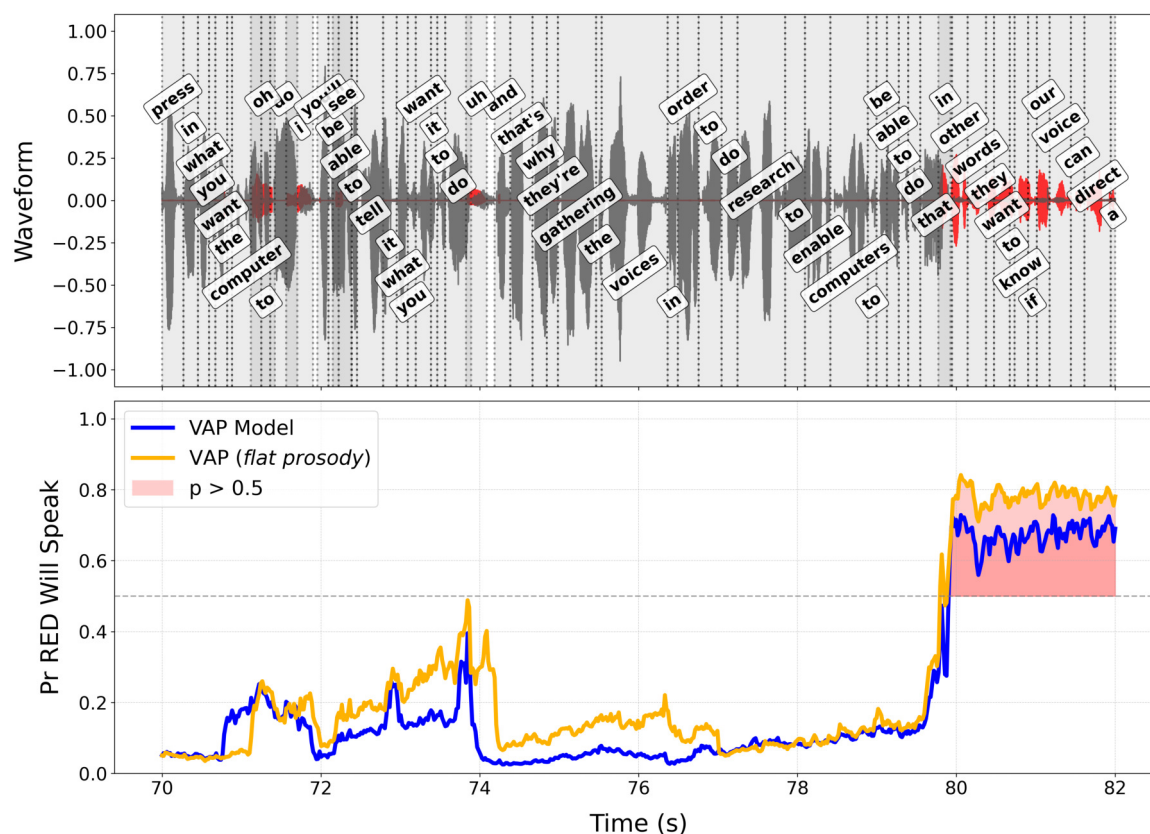


(a) VAP model predictions (audio). The probability of future speech from the interlocutor in red is shown. This interlocutor does not speak during the excerpt, reflected in the flat waveform.

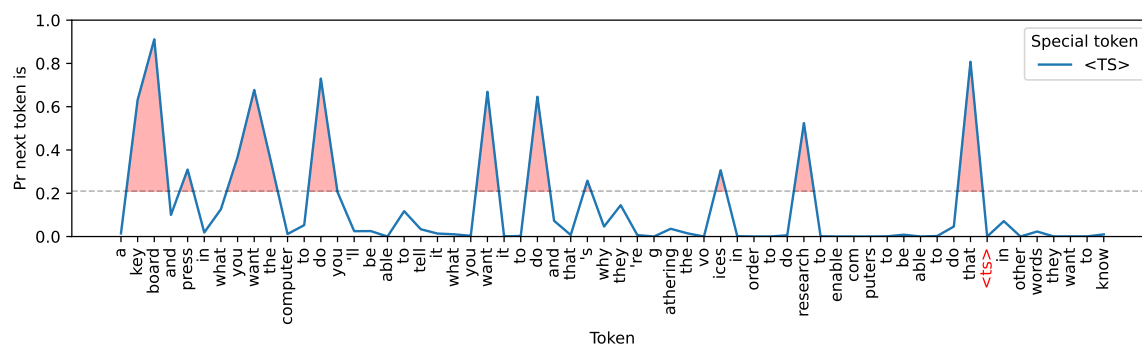


(b) Turn GPT model predictions (text). The marginal probability that the next token is a turn-shift (<TS>) is plotted against the tokens in the sequence. The threshold for a positive turn-shift (<TS>) prediction is indicated by the black line, and regions where the probability exceeds this are highlighted in red.

Figure 5.8 A comparison of VAP and TurnGPT models on the same excerpt from session SW2736 of the Switchboard corpus. Note that the alignment between the models is approximate, given the models use audio and text for prediction.



(a) VAP model predictions (audio). The probability of future speech from the interlocutor in red is shown. This interlocutor begins to speak after around the 80 second mark, visible in the red waveform.



(b) Turn GPT model predictions (text). The marginal probability that the next token is a turn-shift (<TS>) is plotted against the tokens in the sequence. The threshold for a positive turn-shift (<TS>) prediction is indicated by the black line, and regions where the probability exceeds this are highlighted in red.

Figure 5.9 A comparison of VAP and TurnGPT models on the same excerpt from session SW2238 of the Switchboard corpus. Note that the alignment between the models is approximate, given the models use audio and text for prediction.

We compare the TurnGPT model predictions with the audio VAP model from Section 4.3. The predictions in blue show that the model correctly identifies the turn shift in Figure 5.8a, and does not yield any false positive predictions of a speaker transition. The model therefore

does not suffer from the same vulnerability as the TurnGPT model, which utilises lexical cues. Next, we flatten the intensity and $F0$ contours from the speech and run the VAP model again. The predictions are shown in yellow in Figures 5.8a and 5.9a. The predictions of the model are largely not impacted, following the same pattern.

5.4.7 Proposed architectural modifications

Next, we propose two architectural modifications to the Pairwise TurnGPT model. We outline the modifications and then we consider their impact on the performance of the model.

Weight initialisation In the transformer, the tokens are converted at input by projecting the embedding to the model dimensionality (Vaswani 2017). These are learned word embeddings, which form the representation that is fed to subsequent self- and cross-attention layers of the transformer. When the special tokens are added to the model vocabulary of TurnGPT and Pairwise TurnGPT, the size of this matrix is increased to account for the additional tokens. However, as the extra elements of the matrix are randomly initialised, Pairwise TurnGPT must learn 5 word embeddings (<PAUSE>, <YIELD>, etc.) whereas TurnGPT learns just one, <TS>.

We propose a slight modification to Pairwise TurnGPT where, rather than randomly initialising embeddings when adding special tokens <PAUSE>, <YIELD> and <HOLD>, we copy the weights of the <TS> embedding from TurnGPT. Our logic is that firstly, by initialising the weights from a model pre-trained on a related but much simpler task, there is less for the model to learn during training. Secondly, we transfer knowledge by initialising the special token embeddings with those of <TS>, an embedding capturing the end of a turn. This may assist the training process. We ensure to use the same fold of cross-validation when copying the weights, ensuring separation of training and validation data. We call this modification “*weight initialisation*”.

Masked cross-attention The second modification we propose is a simple modification to the cross-attention mask. In the Pairwise TurnGPT model, cross-attention is used to combine information from both speakers (Leishman et al. 2024). At each time t the model is fed two words, one from each speaker. Given two sequences of words, one from speaker 1 $\mathcal{W}_1 = \{w_{1j}\}_{j=1}^N$ and the other from speaker 2 $\mathcal{W}_2 = \{w_{2j}\}_{j=1}^N$, the attention is masked:

$$\text{AttnMask}(w_{1i}, w_{2j}) = \begin{cases} 1, & \text{if } j \leq i \\ 0, & \text{if } j > i \end{cases} \quad (5.10)$$

The $j \leq i$ condition means the model can access past and *current* tokens when computing the attention between speakers. In Pairwise TurnGPT, if we use this setup, when $i = j$, the

Table 5.14 Token alignment illustrating a turn transition, using the Pairwise TurnGPT (Leishman et al. 2024) method. The <YIELD> token (yellow) overlaps with the onset of the response (“who do you work for”).

#	Start (s)	End (s)	Enc	Dec	Enc token	Dec token
1	197.35	198.13	43669	1640	yeah	for
2	197.95	198.22	50260	8727	<YIELD>	who
3	198.22	198.31	50258	4598	<EMP>	do
4	198.31	198.40	50258	5832	<EMP>	you
5	198.40	198.55	50258	1818	<EMP>	work
6	198.55	198.79	50258	1640	<EMP>	for
7	198.79	198.91	50258	5832	<EMP>	you
8	198.91	199.06	50258	1818	<EMP>	work
9	199.06	199.57	50258	1640	<EMP>	for
10	199.57	199.94	50258	50257	<EMP>	<EOT>
11	199.94	200.36	5354	50258	lock	<EMP>
12	200.36	200.78	23616	50258	heed	<EMP>
13	200.78	201.26	50257	50258	<EOT>	<EMP>
14	201.26	201.44	50258	1219	<EMP>	oh
15	201.44	201.62	50258	27485	<EMP>	really

mask is non-zero, meaning the model can copy the current token from the other speaker to make its prediction. Consider the example in Table 5.14. When the model is at the location in grey, it predicts the next pair of tokens in yellow. In Pairwise TurnGPT, each speaker has access to the grey token from the other speaker when predicting the next token in yellow. This is problematic as <YIELD> is used to predict ‘do’ and <EOT> is used to predict ‘lock’ and ‘oh’. Thus, the model already has access to ground-truth information that the turn has ended. Furthermore, as multiple tokens can emerge from a single word, the pairing of tokens is approximate (Leishman et al. 2024); the model could therefore attend to future information from the other speaker.

We propose to modify the attention to ensure that future tokens are predicted from past context, and information cannot be trivially copied from the other speaker. We simply make sure that the mask *AttnMask* is 1 if $j < i$, and not $i \leq j$ as before. We call this modification “*Masked Cross-Attention*”.

Analysis of the two proposed modifications We train the Pairwise TurnGPT model three further times: with the *weight initialisation* modification; with the *masked cross-attention* modification; and then with both modifications applied. Cross-validation is conducted as before. We optimise the F1 on the validation set, as we found the optimising F1 gave a more balanced performance compared with the balanced accuracy. We show the results of the models trained with the proposed modifications in Table 5.15.

We first consider the <EOT> token. For both modifications, the balanced accuracy falls relative to the original Pairwise TurnGPT model (68% vs. 66% and 65%). Inspection of the performance metrics shows this originates in an increase in false negative predictions, as less end-of-turns are predicted. However, when the two modifications are both applied to

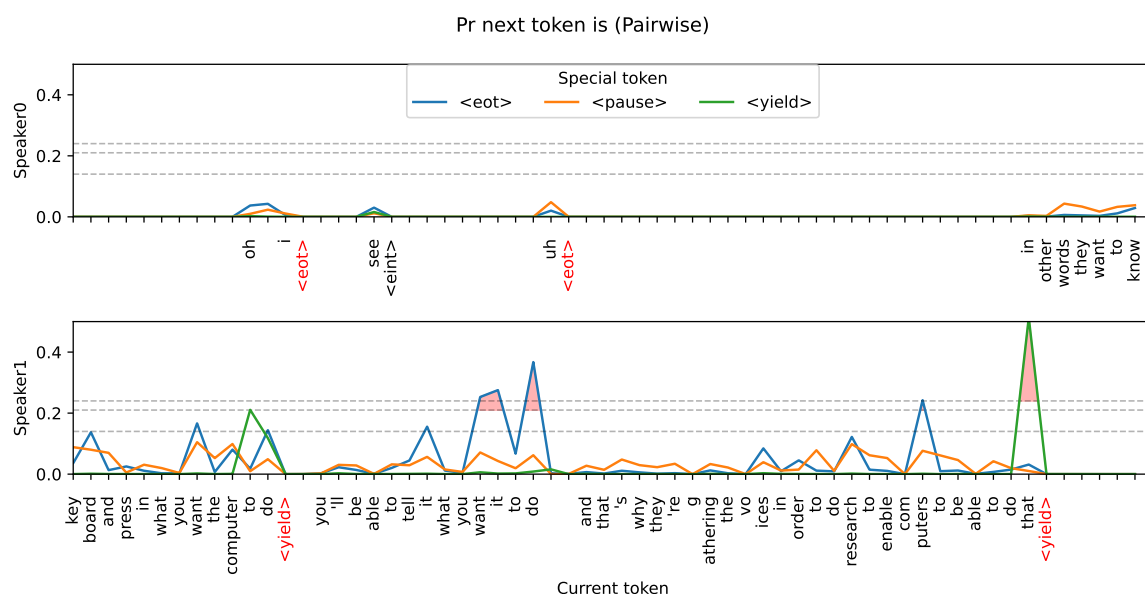
Table 5.15 LLM-based turn-taking model performance metrics across models and event types. The original Pairwise (P/w) TurnGPT model is compared against the proposed Masked Cross attention (C/a) and pre-trained weight initialisation (Weight init) methods. The average of 5 folds. TN=number of true negatives, FP=number of false positives.

Model+Token	TN	FP	FN	TP	Sensitivity	Specificity	Precision	F1	Bal. Acc. (%)
<EOT>									
<i>P/w TurnGPT</i>	374,929	5,850	3,862	2,257	0.37	0.98	0.28	0.32	68
<i>Weight init.</i>	375,451	5,328	4,103	2,016	0.33	0.99	0.28	0.30	66
<i>Mask C/a</i>	373,234	7,545	4,195	1,924	0.31	0.98	0.21	0.25	65
<i>Weight init. + Mask C/a</i>	375,611	5,168	3,509	2,610	0.43	0.99	0.34	0.38	71
<PAUSE>									
<i>P/w TurnGPT</i>	368,410	8,846	6,926	2,716	0.28	0.98	0.24	0.26	63
<i>Weight init.</i>	368,819	8,437	7,119	2,523	0.26	0.98	0.23	0.24	62
<i>Mask C/a</i>	368,186	9,070	6,885	2,757	0.29	0.98	0.23	0.26	63
<i>Weight init. + Mask C/a</i>	367,108	10,148	6,806	2,836	0.29	0.97	0.22	0.25	63
<YIELD>									
<i>P/w TurnGPT</i>	383,465	1,416	982	1,035	0.51	1.00	0.42	0.46	75
<i>Weight init.</i>	383,282	1,599	977	1,040	0.52	1.00	0.40	0.45	76
<i>Mask C/a</i>	382,631	2,250	1,442	575	0.28	0.99	0.20	0.24	64
<i>Weight init. + Mask C/a</i>	383,132	1,749	1,387	630	0.31	1.00	0.27	0.29	65

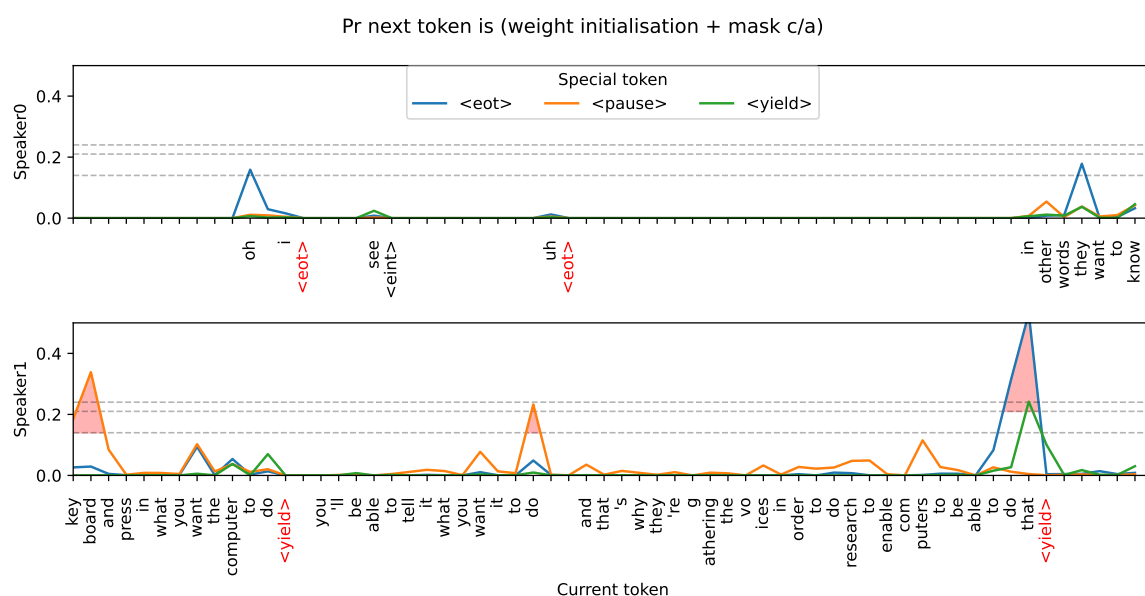
the model, we note an increase in balanced accuracy from 68% to 71%. This originates in a slight decrease in the number of false positives and increased model sensitivity (from 0.28 to 0.34). However, these differences are very small and the overall performance is relatively unchanged. Performance is either unchanged or worsened from the inclusion of the modifications for the <PAUSE> and <YIELD> tokens. Thus the weight initialisation and masked cross-attention have a very marginal improvement on the <EOT> token, but overall there is little benefit.

We show an example the original Pairwise TurnGPT and the model with both modifications in Figure 5.10. The example is the same as previously observed with TurnGPT in Figure 5.9, however, both speakers are modelled in their own token stream. We again note that the Pairwise TurnGPT model suffers from the same issue of false positive end-of-turn predictions (Figure 5.10a). These again correspond to instances where an utterance is grammatically and syntactically complete (e.g. “*you’ll be able to tell it what you want it to do*”). The modelling of the two speakers’ token streams in the Pairwise TurnGPT model has therefore not eliminated this issue.

We note the removal of false positive end-of-turn predictions when we apply the proposed modifications in Figure 5.10b (“*it to do*” and “*research in computers*”). However, on the other tokens have either been unchanged or reduced in confidence. Note that the final <YIELD> prediction is incorrect in the model with proposed modifications, whereas it is correct in the original Pairwise TurnGPT model. This echoes the results in Table 5.15, where we noted a slight increase in the sensitivity of the <EOT> predictions of the model, and unchanged or disimproved performance on the other tokens.



(a) Original Pairwise TurnGPT model



(b) Pairwise TurnGPT model with proposed weight initialisation and masked cross-attention modifications

Figure 5.10 Comparison of the original Pairwise TurnGPT and our modified model trained on the Switchboard corpus. The excerpt is from session *sw2238* from the held-out test set. Thresholds for positive prediction are optimised on the validation set, with token-specific thresholds applied. The regions where the probability exceeds the threshold are highlighted in red.

5.4.8 Discussion

In this section we explored predictive turn-taking with LLMs. We conducted a detailed analysis, and demonstrated that the previously-reported high balanced accuracy of both TurnGPT and Pairwise TurnGPT (Ekstedt and Skantze 2020; Leishman et al. 2024) mask the high number of false positive predictions in both models.

The vulnerability of lexical turn-taking models to syntactic ambiguity has been noted (Bing'er Jiang et al. 2023). In recent work, Skantze and Irfan (2025) note that this can be exploited for deployment. Audio and lexical turn-taking models can be used in conjunction with one another as, in effect, they serve two different goals. The lexical models can be used to identify *transition relevant places*, or points in time where a transition between speakers may occur. The audio models may be used to determine where in the interaction the transition actually does take place. Our qualitative examples show this pattern. Lexical content can predict possible points in time where a turn may end, but it cannot be used to reliably predict the end of speaking turns in a conversational interaction. The ability of the VAP to predict the transition correctly suggests that the local prosodic context, i.e. the $F0$ and intensity contours, along with durational cues such as syllable length, are likely used for disambiguation. This has also been noted in studies of human turn-taking, where end-of-turns are ambiguous when the prosodic contour is altered (Bögels and Torreira 2015).

However, interestingly, we found when we flattened $F0$ and intensity contour, the VAP model did not succumb to the same syntactic ambiguity as the lexical models. If the model was very sensitive to lexical cues, we would expect the predictions to be vulnerable to the same syntactic ambiguity as the TurnGPT language model, but it is not. This may suggest that the lexical information captured by CPC encoder might not be sufficiently rich to capture these syntactic cues. It has been noted that lexical information in S3R is often limited to basic word recognition (Pasad et al. 2021). Alternatively, cues such as syllable and vowel duration, turn duration, and other cues unaffected by the flattening of $F0$ and intensity may be used by the model to overcome the lexical ambiguity. The literature on human turn-taking also raises this question. J.-P. De Ruiter et al. (2006) found that humans can still predict the ends of turns when the prosodic contour is flattened. However, Bögels and Torreira (2015) found the opposite result, demonstrating that prosody is necessary when using utterances elicited from read speech designed to have syntactic ambiguity.

Our analysis therefore demonstrates that speech and audio contain many rich acoustic, prosodic and lexical cues that are leveraged for turn-taking prediction. Disentangling their respective contributions to the turn-taking problem is challenging and requires careful analysis.

5.5 General conclusion

This chapter has explored the contributions of prosodic and lexical cues to turn-taking prediction. In Section 5.1 we conducted an analysis of the Candor corpus, finding statistically significant patterns in $F0$, intensity, jitter and shimmer in the speech prior to holds and shifts.

We then investigated the role these cues play in turn-taking with the VAP model, where analysis is complicated by its S3R encoder. In Section 5.3, we proposed a novel framework based on the WORLD vocoder to isolate prosodic and lexical cues in speech, enabling an analysis of the VAP model's sensitivity to prosodic and lexical cues. We demonstrated that both prosodic and lexical cues contribute to turn-taking prediction, addressing Research Question 3. However, we found performance on noise was remarkably similar to performance on speech with flattened prosody, raising the potential for future models with privacy benefits via the removal of intelligible speech.

In the last part of the chapter in Section 5.4, we explored LLM-based turn-taking models. We found the high reported balanced accuracy of these models in the literature masks their high number of false positive end-of-turn predictions. In a qualitative analysis, we showed that this stems from syntactic ambiguity. We compared the VAP model with LLM-based models, and found it does not suffer from syntactic ambiguity to the same extent as the LLM models. We discussed explanations for this, speculating that as speech is a very rich signal with multiple cues, addressing Research Question 7a. We found our proposed modifications to the LLM-based turn-taking models did show some limited ability to reduce false positive end-of-turn predictions; however, performance on other turn-taking events was either unchanged or slightly worsened.

Chapter 6

Incorporating visual cues in a predictive turn-taking model

Until now, we have explored turn-taking prediction with speech. In Chapter 4, we found that an audio-only turn-taking model was able to predict upcoming transitions between speakers in videoconferencing interactions with high accuracy. We expanded on this analysis and also demonstrated how acoustic, prosodic and lexical cues support prediction.

We now direct our attention to visual cues and the role they play in turn-taking. Although we found good turn-taking prediction on the Candor videoconferencing corpus using audio alone, non-verbal communication is available to interlocutors and could hence be exploited for prediction. This is because the visual modality of communication contains rich non-verbal cues: gaze and head pose are highly linked with turn-taking, as highlighted in our literature review in Chapter 2. Yet despite the wealth of linguistics literature linking such visual cues with turn-taking, prior work in predictive turn-taking models (PTTMs) focusses on speech alone. It therefore remains unclear whether the inclusion of visual cues in PTTMs can improve performance.

In this chapter, we close this gap by investigating whether incorporating visual cues into a PTTM improves the performance of turn-taking prediction. We begin the chapter with an exploration of gaze, head pose and facial expression and their connection to turn-taking in the Candor corpus. We find the listener exhibits increased lip, brow and cheek movement before a turn; a visual cue that might be useful in a PTTM.

We therefore propose multimodal voice activity projection (MM-VAP), a new transformer-based PTTM which integrates visual and acoustic information. The model utilises interlocutor gaze, head pose, facial expression and landmark speakers from two interlocutors in a dyad, along with audio. We propose early and late modality fusion versions of the model, distinguished by their integration of verbal and non-verbal cues. We demonstrate that MM-VAP achieves significant improvements in turn-taking prediction above the state-of-the-art

voice activity projection (VAP) model in a wide range of speaker transitions. We provide a number of examples of model performance, along with a qualitative analysis. We also inspect the transformer attention maps of both VAP and MM-VAP, highlighting how MM-VAP captures a much richer context across speaking turns than the audio-only VAP model.

To our knowledge, our study represents the first comprehensive study of visual cues in multimodal PTTMs, demonstrating that including visual cues improves performance across all types of speaker transition considered. This addresses **Research Question 5** of this thesis, and our work is one of the first to integrate visual cues into a PTTM. The analysis of MM-VAP in this chapter lays the groundwork for Chapter 7, where we explore the practicalities of a real-world PTTM deployment and demonstrate that the visual modality endows the MM-VAP model with the ability to overcome interfering acoustic noise.

Chapter structure

The remainder of this chapter is structured as follows. Firstly, to motivate our work, Section 6.1 presents an analysis of visual cues utilised in the MM-VAP model. Section 6.2 outlines the two proposed versions of the MM-VAP model architecture: *early* and *late* fusion, referring to the integration of visual and audio features. In Section 6.3 we present a thorough evaluation of the MM-VAP model, finding a statistically significant increase in turn-taking prediction performance compared with the audio-only VAP model. An ablation study reveals that facial expression is the largest contributor to model accuracy, followed by head pose.

In Section 6.4 we explore the performance of MM-VAP on examples from the Candor corpus, comparing predictions to the audio-only VAP model. Finally, Section 6.5 discusses our work on the visualisation of attention maps, which demonstrates that visual information from the preceding turn is used in the model prediction.

Research questions addressed

The work in this chapter addresses the following research question:

- **Research Question 5** *Can the incorporation of visual cues improve the performance of PTTM?*

Publications

The work in this chapter has led to the following publications:

- ACL Findings *Visual Cues Enhance Predictive Turn-taking for Two-party Human interaction*, [O'Connor Russell](#) and Harte (2025)

- MMSYM (Multimodal symposium on human interaction) *Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction* O'Connor Russell and Harte (2024)

6.1 Visual features during speaker transitions

We begin our analysis by exploring whether certain patterns of visual activity precede shifts or holds in the Candor corpus. If there are distinct patterns, this would suggest the presence of useful information that could be exploited for turn-taking prediction.

6.1.1 OpenFace feature extraction

We extract visual features from the Candor corpus with OpenFace (Baltrušaitis, Robinson, et al. 2016), which runs without issue on the vast majority of sessions. However, in 243 of the 1,658 sessions, the tracking algorithm fails in the middle of the session. We investigate and find failure is due to various issues e.g. in one session, a participant gets up and leaves the frame to fetch something.

In the overwhelming majority of the corpus, there are no tracking issues (1,414 sessions, equivalent to 710 hours). We therefore remove these 243 sessions and conduct all our training and model analysis on the remaining sessions. We extract facial action units, eye gaze, head pose and facial landmark features and analyse them in the following sections.

6.1.2 Facial action units

We first consider facial action units (FAUs). Each FAU encodes a specific facial movement, with the full list provided in Chapter 2. OpenFace predicts FAU *presence* (binary) and FAU *intensity* (a scalar on 0-5), which are independently predicted by two internal models. We only use the FAU intensity predictions.

We plot FAU intensity for selected FAUs for a single speaker in Figure 6.1, excluding the frames where the FAU is not observed ($= 0$) for clarity. These histograms are representative of the distribution shape of the remaining action units, which also follow long-tailed and non-parametric distributions.

6.1.3 Facial action activity during speaker transitions

Next, we consider the relationship between facial action units and turn-taking, compare the median intensity of each FAU from the speaker and the listener in 500 ms windows before shifts and holds. The full number of shifts and holds extracted is listed in Table 4.5 in Chapter 4.

We first consider 500 ms windows before shifts or holds with > 250 ms between speaking turns (gap condition). We then remove the requirement of a minimum period of silence between turns, i.e. with > 0 ms between speaking turns (no gap condition). We also consider 500 ms windows located before holds and shifts that occur with periods of overlapping speech (overlap condition).

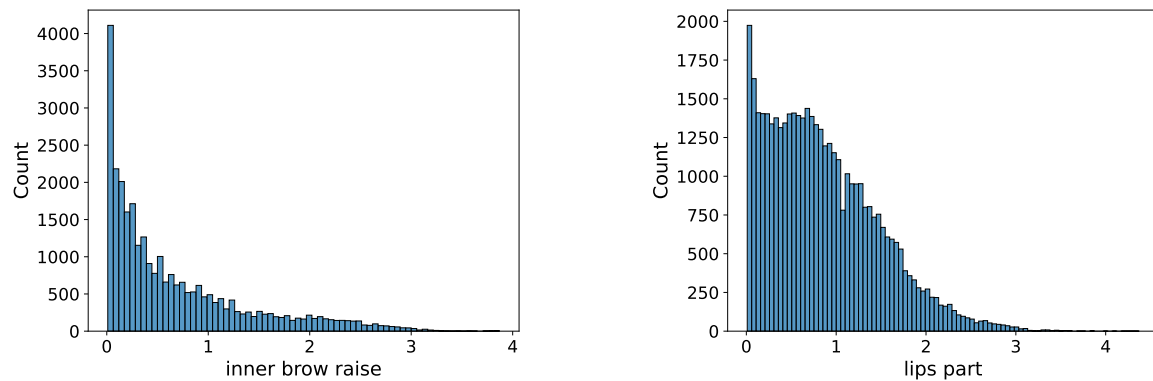


Figure 6.1 Histogram facial action unit intensities. Selected FAUs are shown from session *0a0cf5b9-84f6-4d8d-8001-ec7fd4b7437a* of the Candor corpus. Note that the values are unitless, and describe intensity. Inner brow raise = AU01, lips part = AU25, Table 2.3, pg. 37

Next, we compare FAU median intensity values during turn-taking with those located far away from speaker transitions. To achieve this, we sample an equal number of windows from a currently active speaker during mid-turn speech. Similarly, we sample windows from listeners during periods of silence. We ensure mid-turn speech and silence windows are at least 1500 ms away from the end, and 1000 ms away from the start, of a turn. This enables us to establish a baseline mid-turn value for each FAU when an interlocutor is speaking (random-speech) and listening (random-silence).

We compare the median value of each FAUs using the Mann-Whitney U test (Mann and Whitney 1947).

6.1.4 Analysis of facial action units during speaker transitions

Speaking versus listening We begin by exploring the intensity of FAUs during random periods of speaking and listening. This establishes a baseline level of FAU intensity away from speaker transitions. Figure 6.2 shows a heat map of the median value of each FAU. Random periods of speaking and silence are shown in the centre two columns. Statistical significance is for Mann-Whitney U tests comparing the median of each FAU during random speech and random silence.

The plot shows that, logically, the median values of *lip part* and *jaw drop* are higher when an interlocutor is speaking in comparison to when they are listening. We similarly observe significantly higher eyebrow, cheek and lip activity when an interlocutor is speaking in comparison with when they are listening. The overall pattern shows that the movement of the mouth and jaw when speaking is captured by the FAUs, demonstrating how facial movement is captured by the FAU intensity values.

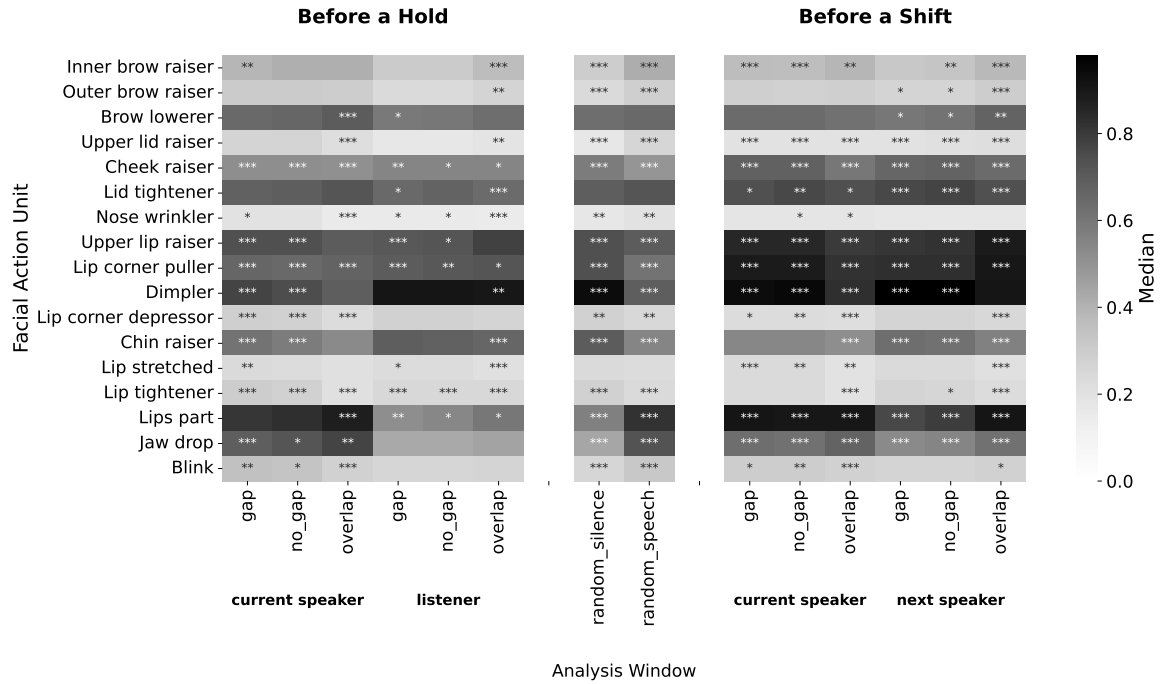


Figure 6.2 Median facial action unit intensity before key turn-taking events in the Candor corpus. Window width is 500 ms. The p-values compare random speech with random silence, random speech the current speaker, and random silence with the next speaker (shift) and listener (hold). $p < 0.001 = ***$, $p < 0.01 = **$, $p < 0.05 = *$

Before a shift We continue by exploring FAUs activity before a transition from one speaker to another (a shift). Median FAUs are shown in Figure 6.2 on the right-hand side. Statistical significance is for Mann-Whitney U tests comparing the current speaker with random speech and the next speaker with random silence.

The figure shows that the upcoming speaker exhibits significantly more FAU activity in the mouth region before a shift in comparison with a random period of silence. The action units *lips part*, *jaw drop*, *upper lip raiser*, and *lip corner puller* are all significantly more intense. We also find significant differences in *inner brow raiser*, *outer brow raiser* and *brow lowerer*.

Next, we compare the FAU activity of the current speaker before a shift. When compared with random speech in the middle of a turn, the current speaker has significantly increased *upper lip raiser*, *lip corner puller*, *cheek raiser* and *lip part* intensity FAU before a shift. There is also significantly more intense *inner brow raiser* and *lip stretch* FAU intensity from the current speaker before a shift in comparison with random speech and before a hold.

Before a hold We continue by exploring FAUs activity before a hold, i.e. when a turn ends with no change in active speaker. Median FAUs are on the left-hand side of Figure

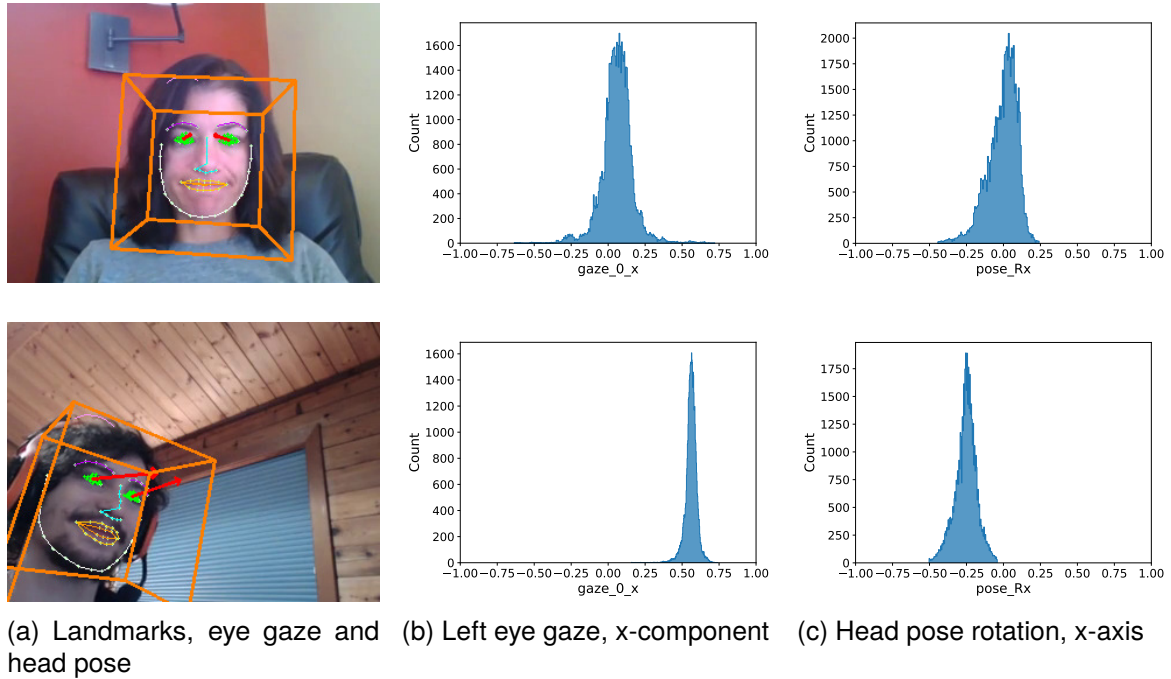


Figure 6.3 OpenFace gaze, head pose and facial landmark tracking for session *a29635a0-d33c-4ddd-ba70-154170b85057* of the Candor corpus (left) with gaze angle (centre) and head pose rotation (right) about the x-axis (horizontal direction).

6.2. Statistical significance is for Mann-Whitney U tests comparing the current speaker with random speech, and the next speaker with random silence.

When compared with random silence, the listener, who is not about to speak, shows no significant difference in brow activity, jaw drop and chin raiser. This means there is less intense activity in the mouth region here, akin to random silence. There are, however, differences in some FAUs e.g. *lid tightener*, *lip raiser* and *cheek raiser*. The current speaker shows increased blink, chin raise, lip movement and lip action unit activity compared with random speech. There is no significant difference in the *lips part* FAU for *gap* and *no-gap* windows. For overlapping speech, there is increased activity of the *lips part* FAU.

6.1.5 Facial landmarks, eye gaze and head pose

The remaining features which we extract from OpenFace are facial landmarks, eye gaze and head pose. We manually check tracking performance on a subset and observe good tracking performance. We show illustrations of these features along with histograms of gaze and head pose in Figures 6.3 and 6.4. The illustrations in Figures 6.3a and 6.4a depict eye gaze vectors in red. Head pose is depicted as a box centred at the head and rotated to the head pose angle.

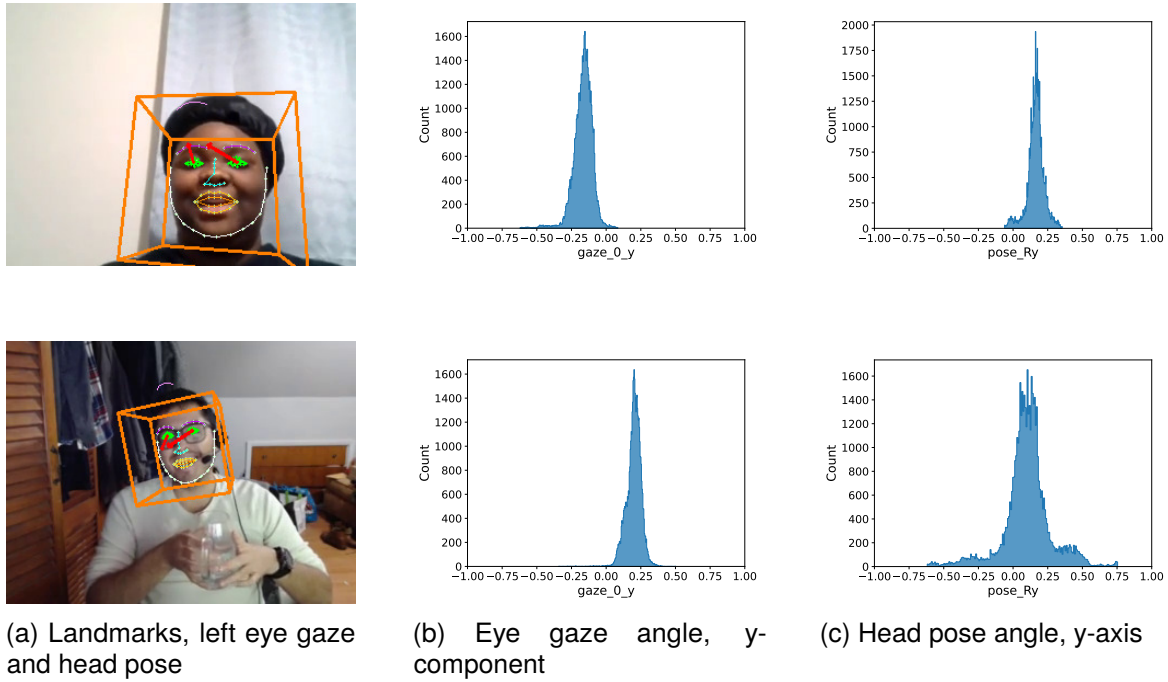


Figure 6.4 OpenFace gaze, head pose and facial landmark tracking for session *c97f8a32-6249-409c-b042-aaed261c3f64* of the Candor corpus (left) with gaze angle (centre) and head pose rotation (right) about the y-axis (horizontal direction).

Comparing across participants We inspect histograms of eye gaze and head pose to note their shape and see how they vary across participants. There is one gaze vector per eye which is a 3 dimensional unit vector. Head pose consists of position (X, Y, Z in mm) and angle (roll, pitch and yaw in radians). Inspection of the histograms shows they are unimodal, e.g. Figures 6.3b and 6.3c. Modal gaze and head pose value differs across participants. For example, the x component of eye gaze differs for the two participants in Figure 6.3b. This reflects the camera angle of one of the participants and is also reflected in the x component of head pose in Figure 6.3c. In the second example in Figure 6.4a, one participant is closer to the screen and the camera is below them, suggesting laptop use. The other participant is farther away from the screen. This leads to differences in modal y values for gaze (Figure 6.4b) and head pose (Figure 6.4c). We find this pattern is repeated for participants in the corpus.

As the Candor corpus is recorded *in-the-wild*, participants join from personal devices in uncontrolled environments. As the position of the participant relative to the camera varies, raw gaze and head pose angle carry little meaning across participants in the corpus. We must therefore be careful to standardise gaze and head pose prior to use, e.g. by subtracting the mean, discussed further on.

6.1.6 Discussion

Our exploratory analysis of FAUs demonstrates that there is a distinct pattern of FAU activity prior to shifts and holds. Immediately prior to the onset of speech, the next speaker exhibits increased mouth and jaw FAU intensity. On the other hand, during a hold where they are not about to speak, listener FAU activity resembles random silence. Similarly, the current speaker shows a differing pattern of FAU activity before a shift than before a hold. These patterns could be exploited in a PTTM, enabling holds and shifts to be predicted ahead of time.

However, our simplified analysis is exploratory. This is because FAUs are highly inter-correlated and combine thousands of ways to generate unique facial expressions (Tong et al. 2007). The temporal relationship between FAU is also richly complex (Tong et al. 2007). Logically, however, we did find that those action units connected with jaw movement were more active during speech than during silence. There was also notably more intense *inner brow raiser* activity from the speaker before a shift, a cue which has been linked to question marking (Nota et al. 2021). Thus, action unit activities might reflect more than just the action of speaking, but also turn-holding and yielding cues. Whatever their underlying role, we conclude from the analysis that FAUs have the potential to be features for turn-taking prediction, which we later confirm in our neural network model MM-VAP.

Our analysis of eye gaze and head pose revealed unimodal histograms, reflecting the dyadic nature of the interactions. The difference in individual participant setups, e.g. screen size and location of the participant relative to the camera, demands careful normalisation across participants. We repeated the analysis in Section 6.1 for eye gaze and head pose, considering velocity, acceleration and jerk within 500 ms windows. These did not reveal any statistically significant patterns, so we omitted heat maps from the analysis. Nevertheless, we include these features in our MM-VAP model as neural networks capture rich patterns.

A limitation of this analysis is that we only consider facial action unit activity in the final 500 ms of a speaking turn, but turn-taking cues can take place much earlier on in a turn (Nota et al. 2021). However, deep learning can capture complex relationships both between the action units and across long durations.

6.2 The multimodal voice activity projection model

Having now established a connection between visual features and turn-taking, we introduce the multimodal voice activity projection (MM-VAP) model. MM-VAP is our novel multimodal PTTM which integrates audio with FAU, head pose, eye gaze and facial landmark features from both interlocutors in a dyadic interaction.

We first describe the feature pre-processing in Section 6.2.1. Next, we outline the model architecture in Section 6.2.2 and the training process Section 6.2.3, before assessing the model performance in Section 6.3.

6.2.1 Feature selection and pre-processing

We first choose a set of video features from OpenFace to include in the model. Our choice is based on our previous exploratory analysis. We include all 17 FAUs intensity values. We also include eye gaze vectors from both eyes, head pose and selected landmark in the MM-VAP model. Eye gaze consists of two 3-dimensional unit vectors. Head pose comprises head position in millimetres (X, Y, Z) and head rotation in radians (roll, pitch and yaw), giving a total of 6 dimensions. We select 15 facial landmarks located on the brow, jaw, nose and lips in pixels (x and y) and a scalar tracking confidence score. In total, there is a 60-dimensional vector per frame of video, extracted at 30 frames per second.

Next, we standardise features at the participant level to remove participant-specific differences. We scale all features to 0,1 by max-min scaling at the participant level, except the scalar confidence score. To account for the previously observed differences in modal eye gaze and head pose position per participant, we zero-mean the head pose and eye gaze vectors at the participant level. We choose not to modify FAU intensity values as they are defined on their own scale. The confidence score is not modified. The video features and scaling operations are summarised in Table 6.1.

For the audio features, we use the same CPC front-end as the VAP model from Riviere et al. (2020), which outputs 1 feature for every 20 frames of 16 kHz audio (equivalent to 50 Hz). However, the video is 30 frames per second (30 Hz). We therefore upsample visual features from 30 to 50 Hz (a factor of 1.67) using linear interpolation to match the frame rate of the audio. This results in approximately time-aligned video and audio features.

6.2.2 Proposed multimodal turn-taking model architectures

Next, we outline the architecture of MM-VAP. We propose early and late modality fusion versions of the model, distinguished by whether audio and video features modalities are combined at an early stage in the model or at a later stage, prior to prediction. However, the

Table 6.1 Summary of features, their dimensionality, and standardisation operations applied to the OpenFace visual features extracted from the Candor corpus and utilised in the MM-VAP model. 0,1=max/min scaling to the 0,1 range. $\bar{x}$ =participant-level mean

Feature	Description	Units	Dimensionality	Standardisation
Facial action units	1 value per unit	custom scale	17	n/a
Eye gaze	3 per eye	unit vector	6	$0,1 - \bar{x}$
Head position	X, Y, Z	millimetres	3	$0,1 - \bar{x}$
Head rotation	roll, pitch, yaw	radians	3	$0,1 - \bar{x}$
Landmarks	x, y	pixels	30	$0,1 - \bar{x}$
Confidence	scalar	0–1	1	n/a
Total			60	

models could be equivalently referred to as late and early speaker fusion models respectively. This is because in each model there are two modalities, audio and video, and two speakers.

In the early modality fusion version, audio and video modalities early on in the model. Information from both speakers is combined at a later stage, in the final cross attention block 6.5a. Hence, the early modality fusion model could be equivalently termed a late speaker fusion model. Similarly, the late modality fusion model could be equivalently termed an early speaker fusion model. However, in this thesis, we use the terms early and late modality fusion to refers to the fusion of the audio and video modalities in the model, and not speaker-level information. This echoes the classic multimodal machine learning paradigm (Baltrušaitis, Ahuja, et al. 2018).

Early modality fusion version We first describe the early modality fusion version of the model. Like the VAP model (Inoue, Bing'er Jiang, et al. 2024) we used in Chapters 4, MM-VAP consists of self- and cross-attention blocks. The model is shown in Figure 6.5a. Audio and video modalities are merged early on within each speaker before any merging of information across speakers.

The audio front-end is the CPC encoder, which is frozen during training. The video features are the 60-dimensional features from OpenFace, projected to match the dimensionality of the audio features (256) via a trainable linear layer. Video and audio are fed to self-attention blocks (as described in Chapter 2). These blocks model temporal patterns within the audio and video modality. Each self-attention block has 1 self-attention layer and is identical to those used in the VAP model.

Self-attention blocks are followed by a cross-attention block, identical to the block used in the VAP model. They consist of three cross-attention layers. The block is fed the output of both the audio and video self-attention blocks. The cross-attention block is given by $N = 3$ stacked cross-attention transformer layers:

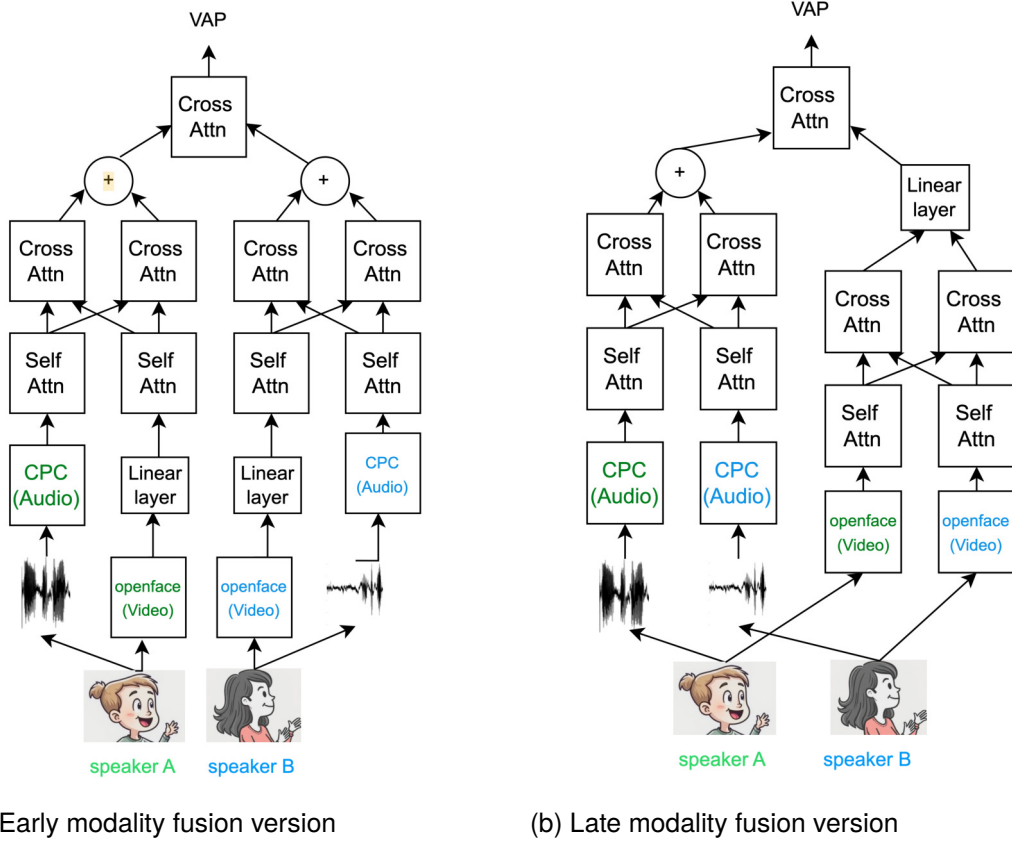


Figure 6.5 Schematic of two versions of the proposed multimodal voice activity projection (MM-VAP) model. Self-attn = self-attention block, cross-attn = cross-attention block, + = sum.

$$\text{CROSSATTNBLOCK}(v, a) = \left(\begin{array}{c} \text{TRANSFORMERLAYER}(q = v, k = a, v = a) \\ + \text{TRANSFORMERLAYER}(q = a, k = v, v = v) \end{array} \right) \times N$$

where a output of the audio self-attention block and v output of the video self-attention block, q, k, v are the query, key and value. The block therefore uses cross-attention to combine audio and video features of each speaker. Both speakers are then combined in a final cross-attention block. The final layer is a linear layer followed by a softmax outputting the same 256 dimensional prediction as the original VAP model.

Late modality fusion version The late modality fusion version of the MM-VAP model, depicted in Figure 6.5b, is very similar; however, video and audio are kept separate until just prior to prediction. Video from both speakers is combined in the first cross-attention block. Separately, audio from both speakers is combined in another cross-attention block. There is

Table 6.2 Multimodal voice activity projection (MM-VAP) model configuration parameters

Model Block	Parameter	Value
Audio encoder	encoder	CPC (Riviere et al. 2020)
	audio sample rate	16000 Hz
	prediction sample rate	50 Hz
	sequence length	20 seconds
	frozen	True
	pre-trained	True
Video encoder	features	OpenFace (Baltrušaitis et al. 2016)
	d_in	60
	d_out	256
	video sample rate	30 Hz
	upsampling factor	1.67
	sequence length	20 seconds
Self-attention block (audio)	d_in	256
	d_out	768
	number of attention heads	4
	number of layers	1
	sequence length	1000 frames
Self-attention block (video)	d_ffn	174
	model dimensionality	60
	number of attention heads	4
	number of layers	1
	sequence length	600 frames
cross-attention block	d_model	256 (or 60 in late modality fusion)
	number of attention heads	4
	number of layers	3

a learned projection layer to bring the video features from 60 to 256 dimensions, matching the audio dimensionality.

The audio and video modalities are joined later on the model by a cross-attention layer just prior to prediction. As before, a linear layer followed by a softmax operation outputs a 256-dimensional prediction, which is the same as the original VAP model.

Technical specification The configuration of both models is detailed in Table 6.2. All transformer attention layers use Alibi attention, which applies causal masking to ensure the model cannot use future information to make predictions (Press et al. 2021).

We use GeLU activation function (M. Lee 2023), dropout (Srivastava et al. 2014), 4 attention heads and set the transformer feed-forward network dimensionality to 768 for the audio block. For the video block, we set the dimensionality to 174. As before, self-attention blocks have 1 transformer layer and cross-attention blocks have 3 transformer layers, matching the VAP model. Audio and video are fed in batches of 20 seconds chunks (1000 audio features and 600 video features prior to upsampling), again identical to the previous audio-only VAP model.

Video-only model To quantify how well video alone can predict turn-taking, we also include a PTTM which uses only video features. The architecture is identical to the audio-only VAP model described in Chapter 2, with the audio encoder removed and the dimensionality of the model changed from 256 to 60 to match the video feature dimensionality.

Table 6.3 Epochs corresponding to the lowest validation loss for each fold of cross-validation.

Model	Lowest Validation Loss Epoch				
	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
MM-VAP (early modality fusion)	2	7	8	1	8
MM-VAP (late modality fusion)	9	6	5	8	8
Video-only	4	7	4	5	1

6.2.3 Training and evaluation procedure

Training We then train the model using a cross-validation procedure identical to the procedure that we used for the VAP model outlined in Section 4.1.3. We train each fold for 10 epochs with a learning rate of $5e-4$. This is a decrease from the earlier learning rate of the VAP model, but we find that reducing the learning rate gives better convergence for the MM-VAP model. The batch size is reduced from 64 to 32 in order to fit on the Nvidia A6000 GPU. The MM-VAP model is larger than the VAP model at approximately 5.8 M trainable parameters. The GPU hours are early modality fusion 110 hrs, late modality fusion 113 hrs.

We show curves of the training and validation losses in Figures 6.6a and 6.6b, where we show the losses alongside the audio-only VAP model for comparison. The loss training and validation loss average is plotted with the standard deviation shaded. Multiple informal attempts were made at learning rate adjustment, but we ultimately found that achieving convergence with MM-VAP was more challenging, as indicated by the wide standard deviation in training and validation losses in Figures 6.6a and 6.6b. Nevertheless, convergence was achieved in some folds of cross-validation. The epoch with the lowest validation loss for each model is shown in Table 6.3, and the weights from these epochs are used in all subsequent analyses.

Evaluation procedure We assess model performance in the same manner as Chapter 4. We compare the mean balanced accuracy, F_1 (Weighted), F_1 (hold) and F_1 (shift) scores of MM-VAP, VAP and video-only models using the and shift prediction performance metrics, fully described in Section 4.2. The results for the VAP model are taken from Chapter 4. Statistical significance is computed with the paired t-test, as before.

Ablation study We conduct an ablation study by training 4 different versions of the MM-VAP model (late modality fusion), each on a different subset of visual features. We divide the visual features into gaze, head pose, facial action unit and facial landmarks groups (6, 6, 17, 30 dimensions respectively, excluding the scalar confidence score). Each model receives audio and a subset of the visual features as input, with the dimensionality of the video features adjusted accordingly. Training and validation follow the same procedure as before.

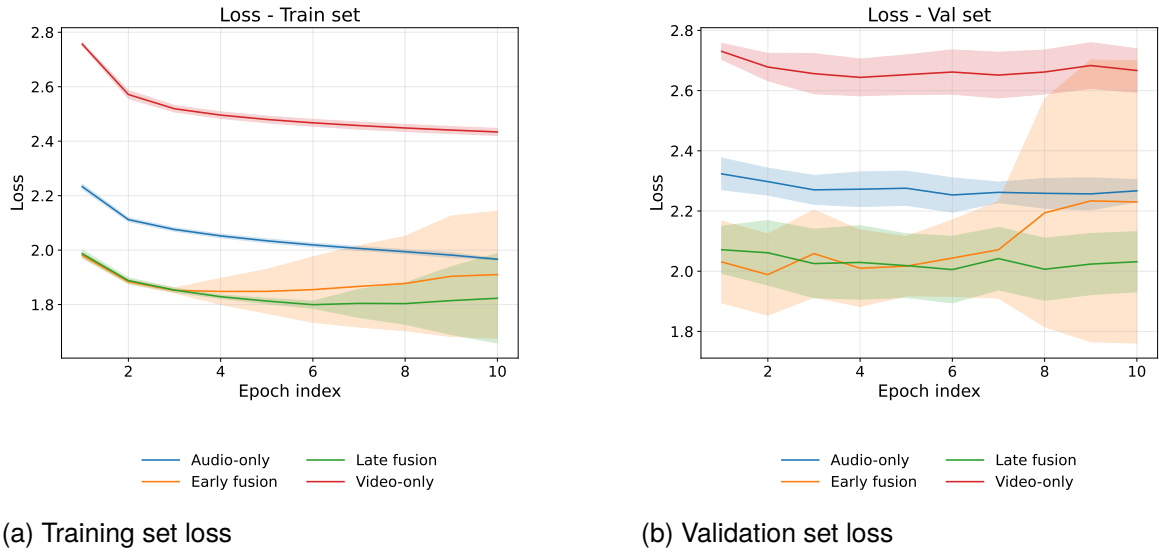


Figure 6.6 Training and validation losses for the MM-VAP early and late modality fusion models, compared with the audio-only VAP model and the video-only turn-taking model, on the Candor corpus. A 5 fold average loss is plotted and the standard deviation of the loss is shaded.

6.3 Assessing the performance of the multimodal voice activity projection model

After training, we establish the performance of MM-VAP relative to audio-only and video-only turn-taking models. This reveals that MM-VAP outperforms both audio-only and video-only models. In our ablation study, we find FAUs are the largest contributors to the increased performance of MM-VAP, followed closely by head pose.

6.3.1 Comparing MM-VAP with VAP

We report the 5-fold average test set performance of MM-VAP (audio+video), VAP (audio) and the video-only model on the Candor corpus in Table 6.4.

During mutual silence We first consider hold/shift prediction during mutual silence. This represents the model’s ability to distinguish holds from shifts immediately after a turn has finished. The late modality fusion model shows a 4% relative increase in the weighted F_1 score and a 5% relative increase in balanced accuracy over the audio-only model VAP (*during mutual silence, $FTO > +250$ ms*, Table 6.4). The early modality fusion model similarly shows a significant increase over the audio-only VAP model, equivalent to a 4% relative increase in balanced accuracy. Paired t-tests comparing VAP with MM-VAP reveal that this performance increase is statistically significant (Appendix B).

Before the end of a turn We then consider hold/shift prediction before the end of a speaking turn. This captures the ability of the model to predict upcoming shifts while the previous speaker is still speaking, akin to human turn-taking (Sacks et al. 1974). Best performance is achieved with the late modality fusion model, with a 5% relative increase in the F_1 shift score and a 4% relative increase in balanced accuracy over the audio-only VAP model (*before end of turn*, $FTO > +250$ ms, Table 6.4). The early modality fusion model also outperforms the audio-only model, with a 3% relative increase in the F_1 shift score and a 3% relative increase in balanced accuracy. Paired t-tests show the performance increase is statistically significant (Appendix B).

Next, we remove the requirement for a minimum period of silence between turns. This increases the number of speaking turns used for evaluation (Section 4.2). For this more comprehensive set of speaker transitions, MM-VAP outperforms the audio-only models, with the late modality fusion model achieving an 8% relative increase in the F_1 shift score and a 6% relative increase in balanced accuracy over the audio-only VAP model (*before end of turn*, $FTO > +0$ ms, Table 6.4). The performance increase is again statistically significant.

Before overlapping speech Next, we consider overlapping speech. Here we find the largest relative increase in MM-VAP performance over VAP. There is a 6% increase in balanced accuracy and a 9% increase in the F_1 shift scores for the late-fusion MM-VAP model. The early modality fusion model achieves a 7% increase in the F_1 shift score (*before overlap*, $FTO < -250$ ms, Table 6.4). Paired t-tests reveal this increase is statistically significant, demonstrating that through the inclusion of visual cues, the MM-VAP model outperforms the audio-only model (Appendix B).

Shift prediction Finally, we consider shift prediction, which is the model's ability to distinguish speech at the end of a turn before a shift with mid-turn speech. The results are shown in Table 6.5. On this metric, the MM-VAP models still outperform the audio-only VAP model, but the performance increase is less pronounced. Again, paired t-tests reveal this difference is statistically significant (Appendix B).

6.3.2 Comparing early and late modality fusion versions of MM-VAP

Next, we compare late and early modality fusion versions of MM-VAP. Both models outperform the audio-only model, however, we find there is no significant difference between the early and late modality fusion MM-VAP models, so in all subsequent work, we use the late modality fusion version of MM-VAP. The paired t-tests are listed in Appendix B.

Table 6.4 Average hold/shift prediction across 5 folds on the Candor corpus (best in bold). a = audio only, v = video-only, and a+v = multimodal; e = early modality fusion, l = late modality fusion. Percentage change relative to the audio-only model is provided as $\uparrow / \downarrow$, with ‘–’ indicating $p > 0.05$.

Hold/shift prediction metrics, Candor corpus						
Evaluation point	Corpus	F_1 (Weighted)	F_1 (Hold)	F_1 (Shift)	Accuracy (Balanced %)	
during mutual silence (FTO > +250 ms)	a	0.83	0.88	0.70	79	
	v	0.72 $\downarrow$ 13%	0.82 $\downarrow$ 7%	0.47 $\downarrow$ 33%	68	$\downarrow$ 14%
	a+v (e)	0.84 –	0.90 –	0.74 $\uparrow$ 6%	82	$\uparrow$ 4%
	a+v (l)	0.86 $\uparrow$ 4%	0.90 $\uparrow$ 2%	0.74 $\uparrow$ 6%	83	$\uparrow$ 5%
before end of turn (FTO > +250 ms)	a	0.81	0.87	0.67	77	
	v	0.70 $\downarrow$ 14%	0.80 $\downarrow$ 8%	0.45 $\downarrow$ 33%	66	$\downarrow$ 14%
	a+v (e)	0.83 $\uparrow$ 3%	0.88 –	0.69 $\uparrow$ 3%	79	$\uparrow$ 3%
	a+v (l)	0.83 $\uparrow$ 3%	0.89 $\uparrow$ 2%	0.70 $\uparrow$ 5%	80	$\uparrow$ 4%
before end of turn (FTO > 0 ms)	a	0.86	0.91	0.66	78	
	v	0.78 $\downarrow$ 9%	0.87 $\downarrow$ 4%	0.43 $\downarrow$ 35%	70	$\downarrow$ 10%
	a+v (e)	0.87 –	0.92 –	0.69 $\uparrow$ 5%	80	$\uparrow$ 3%
	a+v (l)	0.87 –	0.92 –	0.71 $\uparrow$ 8%	83	$\uparrow$ 6%
before overlap (FTO < -250 ms)	a	0.78	0.85	0.57	70	
	v	0.72 $\downarrow$ 8%	0.82 $\downarrow$ 4%	0.42 $\downarrow$ 26%	64	$\downarrow$ 9%
	a+v (e)	0.79 $\uparrow$ 1%	0.87 $\uparrow$ 2%	0.60 $\uparrow$ 5%	72	$\uparrow$ 3%
	a+v (l)	0.80 $\uparrow$ 3%	0.87 $\uparrow$ 2%	0.62 $\uparrow$ 9%	74	$\uparrow$ 6%

6.3.3 Comparing MM-VAP with a video-only model

To establish to what extent video alone can predict turn-taking, we explore the video-only model. The video-only model has the worst overall performance with a 33% relative reduction in the F_1 shift score for hold/shift prediction (*during mutual silence*, (FTO > +250 ms, Table 6.4). In all cases, the MM-VAP (audio+video) outperforms VAP (audio), which in turn outperforms the video-only model.

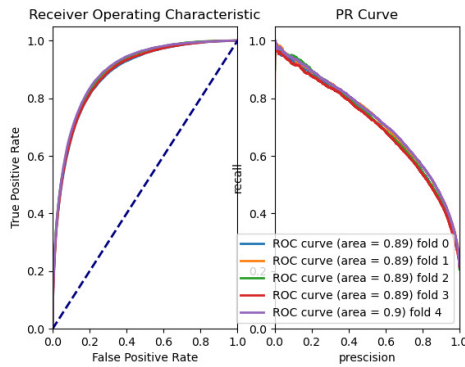
6.3.4 ROC and PR curves

Next, we explore receiver operating characteristic (ROC) and precision-recall (PR) curves. Figure 6.7 shows the curves for the hold/shift prediction task, evaluated before the end of a speaking turn. Comparing MM-VAP with VAP reveals an improvement in the ROC curve. The area under the ROC curve is 0.89 for VAP (Figure 6.7a) compared with 0.91 for MM-VAP (Figure 6.7b). Inspection of the ROC curve shows that for a false positive rate of 0.3, the corresponding true positive rate is 0.87 for VAP vs 0.93 for MM-VAP. Similarly, inspection of the PR curve reveals increased precision of the MM-VAP model. For example, at a precision of 0.4, the recall of VAP is 0.87 compared with 0.92 for MM-VAP. We also consider the shift prediction metric in Figure 6.8. This similarly reveals that the area under the ROC curve has increased slightly, from 0.93 in VAP (Figure 6.8a) to 0.95 in MM-VAP (Figure 6.8a).

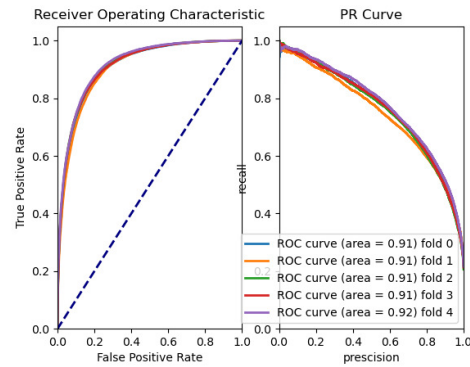
The improved true positive rate and recall of the MM-VAP model reflects the slight increase in prediction performance previously observed in Section 6.3. The analysis shows that the MM-VAP model is able to make more correct predictions than the VAP model. The difference

Table 6.5 Average shift prediction across 5 folds on the Candor (CND) corpus (best in bold). a = audio only, v = video-only, and a+v = multimodal; e = early modality fusion, l = late modality fusion. Percentage change relative to the audio-only model is provided as $\uparrow / \downarrow$, with ‘-’ indicating $p > 0.05$.

Shift prediction metrics, Candor corpus							
Evaluation point	Corpus	F_1 (Weighted)		F_1 (Hold)		F_1 (Shift)	Accuracy (Balanced %)
before end of turn (FTO > +250 ms)	a	0.86		0.84		0.87	85
	v	0.73	$\downarrow 15\%$	0.69	$\downarrow 18\%$	0.76	72 $\downarrow 15\%$
	a+v (e)	0.87	$\uparrow 1\%$	0.86	$\uparrow 2\%$	0.88	87 $\uparrow 2\%$
	a+v (l)	0.87	$\uparrow 1\%$	0.86	$\uparrow 2\%$	0.88	87 $\uparrow 2\%$
before end of turn (FTO > 0 ms)	a	0.86		0.84		0.87	86
	v	0.78	$\downarrow 9\%$	0.87	$\uparrow 4\%$	0.43	70 $\downarrow 19\%$
	a+v (e)	0.88	$\uparrow 2\%$	0.86	$\uparrow 2\%$	0.89	88 $\uparrow 2\%$
	a+v (l)	0.87	$\uparrow 1\%$	0.86	$\uparrow 2\%$	0.89	87 $\uparrow 2\%$



(a) VAP, CND, hold/shift prediction



(b) MM-VAP (late), CND, hold/shift prediction

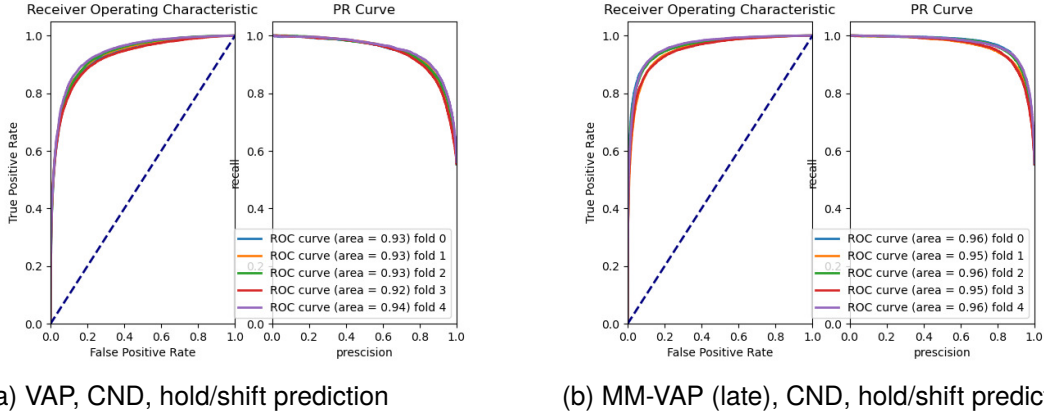
Figure 6.7 Hold/shift prediction, before end of turn FTO > +0 ms

is very slight, which indicates the performance improvement originates from increased performance on a small subset of turns in the corpus.

6.3.5 The effect of transition length

In the previous analysis, we showed that the MM-VAP model is the best-performing model overall. As highlighted, the increased performance is small, indicating an increased performance on a small subset of turns in the corpus. This subset of turns may depend more on visual cues. To explore this further, we group speaker transitions by the period of silence between speaking turns. Our logic is that longer periods of silence might be more visual, for instance, by being accompanied by an averted gaze.

We group turns by varying the minimum floor-transfer offset (FTO) between speaking turns from 0 to 1500 ms in 250 ms increments. We then compute model performance on each subset. We use the balanced accuracy because, unlike the F_1 weighted score, it equally weights holds and shifts. This is important as proportions of shifts relative to holds differ by

Figure 6.8 Shift prediction, before end of turn $FTO > 0$ ms

the minimum FTO between turns, as highlighted in Table 4.5. We also include performance on Switchboard with the VAP model for comparison.

We plot the mean balanced accuracy over 5 folds in Figure 6.9. Balanced accuracies in Table 6.4 correspond with the 0 and 250 ms points. The gap in performance of Switchboard and Candor is consistent with our findings in Chapter 4. Performance falls for both VAP and MM-VAP models as the duration of silence between speaking turns increases, indicating that these turns are more challenging to predict. However, we find that the performance of the MM-VAP models is significantly better than the audio-only VAP models across all FTOs ($p < 0.01$ in all cases). Trends are similar for both VAP and MM-VAP.

The increased performance is therefore not explained by an increased performance on holds and shifts with longer silences between speaking turns. The increased performance of the MM-VAP model is therefore consistent for all FTOs between speaking turns.

6.3.6 Ablation study

Finally, we conduct an ablation study, training the late modality fusion version of the MM-VAP model on audio and gaze, head pose, facial landmarks and facial action unit subsets. We report the 5-fold average performance in Table 6.6.

The best-performing model includes all visual features (a + v (I), Table 6.6). The models trained on the four subsets of the visual features all show reduced performance. We compare the weighted $F1$ score, and find that the gaze and landmark trained models perform significantly worse than the model trained on all visual features. There is no significant difference in the performance of the FAU and head pose trained models on certain metrics (e.g. $p = 0.09$, before end of turn $FTO > 0$ ms Table 6.6). However, on the $F1$ shift score, the FAU trained model achieves the best performance of all models trained on the reduced visual feature sets ($p < 0.05$).

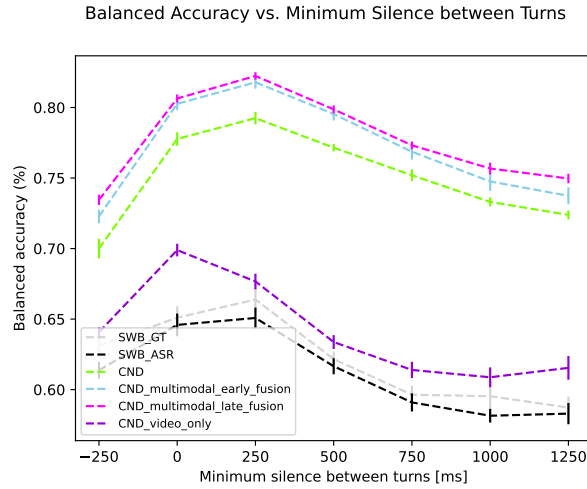


Figure 6.9 Mean balanced accuracy of models averaged over 5 folds on the candor and switchboard (SWB) corpora, $\pm$ standard error in the mean, grouped by a minimum period of silence between turns; i.e. the FTO. The GT=ground-truth / ASR word alignment used for training and testing.

The best-performing model trained on a subset of visual features is therefore the FAU model, closely followed by the model trained on head pose. We therefore conclude from the ablation study that the FAUs and head pose features contribute the most to the increased performance of MM-VAP over VAP.

6.3.7 Discussion

In this section, we have evaluated our proposed MM-VAP PTTM, which integrates video and audio features. We compared its performance to the audio-only VAP model on the Candor corpus of videoconferencing interaction.

In Section 6.3, we found that MM-VAP consistently outperformed VAP across a wide range of turn-taking metrics. We found the performance increase of MM-VAP was statistically significant in each case, e.g. 83% vs. 79% balanced accuracy on hold/shift prediction (Table 6.4). As a video-only model performed worse than VAP and MM-VAP, (Section 6.3.3), this demonstrates that both audio and video cues support turn-taking prediction. We also analysed the model further with an ablation study Section 6.3.6. This revealed that FAU and head pose features contributed the most to the increased performance of MM-VAP over VAP. Gaze and facial landmarks did not contribute much to the model's performance. A strong relationship between turn-taking behaviours and patterns of eye gaze has been noted in the literature (Kendon 1967). Although we did normalise to account for differences across speakers, it could be that videoconferencing distortion of gaze vectors means the model is unable to learn useful patterns across speakers. Alternatively, gaze estimation with a single

Table 6.6 Ablation study results, 5-fold average. a=audio, v=video, a+v (l) = multimodal, late modality fusion. Percentage change relative to the a+v (l) model, which uses all visual features, is shown as $\uparrow / \downarrow$, with '-' indicating $p > 0.05$.

Hold/shift prediction metrics, Candor corpus							
Evaluation point	Corpus	F_1 (Weighted)		F_1 (Hold)		F_1 (Shift)	Accuracy (Balanced %)
during mutual silence (FTO > +250 ms)	a + v (l)	0.86		0.90		0.74	83
	a + gaze	0.74	↓14%	0.83	↓8%	0.51	67 ↓19%
	a + pose	0.84	↓2%	0.89	-	0.72	80 ↓4%
	a + lmks	0.79	↓8%	0.86	↓4%	0.62	74 ↓11%
	a + faus	0.85	-	0.89	-	0.73	81 ↓2%
before end of turn (FTO > 250 ms)	a + v (l)	0.83		0.89		0.70	80
	a + gaze	0.72	↓13%	0.82	↓8%	0.48	65 ↓19%
	a + pose	0.82	-	0.87	-	0.68	78 ↓2%
	a + lmks	0.78	↓6%	0.85	↓4%	0.59	72 ↓10%
	a + faus	0.83	-	0.88	-	0.69	79 ↓1%
before end of turn (FTO > 0 ms)	a + v (l)	0.87		0.92		0.71	83
	a + gaze	0.78	↓10%	0.88	↓4%	0.43	65 ↓22%
	a + pose	0.87	-	0.92	-	0.67	79 ↓5%
	a + lmks	0.83	↓5%	0.90	↓2%	0.57	73 ↓12%
	a + faus	0.87	-	0.92	-	0.68	80 ↓4%
overlap (FTO < 250 ms)	a + v (l)	0.80		0.87		0.62	74
	a + gaze	0.68	↓15%	0.81	↓7%	0.35	58 ↓22%
	a + pose	0.79	↓1%	0.86	-	0.58	71 ↓4%
	a + lmks	0.74	↓8%	0.84	↓3%	0.49	66 ↓11%
	a + faus	0.79	↓1%	0.86	-	0.59	72 ↓3%

camera setup is highly challenging (G. Yang and Sanii 2016), and this could simply reflect noisy or inaccurate gaze vectors.

It is important to note that although the magnitude of the performance increase is small, this does not indicate that visual cues are of limited benefit or utility for prediction. If the performance is on a small number of critical, highly salient turns that rely on visual cues, this could mean the difference between a successful and unsuccessful deployment of a turn-taking model. In Section 6.3.5, we grouped speaker transitions by the period of silence between speaking turns and found the improvement was consistent across all speaker lengths. The performance improvement does therefore not originate from an increased performance on turns with a longer or shorter FTO. A detailed analysis using Conversation Analysis methodology (Clift 2016) should be conducted to uncover the exact role visual cues play in select examples, but this is beyond the scope of this thesis.

Similarly, it is challenging to determine whether visual features contribute to performance simply because they mirror speaking activity, or capture known non-verbal turn-taking cues highlighted in our literature review in Chapter 2. As FAUs contribute the most to performance, aspects such as the increased jaw movement during speaking observed in our analysis of FAUs in Section 6.1 could be learned by the model. Alternatively, more complex relationships could be learned. For instance, Our finding of enhanced lip jaw and cheek movement before speaking indicates the presence of visual cues from the listener prior to the onset of speech (a shift) in Section 6.1 could be exploited. This could explain our finding that visual cues

most benefited turn-taking by improving the performance on shifts (F_1 shift, Table 6.4). Our analysis of distinct patterns of FAU activity before the onset of speaking turns in Section 6.1 further strengthens the argument that the model is exploiting mouth and facial movements prior to the onset of speech, rather than simply correlating with speaking activity.

Limitations There are a number of limitations to address. Firstly, the model is dependent on the performance of OpenFace, which does not handle participants leaving and re-entering the frame. Our solution was to exclude these sessions, which did reduce the amount of data by approximately 200 hours, but we still trained on 710 hrs, significant more data than other studies. We also upsampled the OpenFace features to account for differences in audio feature and video feature frame rates. Although the model is causally masked, an inevitable consequence of the upsampling is the introduction of a slight future 'bleed' as present frames are modified by the next frame through interpolation. Finally, we explored videoconferencing, which is known to alter, distort and reduce non-verbal communication (O'Conaill et al. 1993; Sellen 1995). Interlocutors therefore exhibit altered non-verbal communication in comparison to in-person interaction. However, we did find that including the visual cues improved prediction, which does suggest some role of visual cues in the interaction. We used videoconferencing because at the time of writing no corpus of significant size with audio and video was available to support turn-taking research. However, in the time since this work has conducted, 4,000 hour *Seamless Interaction Dataset* of dyadic casual conversation in-person has been released by Meta (Agrawal et al. 2025). The study should be repeated in-person, as discussed in Chapter 8.

Contributions The work on incorporating visual cues in this chapter builds on Chapters 4 and 5, which explored prosody and lexical cues. The improved performance of MM-VAP demonstrates that, in addition to these cues from speech, visual cues can further improve performance in predictive turn-taking. These findings therefore support the highly multimodal nature of human turn-taking (Holler and Levinson 2019; Holler, Kendrick, et al. 2016).

To our knowledge, our work represents one of the first studies on including visual cues in PTTM for two-party interaction. Although earlier work from Roddy, Skantze, et al. (2018b) did consider visual cues, the dataset size was limited (11 hrs), and the RNN-based architecture is now outdated. Furthermore, their model considered only gaze. In contrast, our work considers a much larger 710 hr dataset, with many more speakers. We also utilise a number of visual cues and conduct a through ablation study. Our work therefore conclusively demonstrates that the inclusion of cues improves model performance, and therefore represents a substantial advancement in the state-of-the-art in predictive turn-taking. A number of studies have been developed independently of our own and at the same time as our work (Saga and Pelachaud

2025; Ishii et al. 2025; Onishi et al. 2024). These studies are discussed further on in Chapter 8.

6.4 Qualitative analyses

Our results have shown the MM-VAP model has a slight but significant increase in turn-taking prediction performance over the current state-of-the-art audio-only VAP turn-taking model. In the following section, we qualitatively examine the prediction of MM-VAP and compare to VAP and video-only models in examples to understand its behaviour. We also visualise the internal transformer attention in the model.

6.4.1 Examples of speaker transitions

To explore the output of the models, we run MM-VAP, VAP and the video-only turn-taking model on a session chosen at random from the Candor corpus. This session was unused during training in the first cross-validation fold, so we run each model trained on the first fold by choosing the epoch with the lowest validation loss. We explore the performance of MM-VAP and VAP by plotting their p_{future} probability: the marginal probability one of the two speakers will speak in the next 800-1200 ms seconds (Chapter 2).

Model probability during a shift We first consider the performance during a transition from one speaker to another (a shift) in Figure 6.10. A transition from one speaker (in grey) to the other speaker (in red) with a short silence between speaking turns is depicted. The model output probabilities are expressed as the probability that the speaker depicted in red will speak and are plotted below (Figure 6.10a). The MM-VAP model is shown in the solid blue line, and the VAP model is shown in the dashed blue line. Where the probabilities exceed 0.5, they are highlighted in red.

The model probabilities show that both models predict the transition to the red speaker ahead of time. However, MM-VAP predicts the transition sooner, approximately 300 ms before VAP and with greater confidence. The video-only model is shown in the dashed black line. The probability is mostly around 0.8, indicating that no speaker transition is predicted. The video only model is depicted by the dashed black line.

We show selected video frames from both participants during the shift in Figure 6.10b, coloured red and grey to match the waveforms. We note the participant shown in red, who is about to speak, has a slightly open mouth prior to the onset of speech. We also note that this participant turns their head and averts their gaze at the start of the speaking turn. Further detail of these behaviours is shown in Figure 6.11.

Model probability during mid-turn speech We continue with an example of model output during mid-turn speech, located away from a speaker transition in Figure 6.12. As before the waveform and model probabilities are depicted in Figure 6.12a.

Comparing MM-VAP (solid blue) with VAP (dashed blue), we note that VAP falsely predicts a transition to the speaker depicted in grey. However, the MM-VAP model correctly does not predict any speaker change. As before, the video-only model does not predict a speaker change; however, it now predicts the incorrect speaker. We also note there is a false positive shift predicted by VAP that occurs during the word “*family*” in the utterance “... *was down there before I moved back home closer to family*”. Exploring the video of participants in 6.12b, we see gaze aversion and head tilt behaviour by the speaker. Although not captured explicitly by MM-VAP as it is restricted to facial features, we also see hand gesturing. We show further detail of the speaker during the word “*family*” in Figure 6.13. Notice the gaze aversion and facial expression.

Model probability during a hold We show an example of model probability during a hold, in Figure C.1 in Appendix C. All models correctly predict that the current speaker will continue after the hold.

6.4.2 Discussion

Our analysis of probability traces illustrates that MM-VAP and VAP have different output probabilities. In the first example, we saw that MM-VAP predicted the speaker transition earlier and with more confidence. Visualisation of the listener behaviour showed gaze aversion and mouth opening prior to the onset of speech. In the second example, we saw that VAP falsely predicted a speaker transition, whereas MM-VAP did not. Again, we saw gaze aversion and facial expression from the speaker during the hold.

This qualitative analysis is illustrative; however, we note that gaze aversion is a strong turn-taking cue from our literature review in Chapter 2. We also found evidence that the next speaker has their mouth open prior to a shift, which we demonstrated in Section 6.1. Moreover, our analysis in Section 6.3 demonstrated that the performance improvement of MM-VAP is evident at the corpus-level in all turn-taking metrics considered. The increased performance may come from increased performance on speaker transitions that involve non-verbal cues, as it is able to incorporate cues from the speaker and listener which indicate whether or not they are about to start speaking.

6.5 Attention analysis

In a final analysis, we visualise the attention of the turn-taking models in a sample pair of turns between two speakers. The purpose of this analysis is to explore how information from both speakers is utilised for turn-taking prediction. We also explore whether models utilise information from the current turn or the previous turn in making their predictions. We choose the example previously depicted in Figure 6.10a, which is a shift from one speaker to another, depicted below:

- **Speaker A** “um would you feel comfortable telling me about what you do for a living”
- **Speaker B** “uh yeah well for the most part like i test and tweak algorithms”

6.5.1 Aggregating word-level attention

We visualise the attention of the model on this example, which offers insights into the contextual and cross-modal dependencies learned by the models. We analyse attention as follows. There is one attention score per frame in the turn-taking model, and the sample rate is 50 Hz. The attention between a query frame and a value frame is given by the formula:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (6.1)$$

where d_k is the dimensionality of the key vectors (256). Each element A_{t_q, t_k} represents the attention weight¹ between a query frame t_q and a key frame t_k . This indicates how much the model attends to a particular key at time t_k . As visualising frame-wise attention is challenging, we aggregate attention scores at the word level using the time-aligned transcriptions. Each word W spans a number of frames. We then compute the mean attention score between pairs of words by averaging the frame-level attention weights:

$$\bar{A}_{ij} = \frac{1}{|W_i||W_j|} \sum_{t_q \in W_i} \sum_{t_k \in W_j} A_{t_q t_k} \quad (6.2)$$

Here, $\bar{A}_{ij}$ denotes the mean attention from the frames spanned by the word i (query) to word j (key). This aggregation provides a word-level representation of attention that can be more easily interpreted than frame-level attention. We conduct this process for each layer of the VAP and MM-VAP models, and then visualise the resulting word-level attention matrices $\bar{A}$ as heat maps.

¹The models have multi-head attention, so we consider the average attention of all heads

6.5.2 The VAP audio-only model

We begin with a visualisation of the first self-attention layer from Speaker B. This is the second speaker, who begins talking after being posed the question “*would you feel comfortable telling me about what you do for a living*” from Speaker A. The model consists of self-attention (per speaker) and cross-attention (combining both speakers) blocks.

Figure 6.14 shows the word-level self-attention map for Speaker B. Query words (W_i) are on the y-axis and Key words (W_j) are on the x-axis. The Figure has a strong diagonal component, showing that in this early self-attention layer, the context is local. When Speaker B is not actively speaking (i.e. during Speaker B’s turn in grey), attention is present. Note that attention scores in the upper-right triangle are all 0. This is because causality is enforced in all turn-taking models in this thesis, meaning there is no attention from the current frame to future frames.

Next, we visualise the last layer of the cross-attention block in Figure 6.15. The plot demonstrates that, as before, the context is mostly local. However, there is now a small attention concentrated on the word “*would you feel comfortable*”. However, the attention between the two speaking turns is limited.

6.5.3 MM-VAP, early modality fusion

We continue by visualising the attention of the early modality fusion version of our MM-VAP model. The early modality fusion approach learns to combine audio and video modalities separately for each speaker, prior to combining information across the two speakers.

We find the initial video and audio self-attention blocks capture local context (Figures C.2 and C.3, Appendix C). However, the attention in later layers reveals longer-term dependencies. The cross-attention from video to audio for Speaker B is shown in Figure 6.16. In addition to the diagonal pattern, which represents local context, strong vertical lines have appeared. This represents attention from the audio of Speaker B’s turn to the video of Speaker B during the previous turn.

The attention between turns is concentrated around “*feel comfortable telling*” and “*do for a living*”. These are both points at which Speaker A’s turn is a complete sentence; however, there is no audio or video from Speaker A at this stage of the network: only audio and video from Speaker B are captured at this stage. Therefore, the model is detecting salient visual information from Speaker B at these points in time. We show how the attention changes in a later layer of the network when full audio and video information from both speakers is present in Figure 6.17. The attention between the turns is now concentrated mostly around the final word “*living*”, where the turn actually ends.

6.5.4 MM-VAP, late modality fusion

We conduct the same visualisation on the late modality fusion version of MM-VAP. This version models dependencies between video or audio streams of the two speakers separately, prior to combining the two modalities late in the network, before prediction. As with the early modality fusion model, we find the initial layers capture local context (Figures C.4 and C.5, Appendix C).

However, from the initial cross-attention layers, attention patterns emerge between turns. Figure 6.18 shows the cross-attention from the video of Speaker B to the video of Speaker A. Frames in Speaker B attend to frames within the word “*you*” in “*would you feel comfortable*”. There is also some attention in the final words of the turn from Speaker B to the start of Speaker A’s turn. This suggests salient information in the visual channel of Speaker A, which the cross-attention captured. The audio attention does not show this pattern (Figure C.4, Appendix C). In contrast with the audio, the video stream contains salient information when the person is not currently speaking.

In the last block of the model, which has full access to both audio and video cues from both speakers, we note that the attention pattern has shifted considerably. The attention between Speaker B’s turn and Speaker A’s turn is concentrated on the final words “*a living*”, where the turn actually ends.

6.5.5 Discussion

In this analysis, we compare the attention scores for VAP and MM-VAP models. The dependency of the model across turns has been shown for the text-based model, TurnGPT (Ekstedt and Skantze 2020), but to our knowledge the VAP model has not been explored to date.

In our analysis, we show that longer-term dependencies are present in the attention scores of MM-VAP models than in the VAP model. This reveals that both models utilise context across turns. We showed that MM-VAP has higher attention scores across turns than VAP. This indicates that the model utilises more information in the previous turn to make its predictions about upcoming speaking activity.

Although this analysis is brief and on a single example, we speculate that this shows the MM-VAP model is capable of better modelling of context across speaking turns, which might contribute to its increased performance. Our reasoning is that the inclusion of non-verbal cues in the model enables it to develop this richer context. In the audio only model, the listener is silent, meaning their audio channel has little useful information for predictive turn-taking. However, in the MM-VAP model, both audio and video is used. This means that when the listener is not speaking, their non-verbal information is still present in the features. Non-verbal cues from the listener such as the mouth opening prior to the onset of speech that were

explored in Sections 6.1 and 6.4 are therefore present and can be exploited for turn-taking prediction.

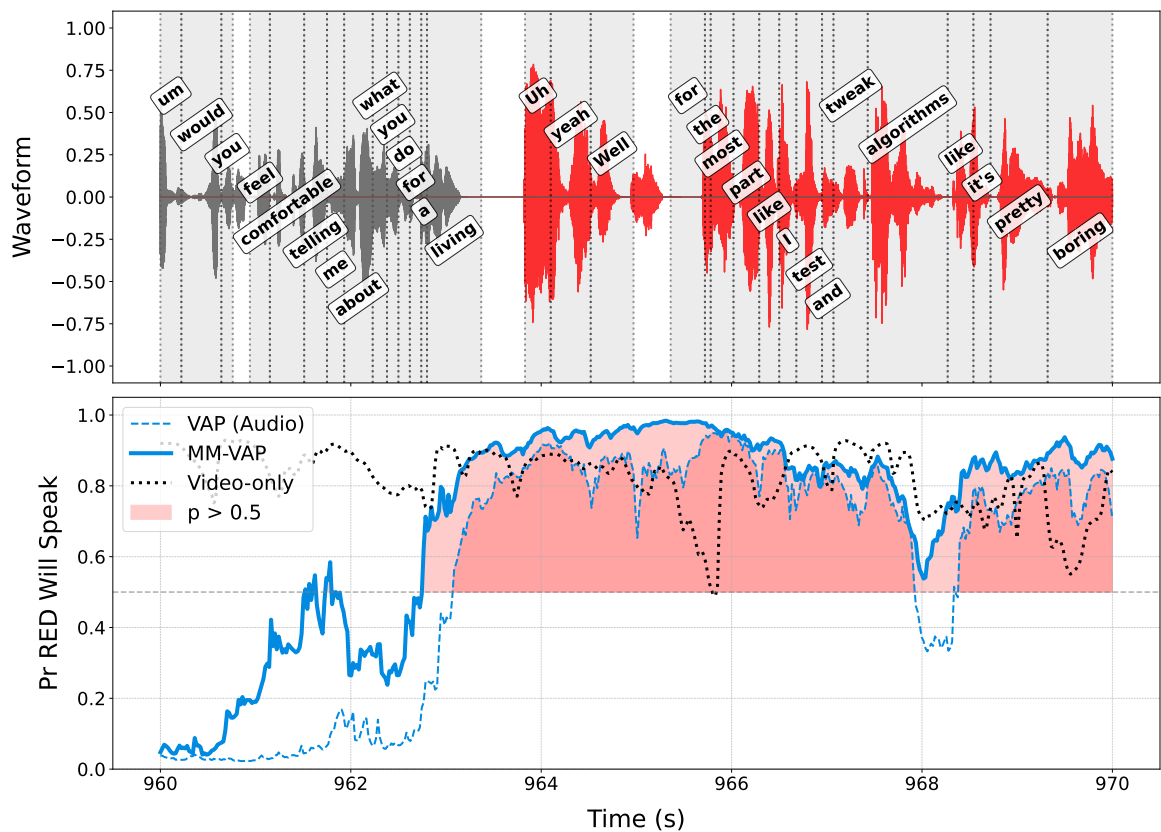
6.6 General conclusion

In this chapter, we have explored the inclusion of the visual modality into turn-taking prediction. In Section 6.1 we conducted an analysis of visual features in the Candor corpus with OpenFace, and found that the next speaker exhibits facial expressions resembling speech prior to the onset of speech. Motivated by these findings, in Section 6.2 we proposed MM-VAP, a multimodal PTTM which utilises video as well as audio cues.

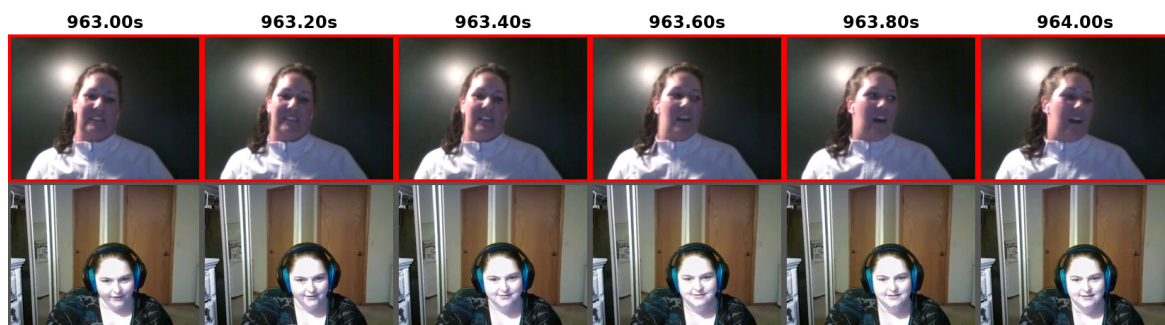
In Section 6.3 we explored the performance of the model, finding improved performance over the audio only state-of-the-art VAP model across a comprehensive set of turn-taking metrics, equivalent to a +3-6% relative increase in balanced accuracy. Performance of a video-only model was significantly worse than both models, demonstrating that both audio and video support turn-taking precondition. In Section 6.4 we explored turn-taking prediction in the models with a qualitative analysis. Finally in Section 6.5, we conducted an exploratory analysis of attention in both VAP and MM-VAP, demonstrating how MM-VAP utilises a richer context across multiple turns. We speculate that this might contribute to its increased performance.

Our work in this chapter therefore shows that visual cues support turn-taking prediction, and therefore addresses Research Question 5 of the thesis. In the following chapter, we build on this finding by exploring the performance of MM-VAP and VAP in noise, to understand if visual cues can assist the model in overcoming the interfering noise.

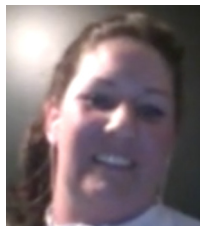
Figure 6.10 Transcription and qualitative comparison MM-VAP, VAP and video-only turn-taking models on session *1f7e582c-c6bc-46b6-b5a4-e5d78e8a46ac* of the Candor corpus, along with an illustration of participant non-verbal behaviour.



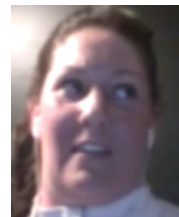
(a) Waveform, transcription and probability trace of model outputs showing the probability that the speaker depicted in red will speak in the next 1.2-2.0 sections (p_{future}).



(b) Frames illustrating participant behaviour immediately prior to and during the mutual silence between speaking turns, with timestamps above in seconds.



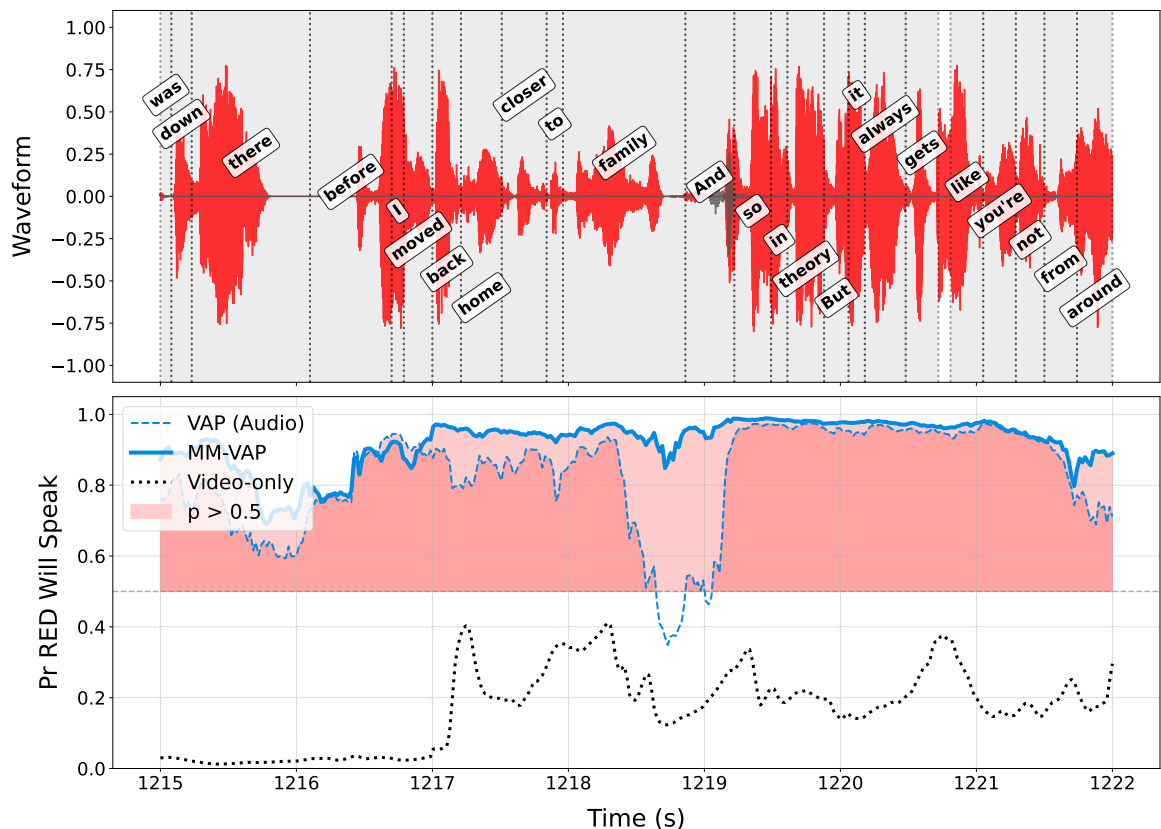
(a) Mouth opening, head tilted. Timestamp = 963.0 seconds



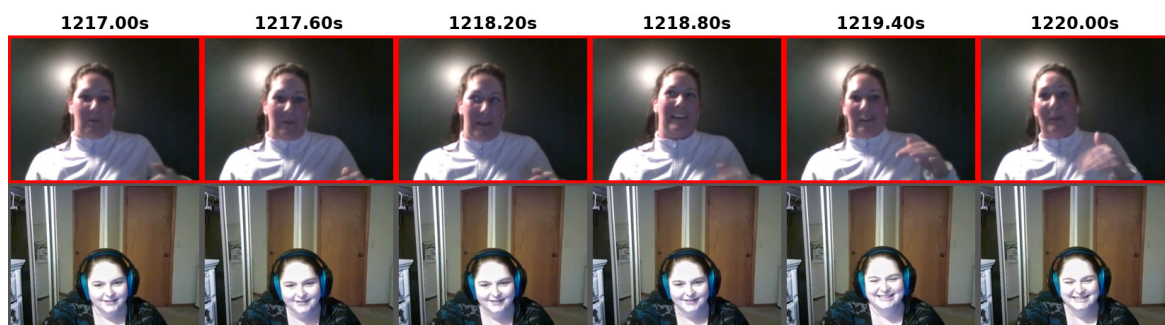
(b) Mouth open, gaze aversion. Timestamp = 963.5 seconds

Figure 6.11 Detail of open mouth and gaze aversion behaviour from the next speaker prior to the onset of speech in session *1f7e582c-c6bc-46b6-b5a4-e5d78e8a46ac* of the Candor corpus.

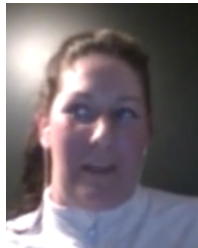
Figure 6.12 Transcription and qualitative comparison MM-VAP, VAP and video-only turn-taking models on session *1f7e582c-c6bc-46b6-b5a4-e5d78e8a46ac* of the Candor corpus, along with an illustration of participant non-verbal behaviour.



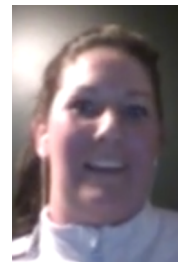
(a) Waveform, transcription and probability trace of model outputs showing the probability that the speaker depicted in red will speak in the next 1.2-2.0 sections (p_{future}).



(b) Frames illustrating participant non-verbal behaviour, with timestamps above in seconds.



(a) Averted gaze



(b) Mouth open, eye contact

Figure 6.13 Detail *1f7e582c-c6bc-46b6-b5a4-e5d78e8a46ac* of the Candor corpus.

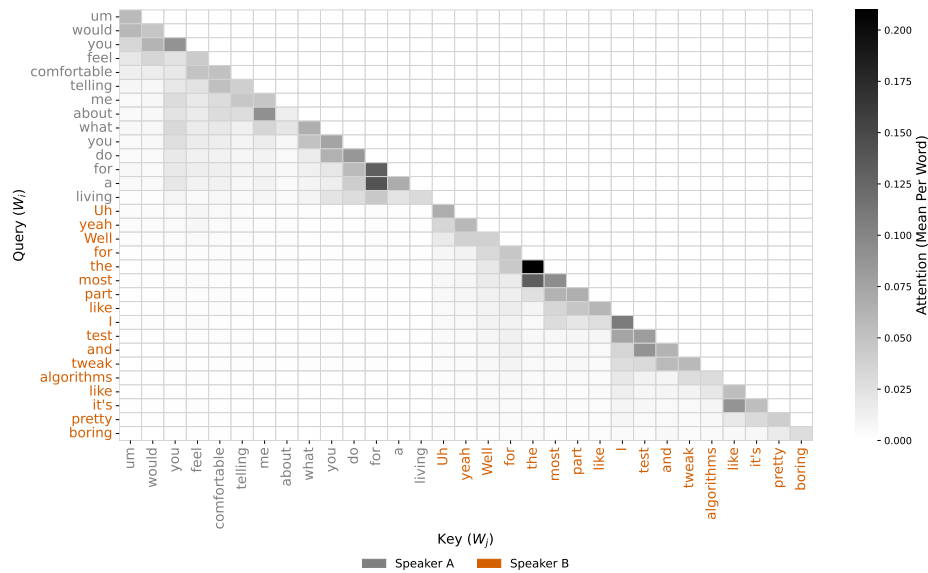


Figure 6.14 **VAP model** attention map. The aggregated self-attention scores Speaker B (red) are shown.

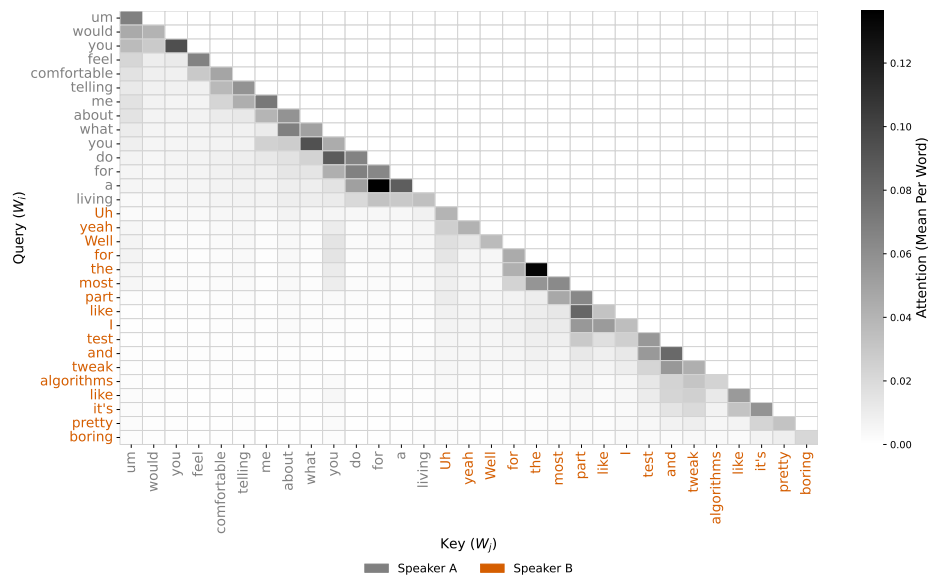


Figure 6.15 **VAP model** attention map. The aggregated cross-attention scores for the attention from Speaker B (red) to Speaker A (grey) is shown.

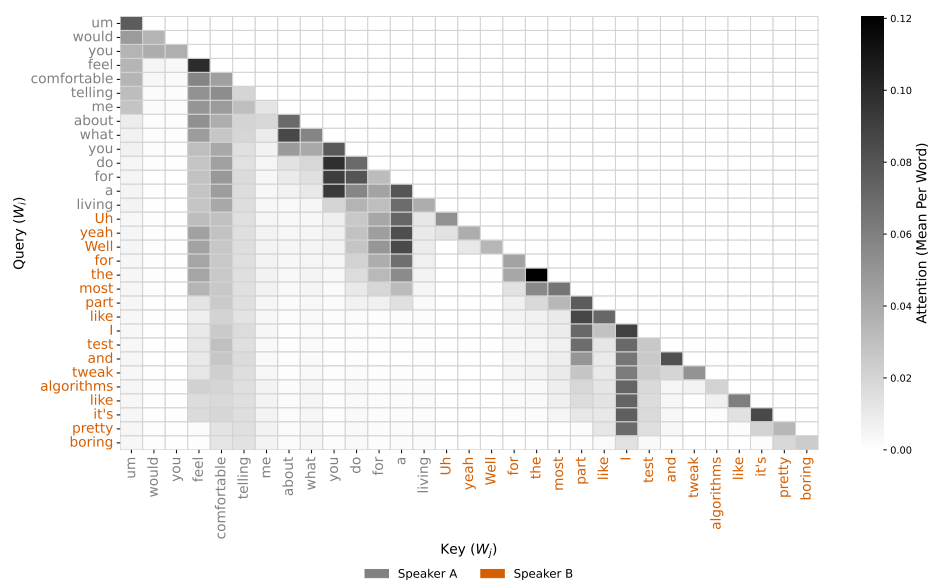


Figure 6.16 **MM-VAP (early modality fusion)** attention map depicting cross-attention from Speaker B Audio $\rightarrow$ Speaker B video cross-attention (layer 0, aggregated by word)

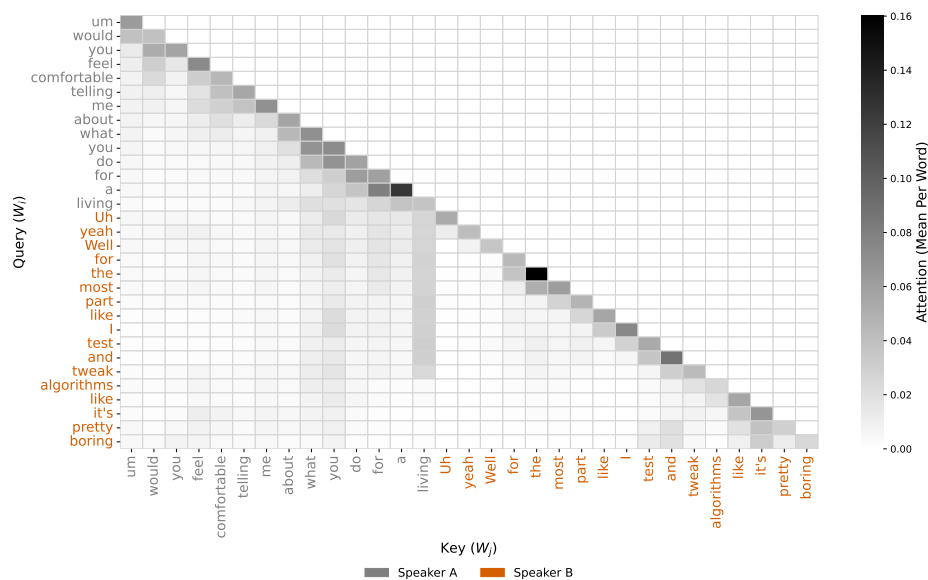


Figure 6.17 **MM-VAP (early modality fusion)** attention map depicting Speaker B $\rightarrow$ Speaker A attention in from final cross-attention layer of the model (aggregated by word)

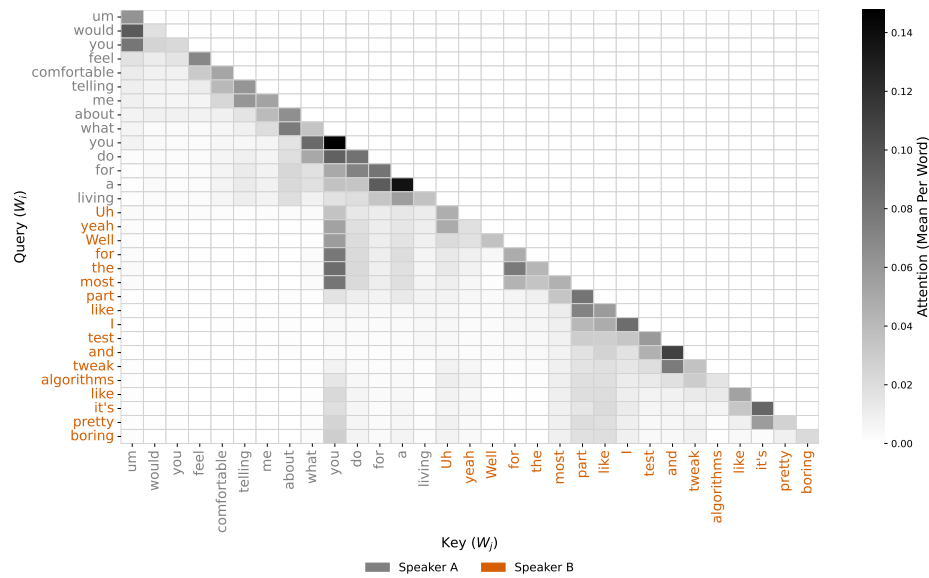


Figure 6.18 **MM-VAP (late modality fusion)** attention map depicting cross-attention from Speaker B video to Speaker A video (layer 3, aggregated by word)

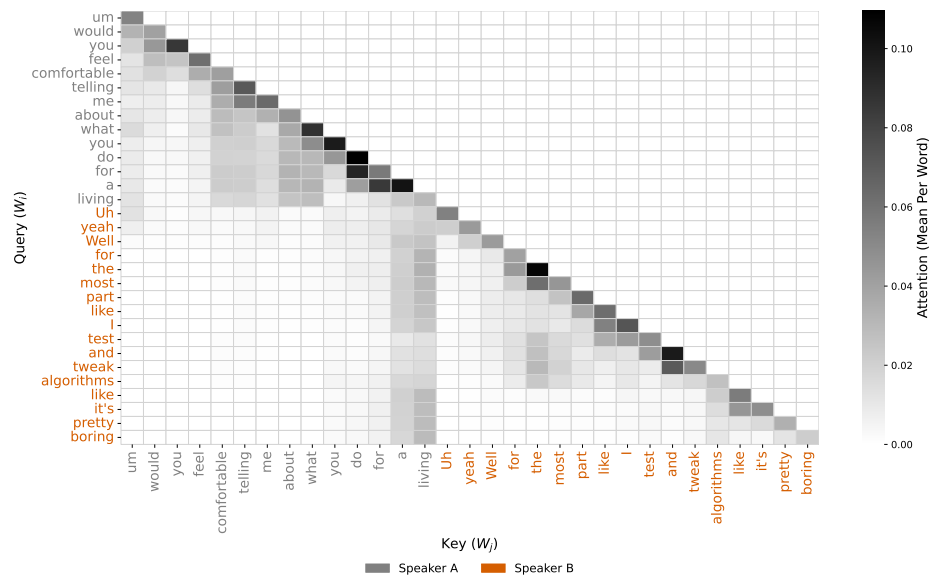


Figure 6.19 **MM-VAP (late modality fusion)** attention map depicting cross-modal Audio→Video attention in from final cross-attention layer of the model (aggregated by word)

Chapter 7

Towards real-world deployment of predictive turn-taking models

In line with prior work this thesis uses corpora to train PTTMs. As corpora were recorded in controlled, noise-free environments, there is little analysis of PTTM performance in real-world conditions. In this chapter, we address this by demonstrating how multimodality makes turn-taking models more robust to acoustic noise and lexical ambiguity.

Our first analysis focusses on PTTM performance in noise. A deployed PTTM in a domestic setting might encounter multiple noises, such as simultaneous conversations, music from a stereo, and so on. Yet to date, PTTMs have only been trained and tested in noise-free interactions (Skantze 2021; Onishi et al. 2024), and the impact of noise on their performance remains unknown. We address this gap by adding environmental noise to recorded interactions at controlled signal-to-noise ratios (SNRs). We then analyse the performance of VAP and MM-VAP models. We find noise degrades the performance of both models. However, the inclusion of visual cues gives MM-VAP a substantial performance advantage, but only when examples of noise are included during training. Thus, the inclusion of visual cues gives our MM-VAP model a major advantage in noise, addressing **Research Question 6** of the thesis.

Notably, the VAP model performs poorly even in conditions where speech remains intelligible despite the interfering noise. We investigate this further. We compare performance in environmental noise and synthetic noise which follows the prosodic contour of the original speech signal. We find performance remains above chance, despite speech becoming progressively unintelligible. This demonstrates that the model has a large degree of redundancy, as it can compensate for the loss of one cue (lexical) by relying on another (prosody). This furthers our understanding of multimodal cues in the model, and reinforces the findings of Chapter 5.

Next, we explore the inclusion of multimodal cues in a large language model (LLM) based turn-taking model. Previously, we found that text alone was insufficient for accurate turn-taking prediction, in large part due to lexical ambiguities (Chapter 5). We therefore propose Audio TurnGPT, an LLM-based turn-taking model which uses audio and text. A major practical challenge as text and audio are highly heterogeneous representations (Baltrušaitis, Ahuja, et al. 2018). We design a cross-attention mechanism which causally aligns text with audio frames. Audio TurnGPT achieves significantly higher accuracy and precision than Pairwise TurnGPT (Leishman et al. 2024), the current text-only state-of-the-art. This demonstrates the role of multimodality in prediction, addressing **Research Question 7a**.

Finally, we discuss a number of outstanding challenges which must be solved in order to deploy the PTTMs from the thesis in human-robot interactions. We conclude with a discussion on how our work advances the state of the art in predictive turn-taking, bringing us closer towards a real-world robust deployment of multimodal predictive turn-taking.

Chapter structure

In Section 7.1, we outline our methodology for adding noise to speech. Section 7.2 presents our analysis of the performance of the VAP model and our MM-VAP model in noise. Section 7.3 considers the role of prosodic and lexical cues in the models. In Section 7.4, we present our Audio TurnGPT model and outline the cross-attention mechanism used to fuse audio and text modalities. Finally, Section 7.5 discusses the necessary steps in making the models in this thesis work in real-world deployments, and highlights remaining unsolved issues.

Research questions addressed

The work in this chapter addresses the following research questions:

- **Research Question 6** *How is the performance of PTTM impacted by noise?*
- **Research Question 7b** *Can PTTM utilise multimodal cues in the same way as humans?; For instance, by using visual cues to overcome noise*

Chapter contributions

The work in Sections 7.1 and 7.2 is the basis for our INTERSPEECH publication, and the work in 7.3 is included in our ICASSP publication.

- O'Connor Russell, Sam and Harte, Naomi “*Visual Cues Support Robust Turn-taking Prediction in Noise*”. Proc. Interspeech 2025, 1073-1077

- O'Connor Russell, Sam and Delphine Charuau and Naomi Harte. "*The Role of Prosodic and Lexical Cues in Turn-Taking with Self-Supervised Speech Representations*" ICASSP 2026, Barcelona, Spain

7.1 A framework for exploring predictive turn-taking in noise

PTTMs must be able to maintain robust performance in environmental noise: in a real-world setting, music may be playing or people talking in the same room. However, an analysis of PTTM performance in noise has not been considered in the literature. We therefore conduct a thorough evaluation of PTTM performance in noise. Our method involves adding noise to speech at controlled SNRs and evaluating model performance. We first outline our methodology, prior to considering the performance of the audio-only VAP and our multimodal MM-VAP model in noise. We use the Candor corpus for all training and analysis.

7.1.1 Synthesising environmental noise

We generate three types of environmental noise: babble, music and speech. An overview is provided in Table 7.1. These noises are representative of the types of interference a PTTM is likely to encounter in-the-wild, once deployed.

Babble noise We synthesise babble noise to simulate multiple interfering conversations. We use the LRS3 corpus (Afouras et al. 2018) to source samples of speech (118,516 train, 1,452 test). We select 30 utterances at random that are at least 15 seconds in length. At each frame i , the babble noise is an average of the frames from the 30 utterances:

$$x_{babble,i} = \frac{1}{30} \sum_{k=0}^{30} x_{i,k} \quad (7.1)$$

Where $x_{i,k}$ is the i^{th} frame from utterance k . We generate 65,000 unique 30 speaker babbles. We follow the LRS-3 train and test partitions, generating 46 hours of training noise and 14 hours of test noise recordings.

Music noise Next, we select music noise to simulate acoustic interference in the same room as the PTTM. We source samples of music from the MUSAN corpus (Snyder et al. 2015). There are multiple genres of music included in the corpus (classic, jazz, rock, etc.). We use the *Jamendo* genre for training (15 hours) and the remaining genres for testing (70 hours).

Speech noise Finally, we simulate interference from other speakers (speech noise) by overlaying another speaker from the Candor corpus. We remove silences from the recording of each interlocutor to generate speech noise samples. During cross-validation and testing of models, we ensure that the separation of train, test and validation partitions of the corpus

Table 7.1 Types of additive acoustic noise and descriptions

Category	Noise	Description
Environmental	Babble	65,000 30-speaker babbles from LRS-3 (Afouras et al. 2018)
	Music	Samples from the MUSAN corpus (Snyder et al. 2015)
	Speech	Random utterances from the Candor corpus, overlaid
Prosodic Manipulation (Section 5.2.3)	Flat F0	Original speech with flat F0 contour
	Flat Intensity	Original speech with flat intensity contour
	Flat F0 + Intensity	Original speech with intensity + F0 contours
Prosodic Noise (Section 5.2.3)	Average spectrogram (utterance-level)	Flat F0, intensity contour preserved
	Average spectrogram (utterance-level)	F0 contour preserved, flat intensity
	Average spectrogram (utterance-level)	Flat F0 and intensity contours
	Average spectrogram (utterance-level)	

is maintained when adding noise by selecting speakers from other sessions within the same partition of the corpus.

7.1.2 Prosody-matched noise and manipulation

We compare PTTM performance in environmental noise with performance in acoustic signals generated with our vocoder-based method from in Chapter 5. We generate prosodic noise and prosodic manipulation signals using the method described in Section 5.2.3. For simplicity, we collectively refer to these signals as *noises*.

A complete overview of the generated noises is provided in Table 7.1.

7.1.3 Controlling the strength of additive noise

After generating the noises, we then add them to the Candor corpus at controlled SNRs. We add the noises separately to each channel. We follow the standard method from the automatic speech recognition (ASR) literature (Shi et al. 2022). An important distinction is that in prior work, noise was typically added to isolated speech segments. However, we use the Candor corpus, which consists of conversational speech and therefore includes long pauses. As the standard method of adding controlled noise to conversational speech has not been reported, we add noise at a given SNR by considering signal energy during speaking time as follows.

We first generate a noise track $n(t)$ that is the same length as a full recording of a speaker from the Candor corpus, denoted $s(t)$. To match the length of the environmental noise tracks to the recording, we concatenate noise samples sampled at random. Prosodic noise and manipulation tracks are already equal to the length of the recordings, as they are manipulated versions of the original recordings. We then remove all silences from the speech ($\tilde{s}(t)$) using

the time-aligned ASR transcription of the Candor corpus. The RMS energy of the noise $n(t)$ and the that of speech signal with silences removed $\tilde{s}(t)$ are:

$$\text{RMS}_{\tilde{s}} = \sqrt{\frac{1}{N} \sum_{t=1}^N \tilde{s}(t)^2} \quad \text{RMS}_n = \sqrt{\frac{1}{M} \sum_{t=1}^M n(t)^2} \quad (7.2)$$

We then compute a scaling factor, α , used to add noise to the original speech signal at a target SNR in the range of -10 to $+10$ dB in 2.5 dB increments:

$$\alpha = \frac{\text{RMS}_{\tilde{s}}}{\text{RMS}_n 10^{\text{SNR}/20}} \quad (7.3)$$

We then scale the noise and add it to the signal:

$$s'(t) = s(t) + \alpha n(t) \quad (7.4)$$

This results in a version of the original signal corrupted with noise at the given target SNR. We then peak-normalise the output to avoid clipping. Natural variation in amplitude across utterances means this yields an average SNR over all utterances in a recording.

7.1.4 Evaluating PTM performance in noise

To assess turn-taking prediction with the VAP and MM-VAP models, we generate test sets from the Candor corpus, using the same train and test split as in previous Chapters. We add each category of noise/manipulation in Table 7.1 to the Candor test set at -10 to $+10$ dB SNRs in 2.5 dB increments. The resulting test sets are overviewed in Table 7.2.

For all evaluations, we use the shift/hold prediction metric (during mutual silence) previously outlined in Section 4.2.2 of Chapter 4. For clarity, we consider only the balanced accuracy score in our evaluations in this chapter. All tests are conducted using the paired t-test to compare the 5-fold average balanced accuracy on the common held-out test set.

7.1.5 Augmenting the training set with noisy data

VAP and MM-VAPs models were not exposed to any noise during training. We therefore retrain the models on a mixture of clean speech and speech corrupted with noise, a process which we refer to as augmentation. We follow Shi et al. (2022) and add 0 dB noise to 25% of sessions in the training set. We train with the same 5-fold cross-validation procedure outlined in Chapter 4. The models trained on these augmented training sets are listed in Table 7.3.

Table 7.2 Test sets generated by adding noise to the Candor corpus and descriptions

Test set name	Test set description
Original test set	Original test set consisting of 70 sessions
Environmental (babble)	Original test set + babble noise
Environmental (music)	Original test set + music noise
Environmental (speech)	Original test set + speech noise
Prosodic noise (F0)	Original test set + prosody-matched noise preserving F_0
Prosodic noise (intensity)	Original test set + prosody-matched noise preserving intensity
Prosodic noise (F0+intensity)	Original test set + prosody-matched noise preserving both F_0 and intensity
Prosodic manipulation (F0)	Original test set + prosodic manipulation, flattened F_0
Prosodic manipulation (intensity)	Original test set + prosodic manipulation, flattened intensity
Prosodic manipulation (F0+intensity)	Original test set + prosodic manipulation, flattened F_0 and intensity

Table 7.3 Overview of the VAP and MM-VAP model Candor corpus training conditions

Condition name	Model	Trained on	Training data detail
clean	VAP	Clean speech	clean speech (i.e. the model from Chapter 4)
augmented	VAP	Clean speech + environmental noise	75% clean speech + 25% 0 dB babble, music or speech noise
	VAP	Clean speech + prosody-matched noise	75% clean speech + 25% 0 dB noise matched to F_0 and intensity
	VAP	Clean speech + prosodic manipulation	75% clean speech + 25% 0 dB speech with flattened F_0 and intensity
clean	MMVAP	Clean speech	clean speech (i.e. the model from Chapter 6)
augmented	MMVAP	Clean speech + environmental noise	75% clean speech + 25% 0 dB babble, music or speech noise

7.1.6 Impact of the noise on the WER

We then evaluate the effect of each noise at each SNR on the intelligibility of speech, using ASR performance as a proxy for intelligibility as described in Chapter 5. We transcribe the original test set and the test sets with -10 to $+10$ dB SNR noise with Whisper¹ (Radford, Kim, et al. 2023). We compute and report the word error rate (WER) at each SNR. Following Chapter 3, we consider an ASR transcription to contain useful information if its word error rate (WER) is below 30%. As a rough indicator for intelligibility, we also use this 30% threshold.

7.2 Turn-taking prediction in environmental noise

7.2.1 The impact of environmental noise on PTM performance

We first consider the impact of environmental noise on turn-taking prediction. The first two rows in Table 7.4 show the performance of both VAP and MM-VAP models at each SNR. The bar chart in Figure 7.1 shows the average shift/hold prediction accuracy on the test data corrupted with environmental noise, where solid colours are VAP (audio-only) and hatched are MM-VAP (audio and video). The performance in noise-free conditions from Chapter 6 is shown in green.

The plot in Figure 7.1 reveals that environmental noise has a strong impact on PTM accuracy. This is most severe in speech and music noise, where the accuracy falls to 51–52% at 10 dB SNR, i.e. a random hold/shift prediction. There is a smaller impact on

¹We use Whisper for this analysis as it is free-to-use ASR and our access to commercial ASR is limited

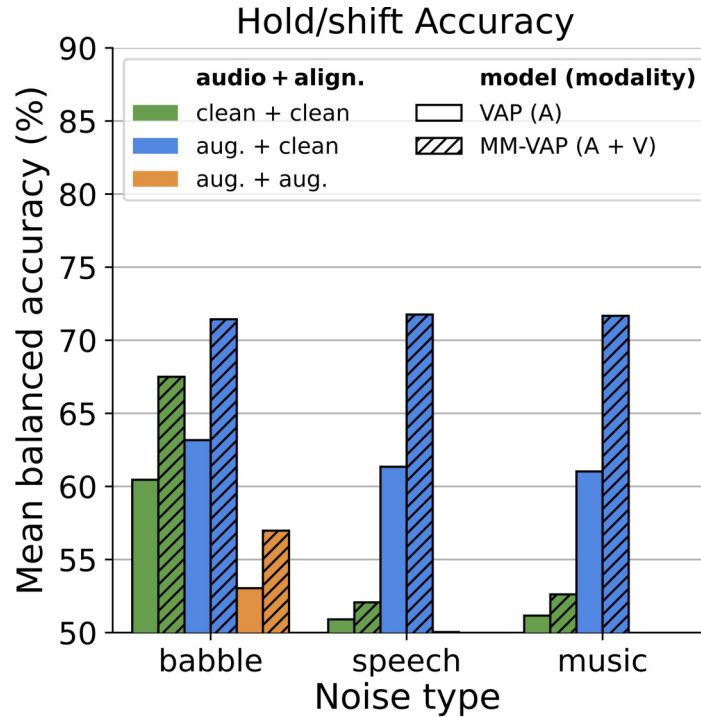


Figure 7.1 Average hold/shift prediction accuracy of VAP (audio-only) and MM-VAP (audio+video) models in each noise type across the -10 dB, +10 dB SNR range. **Audio** refers to the training condition, detailed in Table 7.3. The clean training set contains no examples of noise. **Alignment (align)** refers to whether the word alignment is sourced from ASR transcriptions of the clean or augmented (aug.) training set.

performance from babble noise, with 67% balanced accuracy for MM-VAP, compared with 60% for VAP. Figure 7.2 shows that above 0 dB SNR, the average accuracy increases as the SNR increases. Below 0 dB SNR, accuracy levels off to 50–54%. MM-VAP shows a slightly improved performance over VAP, equivalent to an SNR gain of MM-VAP over VAP of +3.75 dB SNR. However, a breakdown of average accuracy by noise type reveals the superior performance of MM-VAP originates from accuracy in babble noise, as accuracy in music and speech is always 51–54% (Table 7.4).

The results therefore show that neither VAP nor MM-VAP is robust to environmental noise interference. Moreover, the poor performance of the MM-VAP model demonstrates that visual cues do not lead to any meaningful performance increase over VAP in noise.

7.2.2 Augmenting the training set with noisy examples

We then consider the performance of both models when a portion of the training data is augmented with environmental noise, using the procedure outlined in Section 7.3. The

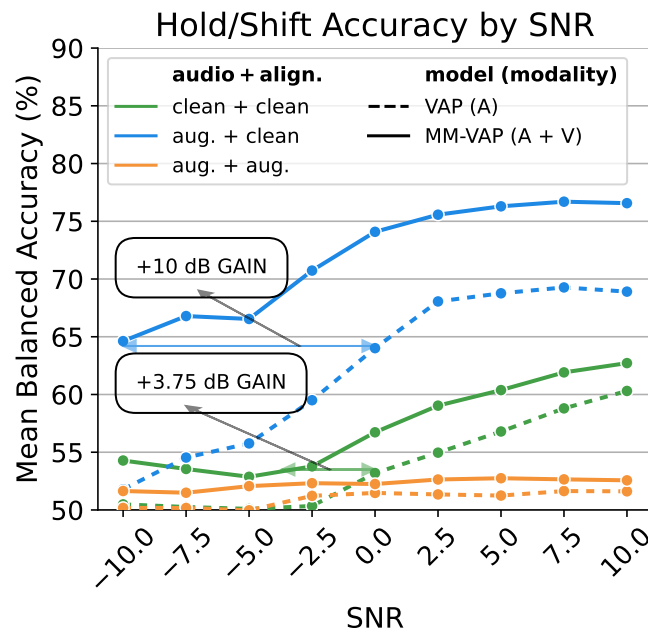


Figure 7.2 Average hold/shift prediction accuracy of VAP (audio-only) and MM-VAP (audio+video) models at each SNR. An average accuracy in speech, music and babble noise is shown. **Audio** refers to the training condition, which is either clean or augmented (aug.) as detailed in Table 7.3, where the clean training set contains no examples of noise.

second two rows in Table 7.4 show the performance of both re-trained models. Shift/hold prediction accuracy of the re-trained models is shown in blue in Figure 7.1.

Comparing blue and green bars in Figure 7.1 shows re-training has led to a clear performance increase, e.g. average accuracy in speech is 61% for VAP and 72% for MM-VAP, up from 51–52% ($p < 0.01$). Not only has performance increased, but a large gap in the accuracy of VAP and MM-VAP models has also emerged. For instance, in music noise previously performance of both models was low (52–52%), but now MM-VAP clearly outperforms VAP by a wide margin (72% vs 61%, $p < 0.01$). The blue lines in Figure 7.2 show the performance breakdown at each SNR. At -10 dB SNR, the average accuracy of VAP is 50% whereas MM-VAP is 65% ($p < 0.01$). At +10 dB SNR of speech noise, the average accuracy is 64% for VAP and 75% for MM-VAP (second two rows of Table 7.4), up from 51% and 52% previously ($p < 0.01$). At -10 dB SNR (representing an extreme level of interference), VAP has a 50% accuracy. On the other hand, MM-VAP backs off to the accuracy of a PTTM model trained exclusively on video features (67% vs 66% in -10 dB SNR music noise).

The results show that MM-VAP exploits visual information to improve accuracy in acoustic interference. However, to achieve this performance, it is necessary to augment the training data with examples of noisy speech, as without this step, MM-VAP accuracy in noise is low (54% in green in Figure 7.2). This shows that the MM-VAP model does not have an inherent ability to exploit visual cues to overcome noise: training is necessary. Importantly,

Table 7.4 Average hold/shift prediction accuracy (%; $\bar{x}$ =mean all SNRs) across 5-folds of cross-validation of VAP (audio, A), MM-VAP (audio+video, A+V), and video-only (V) models on the same test set at varying noise levels (2.5 dB increments, 5 dB shown for brevity). Training is conducted on the clean and augmented (aug., 25% of sessions corrupted with 0 dB noise) audio and ASR alignments.

Model	Training		Input Modality	Test set + babble @ SNR dB =						Test set + speech @ SNR dB =						Test set + music @ SNR dB =						Clean
	Audio	Alignment		-10	-5	0	5	10	$\bar{x}$	-10	-5	0	5	10	$\bar{x}$	-10	-5	0	5	10	$\bar{x}$	
VAP	clean	clean	A	61	50	58	69	76	60	51	51	51	51	52	51	51	50	51	52	51	79	
MM-VAP	clean	clean	A+V	57	56	68	76	80	68	53	52	52	53	54	52	53	51	51	52	53	52	83
VAP	aug.	clean	A	49	50	65	74	77	63	51	59	65	66	64	61	55	58	62	66	65	61	84
MM-VAP	aug.	clean	A+V	61	62	75	79	80	71	66	69	74	74	75	72	67	68	74	76	74	72	86
VAP	aug.	aug.	A	50	50	55	55	56	53	50	50	50	49	49	50	50	50	50	50	50	50	57
MM-VAP	aug.	aug.	A+V	55	56	57	58	58	57	50	50	50	50	49	50	50	50	50	50	50	50	57
video-only	—	clean	V	66						66						66						66

as re-training the audio-only VAP model yielded limited performance benefit, the VAP model has a very limited ability to ‘learn’ to separate speech from noise. This means visual cues must be included in order to achieve robust performance in noise, in conjunction with careful training on the model on examples of noisy speech.

7.2.3 The role of the word alignment in training

Previously, the word alignment for training the models was sourced from ASR transcriptions of clean, noise-free speech. Noise was then added into the recording post-transcription. However, in a real-world setting, the interactions would be recorded in noisy environments, and only noisy recordings would be available for transcription. To assess whether this impacts performance, we re-transcribe the noisy training sessions with Amazon ASR². The bottom two rows in Table 7.4 show the results of the analysis, and the orange bars in Figure 7.1 reveal a stark fall in the accuracy of VAP and MM-VAP models to 50-58%. Figure 7.2 shows consistently poor performance across all SNRs. PTTM training is therefore dependent on accurate transcriptions, with either manual (Ekstedt and Skantze 2022b) or obtained from ASR and clean recordings (Chapter 4).

7.2.4 Predicting a shift between speakers

We show an example of the models predicting a shift between speakers in different environmental noise conditions in Figure 7.3, where MM-VAP solid and VAP dashed. The p_{future} probability is the probability that the speaker depicted in the red waveform (A) will speak within the next 0.8-2 seconds, with the region greater than 0.5 highlighted red.

In clean audio (B), the shift between speakers is predicted ahead of time by MM-VAP. In 0 dB babble noise (C), MM-VAP predicts the shift, whereas VAP always outputs approx 0.5,

²Elsewhere in the thesis, we use Speechmatics transcription for PTTM training, but this work uses Amazon ASR due to constraints imposed by our academic access to these ASR platforms

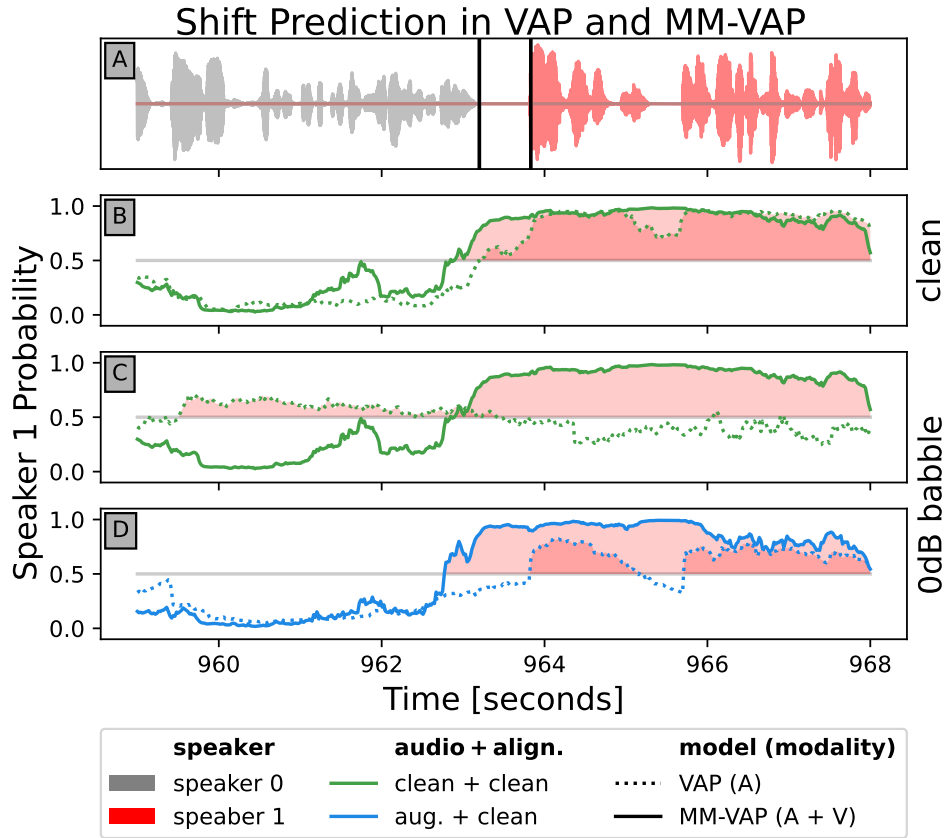


Figure 7.3 Model output during a shift in the Candor corpus (unseen during training). Speaker 0 (grey): "So uhm would you feel comfortable telling me about what you do for a living?", Speaker 1 (red) "Uhm yeah well for the most part I test and tweak algorithms". VAP and MM-VAP model output is shown in clean speech (B) and 0 dB babble speech (C) and (D). The audio is either clean speech or augmented with 0 dB noise. A=audio, V=video.

demonstrating the impact of noise. MM-VAP predictions in noise (C) resemble predictions in clean audio (B), further demonstrating the exploitation of visual cues by MM-VAP. Re-training MM-VAP on data augmented with 0 dB noise (D) makes the prediction more confident, highlighting how the inclusion of noise during training improves exploitation of visual cues. Importantly, comparing VAP predictions in (C) and (D) shows that re-training the audio-only model on noise does not improve prediction significantly. This shows the audio-only VAP model does not learn the characteristics of the noise; visual cues must be included in the model for robust prediction to be achieved.

7.2.5 Generalisation to unseen noises

Previously, in Section 7.2.2, the training set contained a mixture of the three environmental noise types (Table 7.3). In order for a PTM to have robust performance, it should be able to perform well across different types of noise that were unseen during training (generalisation).

Table 7.5 Shift/hold prediction accuracy (%) of MM-VAP trained with speech, babble, music, or a mixture (augmented) at 0 dB SNR.

		Test set, 0 dB SNR			
		<i>babble</i>	<i>music</i>	<i>speech</i>	<i>clean</i>
Train set	<i>babble</i>	71	51	52	85
	<i>music</i>	58	75	57	86
	<i>speech</i>	70	77	76	86
	<i>augmented</i>	75	74	74	86

Table 7.6 Average shift prediction accuracy (%; $\bar{x}$ =mean all SNRs) across 5-folds of cross-validation of VAP (audio, A), MM-VAP (audio+video, A+V), and video-only (V) models on the same test set at varying noise levels (2.5 dB increments, 5 dB shown for brevity). Training is conducted on the clean and augmented (aug., 25% of sessions corrupted with 0 dB noise) audio and ASR alignments.

Model	Training		Input Modality	Test set + babble @ SNR dB =							Test set + speech @ SNR dB =							Test set + music @ SNR dB =							Clean
	Audio	Alignment		-10	-5	0	5	10	$\bar{x}$	-10	-5	0	5	10	$\bar{x}$	-10	-5	0	5	10	$\bar{x}$	∞			
VAP	clean	clean	A	50	50	59	72	79	62	50	50	50	50	52	51	50	50	50	51	52	51	85			
MM-VAP	clean	clean	A+V	50	50	69	72	79	64	50	50	50	51	52	51	50	50	50	51	52	51	87			
VAP	aug.	clean	A	50	50	73	79	79	66	50	63	67	68	66	63	51	62	69	70	67	64	86			
MM-VAP	aug.	clean	A+V	50	50	73	80	80	66	50	67	74	74	76	68	50	67	74	77	75	68	86			
VAP	aug.	aug.	A	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	56			
MM-VAP	aug.	aug.	A+V	51	51	52	52	52	52	50	50	50	50	50	50	50	50	50	50	50	50	57			
video-only	—	clean	V	72							72							72							72

We assess generalisation by retraining MM-VAP three times with **either** 0 dB SNR speech, music **or** babble noise in 25% of training sessions³. The balanced accuracy of shift/hold prediction in Table 7.5 shows that when the training set includes speech or babble noise, accuracy is poor in other types of noise: e.g. 71% for the babble-trained model in babble noise but only 54% in music. However, when trained on speech noise, similar accuracy is obtained in all types of noise 74-75% (Table 7.5, second-last row). The model’s ability to exploit visual cues to overcome noise therefore shows limited ability to generalise to types of noise not encountered during training.

7.2.6 Impact on shift prediction

For completeness, we also report the results of the analysis with the shift prediction metric from Chapter 4. This compares the ability of a turn-taking model to accurately distinguish between mid-turn speech and speech at the end of a turn, before a shift. The results in Table 7.6 show a similar trend to the shift/hold prediction accuracy, which we explored in detail.

7.2.7 Discussion

Our analysis comprises what is to our knowledge the first exploration of PTTM performance in noise. Our results revealed that PTTMs do not maintain high accuracy in the types of noises

³We revert to the original word alignment derived for this analysis

they are likely to encounter in-the-wild, with performance falling from 83% to 54% balanced accuracy in +10 dB speech noise (Section 7.2.1). However, we did find that including noise during training dramatically increased performance to 75% balanced accuracy in +10 dB speech noise compared with 64% balanced accuracy audio-only VAP model at the same noise level (Section 7.2.2). Augmenting training data with noise is a common technique in audiovisual speech recognition (AVSR) (Shi et al. 2022) that has not previously been considered in turn-taking. Our findings therefore show that the PTTMs in this work are not robust to noise without the prior inclusion of noise in training, addressing Research Question 6 of this thesis.

Training on data augmented with 0 dB SNR noise led to considerable improvements in the accuracy of MM-VAP across the full SNR range. At extreme levels of noise interference (e.g. -10 dB music noise) VAP fails (50%, i.e. chance) but MM-VAP backs off to the accuracy of the video-only model (67%). However, we found the improved performance of MM-VAP does not necessarily generalise to new sources of noise, except for the model trained with speech noise (Section 7.2.5). Our working hypothesis is that mixing in random speakers over the speech signal training process ‘forces’ MM-VAP to use visual information, rather than simply relying on the acoustic properties of the noise. These findings therefore address Research Question 7b of this thesis: PTTMs can achieve robust performance in noise via the inclusion of visual cues, but this is intimately related to the training process and does not inherently generalise to new noise types.

We acknowledge some limitations of our work. Firstly, we artificially add noise to speech recorded in noise-free settings. Although this is commonplace (Marxer et al. 2018; Shi et al. 2022; Lin and Harte 2025), it means we do not capture changes to verbal and non-verbal (J. Trujillo et al. 2021) communication and turn-taking (Sørensen et al. 2020) in noise. A further limitation is that our mixing of noise at a constant energy level means the audio-only VAP model may be partly capitalising on the natural variations in SNR in long recordings of speech. As previously noted, another limitation is our use of videoconferencing. However, no in-person corpora match the size and quality of Candor. In future work, training techniques for multimodal PTTMs should be further explored to ensure the exploitation of visual cues generalises to new noise types. As AVSR can exploit visual information in datasets unseen in training, models and training techniques from the ASVR literature forming a natural starting point for this work (Sterpu, Saam, et al. 2020). There are many other turn-taking events (e.g. backchannels) which require future consideration.

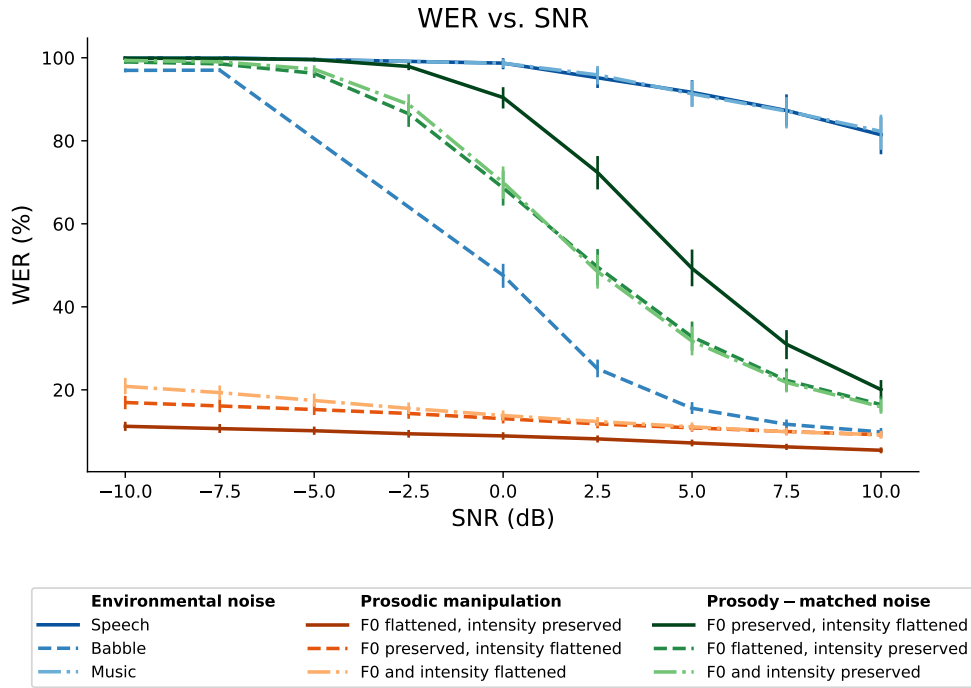


Figure 7.4 Average speaker-level word error rate (WER) for each manipulation (clamped to 100%) at each signal-to-noise ratio (SNR) in decibels (dB). P=pitch, I=intensity. Mean per speaker $\pm 95\%$ c.i. Following our work in Chapter 3, we consider 30% WER as a threshold below which the transcription contains useful information for word recognition.

7.3 Uncovering the role of acoustic and lexical cues in turn-taking prediction

In the previous section, we found background noise severely impacts turn-taking prediction, even when the noise was relatively weak in comparison to the signal, e.g. +10 dB SNR. The speech should remain intelligible here; so why does the performance fall to such an extent? In this section, we explore this, delving deeper into the role of prosodic and lexical cues in the VAP model. The work in this section therefore expands on our findings from Chapter 5.

7.3.1 The impact of noise on speech intelligibility

We begin by considering how each noise impacts the WER, a common proxy for intelligibility in the ASR literature (Choi et al. 2023), as described in Section 7.1.6. The WER in environmental noise, prosodic noise, and prosodic manipulation at each SNR is shown in Figure 7.4 (the maximum WER is clamped to 100%).

Environmental noise The blue lines in Figure 7.4 show the impact of environmental noise on the WER. Speech and music noise have a high WER even in light interference, e.g. $\approx 85\%$ in $+10$ dB SNR. Below 0 dB SNR, the WER exceeds 100%, showing little useful phonetic information is present. Babble noise has a much lower impact on the WER at high SNR. The WER remains below 30% for SNR > 2.5 dB, before rising above our 30% threshold, beyond which no useful phonetic information is retained (Section 7.1.6).

Environmental noise therefore removes useful phonetic information for word recognition, degrading lexical content. The level of the degradation is higher as the strength of the noise increases. Informal listening reveals that speech remains intelligible at SNR > 2.5 dB SNR music and speech noise, despite the WER exceeding 80%. Although informal, this highlights that ASR accuracy tracks, but does not completely align with human intelligibility. We also note that adding environmental noise to speech disrupts the prosodic contour (F_0 and intensity), as an unrelated recording of environmental noise is added to the speech signal.

Prosodic noise The green lines in Figure 7.4 show that when we add prosodic noise, the WER progressively increases. Light prosodic noise in the $+10$ to $+7.5$ dB range has minimal impact on the WER. However, this quickly rises to a very high $> 40\%$ WER at $+2.5$ dB SNR, before saturating at $> 100\%$ WER. As the structured prosodic noise follows either the F_0 contour, the intensity contour, or both, the analysis shows we can progressively remove the lexical information from speech whilst preserving the F_0 and/or intensity contour of the original speech signal.

Prosodic manipulation Finally, the orange lines in Figure 7.4 show that when we add prosodic manipulation to speech, the WER is minimally impacted. At -10 dB SNR, which represents a very high interference, the WER is elevated but remains low $\approx 10\text{--}20\%$. The increase in the WER likely originates in distortion, as the vocoder used to synthesise speech with flattened prosody has a 10 ms analysis window. We speculate that Whisper ASR is sensitive to these distortions. As informal listening reveals speech remains readily intelligible, the analysis shows we can manipulate the prosody of speech, whilst (mostly) preserving lexical information.

Summary The types of noise described here are summarised in Table 7.7. The noises are grouped by their preservation or removal of the lexical content, and F_0 and intensity contours.

Table 7.7 Types of additive acoustic disruptions used in this chapter, grouped by their preservation of the lexical content, $F0$ and intensity contours ($\checkmark$ =preserved, x =removed)

Type of disruption	Lexical	$F0$	Intensity
<i>prosodic noise</i>	x	x	$\checkmark$
	x	$\checkmark$	x
	x	$\checkmark$	$\checkmark$
<i>prosodic manipulation</i>	$\checkmark$	x	$\checkmark$
	$\checkmark$	$\checkmark$	x
	$\checkmark$	x	x
<i>environmental noise</i>	x	$\checkmark$	x

7.3.2 Exploring turn-taking prediction performance in noise

We then investigate the performance of the VAP model in each type of noise. As the WER analysis shows that we can separately disrupt intelligibility and the prosodic contour of speech, our method enables us to explore how prosodic and lexical cues contribute to PTM performance. The strength of the disruption is governed by the SNR of the additive noise.

Performance in prosodic noise and manipulation The balanced accuracy of the VAP model on the shift/hold and shift prediction metric is shown in Figure 7.5a. The model was trained on clean speech and tested under the described noises, summarised in Table 7.7. The orange lines show the impact of prosodic manipulation, which we found alters prosody and mostly preserves lexical content and hence, intelligibility. In +10 dB of prosodic manipulation, the balanced accuracy ranges from 77-81%; close to the 79% balanced accuracy achieved on clean speech, revealing minimal to no disruption. Next, we consider the lines in green in Figure 7.5a which show the impact of prosodic noise, which we found disrupts intelligibility whilst preserving $F0$ and/or intensity. The green lines show +10 dB of prosodic noise results in a slightly larger drop to 72-76%. This manipulation preserves $F0$ and/or intensity, whilst disrupting intelligibility.

Performance in environmental noise On the other hand, the blue lines in Figure 7.5a show the impact of environmental noise on shift/hold prediction. In +10 dB of music and speech noise, the balanced accuracy falls to 51-52% (chance accuracy). As noted earlier in Section 7.2.1, the models are more robust to the impacts of babble noise. We speculate that this is because the CPC audio front-end of the models may have been trained with some data augmentation from babble noise, though this is not explicitly stated (Riviere et al. 2020).

Increasing the manipulation strength The findings therefore reveal the model is robust to light (+10 dB) disruption of either prosody (prosodic manipulation); lexical cues (prosodic noise); but not both (speech and music noise). As the SNR decreases, we see this trend

continue. We observe a much smoother roll off in balanced accuracy for prosodic manipulation and prosodic noise (orange and green) than for environmental noise (blue). At -10 dB, prosodic and environmental noise, the speech has almost totally been replaced by unintelligible noise. However, in prosodic noise, there is a 67% balanced accuracy. This contrasts with environmental noise, where performance is chance at $\approx 50\%$. Therefore, even though both prosodic noise and environmental noise make speech progressively less intelligible as the SNR decreases (Figure 7.3.1), turn-taking performance is far less impacted in prosodic noise, where the prosody is preserved. On the other hand, prosodic manipulation achieves a $\approx 76\%$ accuracy in -10 dB noise, indicating minimal performance deterioration.

This analysis therefore demonstrates that prosodic cues can be exploited to achieve performance where lexical cues are removed and vice-versa although accuracy does fall substantially. On the other hand, performance in environmental noise, which disrupts both cues, is poor. The results therefore show that either the prosody or lexical cues must be preserved to support above-chance turn-taking prediction. A notable finding is that the model was not trained on any noisy data, but unintelligible prosodic noise matching the prosodic contour supports above chance performance. We discuss this finding further on, in Section 7.3.3.

Re-training the model on noisy data Next, we re-train the VAP and MM-VAP models with examples of noise in their training data (full list of models is Table 7.3). Figure 7.5b shows the balanced accuracy of shift/hold prediction. Performance in prosodic noise (green) rises slightly, though the trend remains largely the same as before. A similar trend is noted for prosodic manipulation (orange). This demonstrates that the training process enables models to better leverage prosodic and lexical cues in speech, achieving better accuracy.

For environmental noise (blue), retraining the models has had a slight improvement in performance, but performance remains low. As environmental noise disrupts intelligibility and disrupts the prosodic contour of speech, turn-taking prediction performance falls to chance in high noise levels (< -2.5 dB). The re-training has therefore not led to a meaningful impact on performance under high levels of noise. This demonstrates that either prosodic or lexical cues must be preserved in order to maintain robust performance, and re-training does not enable the model to overcome the disruption of environmental noise at low SNRs.

Performance on the shift prediction metric Finally, for completeness, we repeat the analysis on the shift prediction metric from Chapter 4. The results are shown in Figures 7.5c and 7.5d. These results show similar performance trends to the shift/hold prediction metric.

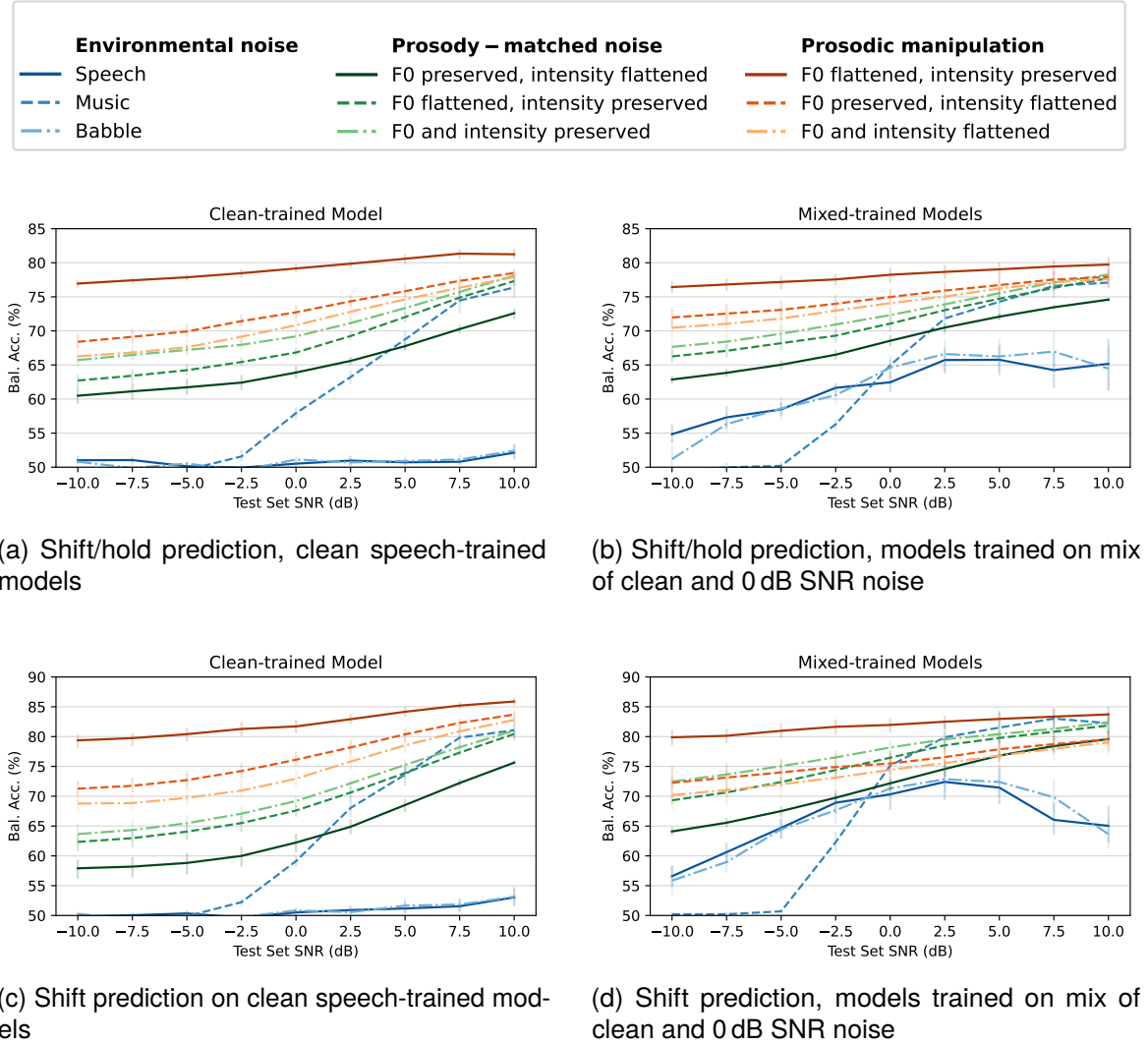


Figure 7.5 Balanced accuracy of MM-VAP and VAP models under different environmental noises, prosodic noises, and prosodic manipulation. The noises and manipulations are applied to the test set at different signal-to-noise ratios (SNRs), ranging from -10 to +10 dB. The average test-set accuracy over 5 folds of cross-validation is shown, $\pm 95\%$ c.i. The metrics are shift/hold prediction (during mutual silence) and shift prediction (before the end of a turn). Models are either trained on clean speech, or a mixture of clean and 25% (see Table 7.3). Models are always trained and deployed on the same type of manipulation (e.g. clean speech + prosodic noise is tested on prosodic noise at different SNRs).

7.3.3 Discussion

In this section of the thesis, we probed the role of prosody and lexical cues in turn-taking prediction by selectively removing them from speech. We used our method of generating synthetic signals from Chapter 5, which selectively preserved/degraded certain cues in speech. We then added them to the original, unmodified speech signal at controlled SNRs, and compared their impact on performance to that of environmental noises such as music.

We found that additive environmental noise and prosodic noise lead to a high WER, indicating speech was unintelligible. However, prosodic noise preserves the $F0$ and intensity contour. Additive prosodic manipulation, in contrast, largely preserved lexical content, with WER staying below 30%, while flattening both $F0$ and intensity. This enabled us to analyse the impact of the removal of cues from speech on turn-taking prediction with the VAP model. We found that the removal of prosodic cues (prosodic manipulation) has a less severe impact than the removal of lexical information (prosodic noise). This demonstrates that lexical cues are the dominant turn-taking cue in the model, but either prosodic or lexical cues can support turn-taking prediction, as one can be removed, whilst the model flexibly uses intact cues. If both are removed, as is the case with environmental noise, performance falls to chance accuracy. Concerning the removal of individual prosodic cues, removal of the intensity contour has a greater impact on turn-taking prediction than removal of $F0$ (solid and dashed green lines in Figure 7.5), suggesting intensity is the dominant turn-taking cue.

The work in this section has therefore expanded upon our exploration of turn-taking in noise in Section 7.2. These findings are consistent with our results in Chapter 5, but the comparison with environmental noise and across a range of SNR shown here deepens these earlier findings. As discussed in Chapter 5, a limitation of the work is that our vocoder methodology does not fully isolate lexical cues as syllable rate, vowel duration, etc. are preserved in prosodic noise and may hence be preserved and exploited by the model for prediction.

7.4 Adding speech cues to an LLM-based turn-taking model

The previous section illustrates how the VAP model can rely on prosody and lexical cues: when either becomes unavailable, the model can exploit the other, without re-training. This means the model is robust as there are multiple cues that can be exploited for prediction when some are removed, such as lexical cues, $F0$ and so-on. This is in contrast with our findings in Chapter 5, where we explored LLMs-based turn-taking models. We found a large number of false positive end-of-turn predictions, as the model has no cue beyond text to rely upon when it encounters instances of syntactic ambiguity.

In this section, we explore the incorporation of another modality, audio, into an LLM based turn-taking model. Our motivation is that in human turn-taking, cues from audio such as *F0* are exploited where syntax alone is unreliable (Section 2.2.2). We therefore investigate whether an LLM may similarly exploit audio by proposing Audio-TurnGPT, an LLM turn-taking model which includes audio. This work builds on the work of Section 7.2, where the inclusion of the visual modality led to increased robustness in acoustic noise.

7.4.1 Integrating audio cues into an LLM-based turn-taking model

We propose the inclusion of audio cues into an LLM-based turn-taking model by extracting features from audio using layers from VAP model, and aligning text with audio using transformer cross-attention. We call the model Audio-TurnGPT. Like the text-only Pairwise TurnGPT model from Leishman et al. (2024) which we explored in Chapter 4, our proposed model is an LLM. However, it additionally contains information from the audio channel. As aligning two modalities with different frame rates and timescales is highly challenging, we design a masked cross-modal attention framework for the model that ensures that text aligns correctly with frames of audio while maintaining the causality of the model.

Section 7.4.2 outlines the cross-modal attention mechanism. In Section 7.4.3 we outline the proposed Audio-TurnGPT model architecture, and in Section 7.4.5, we review its performance compared with a text-only baseline. For consistency with the work in Chapter 5, the Switchboard corpus is used for this work.

7.4.2 Aligning modalities with masked cross-model attention

In the MM-VAP model, we utilised the cross-attention mechanism within the transformer (Vaswani 2017) architecture to combine audio and video modalities. We use the same mechanism to incorporate audio cues into an LLM-based turn-taking model in a similar manner. However, a key difference arises from the challenge posed by the different timescales and dimensionalities of the audio and text modalities.

The audio features are 265-dimensional feature vectors, with one feature at a 50 Hz rate. On the other hand, text is modelled using GPT-2 (Radford, Wu, et al. 2019). Text features are 786-dimensional feature vectors, and a single token can span hundreds of milliseconds of audio. This means there is a one-to-many mapping from tokens to audio features. Moreover, different words have different lengths, making some tokens align with more / less audio frames than other tokens. Furthermore, as the model is an LLM, a prediction of the next token is made each time an input token is received (and not at 50 Hz, like the VAP and MM-VAP models). A simple attention mask cannot capture these dependencies.

To align audio and text, we also use cross-attention. There are multiple possibilities. One option to align tokens with audio is to allow each token to attend to all audio frames in the

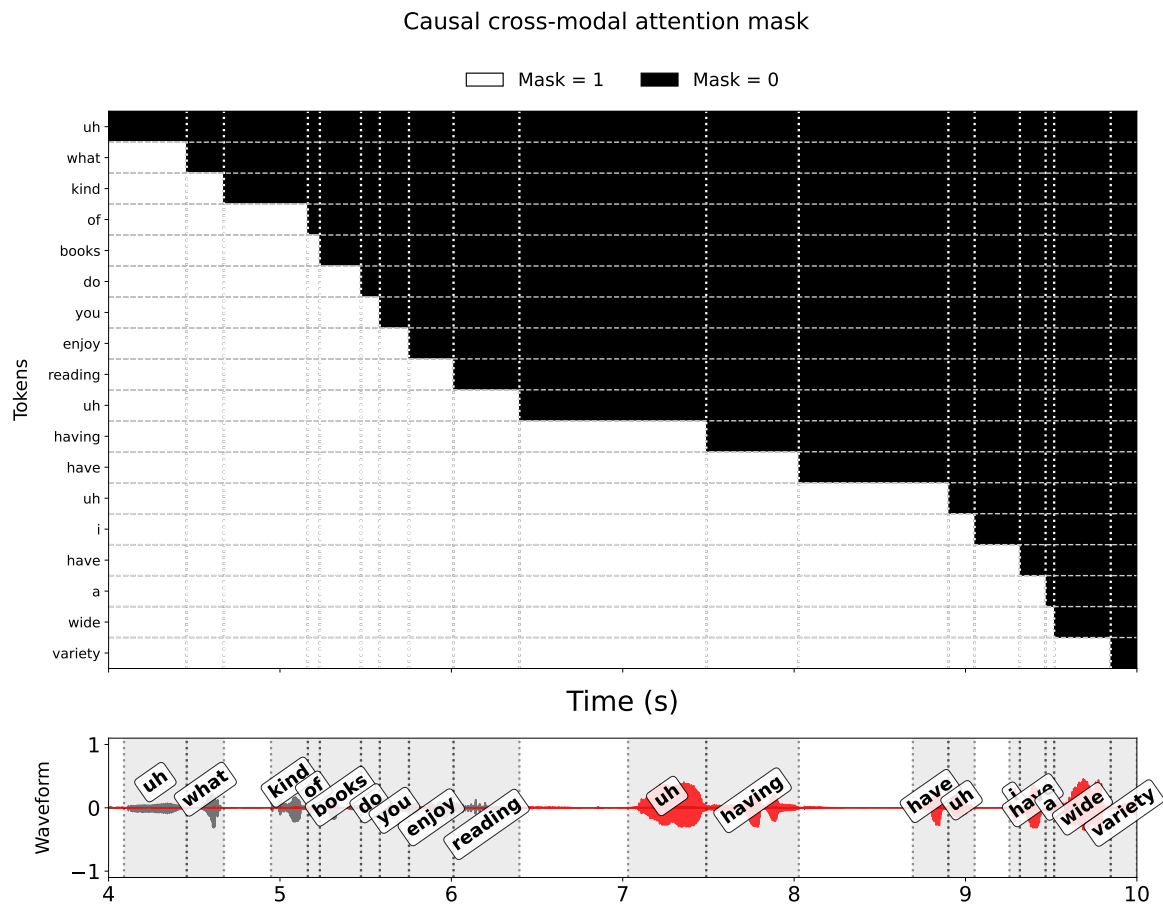


Figure 7.6 An example showing a causal, cross-modal attention mask used to align audio (50 Hz) with text in our Audio-TurnGPT model. Tokens on the y axis are aligned with audio on the x axis. Tokens can only attend to audio frames from previous tokens.

current token. However, this would mean that a word has immediate access to all audio frames as soon as it occurs. We therefore ensure that text tokens can only attend to audio frames from previous tokens in the sequence, and not the current token being fed to the model. We mask the cross-attention layers in the model to achieve this, and we visualise the attention mask in Figure 7.6 for an example from the Switchboard corpus. Black indicates the attention is masked to 0. Tokens are on the y axis, and the audio is shown on the x axis. The staircase-like pattern reflects the many-to-one mapping from tokens to audio frames. Note that for each token, the mask is only non-zero for audio frames occurring during the previous token in the sequence.

7.4.3 The Audio-TurnGPT Model

Having devised a method to align tokens with audio frames, we now outline the proposed Audio-TurnGPT Model. The model follows the same objective as Pairwise TurnGPT (Leish-

man et al. 2024), which we explored and modified in Chapter 5. It continually predicts the next token for two speakers in a dyad, with special tokens inserted for turn-taking events, e.g. end-of-turn. However, it also contains information from the audio channel.

The Audio-TurnGPT model is shown in Figure 7.7. The model consists of a pre-trained Pairwise TurnGPT⁴ model⁵ and a pre-trained VAP model with the final prediction layers removed. Both models are pre-trained⁶ on Switchboard, as described in Chapter 4. We use pre-trained layers from the VAP model as our logic is that these layers have learnt useful information relevant for turn-taking from the audio channel. The audio from both speakers is fed into these layers, which are frozen during all subsequent training.

The audio block is followed by a single hidden-layer neural network, which projects the audio from 256 to 768 dimensions in order to match the dimensionality of the Pairwise TurnGPT model (hidden layer dimensionality 512). Audio and text streams are then combined with a 2-layer cross-attention block where query=tokens and k=v=audio. We apply the masking procedure from Section 7.4.2 to align tokens with audio frames. This is followed by an LLM prediction layer and a softmax function, which outputs a probability vector over all possible words in the vocabulary (50,256 + 5 special tokens, Chapter 5).

7.4.4 Training the Audio-TurnGPT model

For training, we batch the text into chunks of 256 tokens, sampled at random. We combine text with a corresponding 20-second chunk⁷ of audio, where the last frame of audio aligns with the start time of the last token in the sequence. We train with a batch size of 20, which is the maximum in our Nvidia RTX 6000 GPU. We train the model with the same method as Pairwise TurnGPT models, outlined in Section 5.4.4 of Chapter 5.

For inference, we discard the first 240 predictions in a batch as only the final tokens are aligned with audio. We then report performance on special tokens <EOT>, <PAUSE>, <YIELD>, and <EINT> that are inserted into the dialogue at key turn-taking events (Leishman et al. 2024). These tokens and their definitions are outlined in Section 5.4.3 of Chapter 5.

7.4.5 Predictive turn-taking with Audio-TurnGPT

The test set performance on special tokens of both models on the Switchboard corpus is shown in Tables 7.8 and 7.9. We first focus on the performance when the validation set balanced accuracy is maximised, shown in Table 7.8. For end-of-turn (<EOT>) prediction, Pairwise TurnGPT achieves a balanced accuracy of 85%, compared with Audio-TurnGPT,

⁴We use the version of Pairwise TurnGPT that has our proposed weight initialisation and masked cross-attention modifications, discussed in Section 5.4.7.

⁵We use our weight initialisation and masked cross-attention modifications described in Chapter 4

⁶We tried training these layers from scratch, but convergence was challenging to achieve

⁷Due to GPU memory constraints, we could not align each of the 256 tokens with the corresponding audio

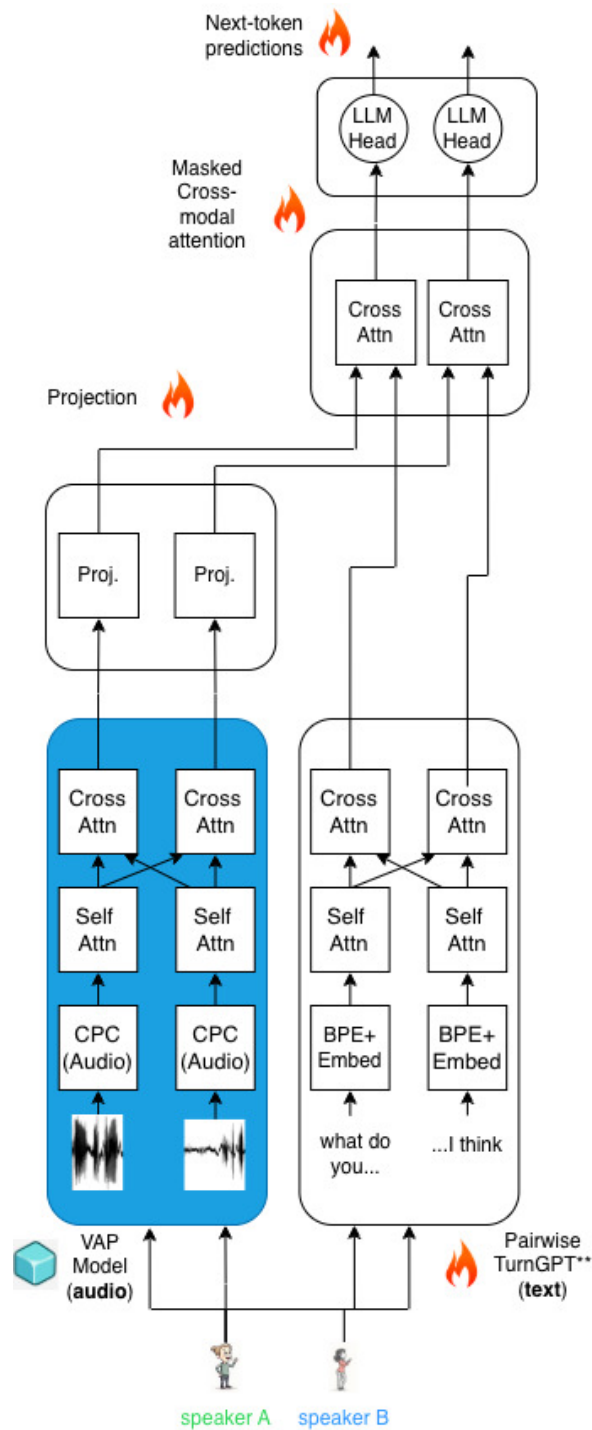


Figure 7.7 Architecture of the Audio-TurnGPT model. Pre-trained weights are loaded from the VAP model ((Ekstedt and Skantze 2022b), Chapter 4) and the Pairwise TurnGPT ((Leishman et al. 2024), Chapter 5) models. The VAP model layers are frozen during training, whereas the Pairwise TurnGPT layers are not. The audio features are projected (proj.) to match the text feature dimensionality via a single hidden-layer neural network. Text and audio modalities are combined via masked cross attention (Section 7.4.2) prior to prediction. The next token in the sequence is predicted via an LLM head (linear layer + softmax, vocabulary size=50,256 + 5 special turn-taking tokens).

which achieves an accuracy of 90%. This difference is significant as per a paired t-test ($t(4) = 23.19, p < 0.001$). Inspection of metrics reveals that this originates in a substantial reduction of false negatives, from 1,206 to 715. The number of true positives is also up, increasing from 4,913 to 5,428. The overall precision remains low, but the increase from 0.12 to 0.14 is significant ($t(4) = 4.64, p = 0.0097$). The performance of Audio-TurnGPT on the other special tokens <PAUSE>, <YIELD>, and <EINT> is also increased over that of Pairwise TurnGPT, with paired t-tests confirming $p < 0.001$ in each case.

This demonstrates that Audio TurnGPT achieves a higher prediction performance than TurnGPT. In each case, the increased accuracy stems from Audio-TurnGPT giving an increase in true positive predictions and a decrease in false negative predictions. This is also the case when we report the performance by maximising the validation set F1 score, shown in Table 7.9. As discussed in Section 5.4.5 of Chapter 5, this method gives a more balanced performance between minority (special token) and majority (all other tokens in the vocabulary) classes. We again note Audio-TurnGPT performance is increased over that of Pairwise TurnGPT, with paired t-tests confirming $p < 0.001$. Here, the performance increase stems from a decrease in false positives and an increase in true positive predictions. However, we note that the overall precision and sensitivity of both models are relatively low (e.g. Audio-TurnGPT, <EOT> token, sensitivity=0.48 and precision=0.39).

7.4.6 Comparing the predictions of Audio-TurnGPT with Pairwise TurnGPT

We show an example of the predictions of special tokens with Audio TurnGPT in Figure 7.8. The example is the same as previously explored in Section 5.10. Despite the significant increase in balanced accuracy, our analysis here notes no qualitative improvement of Audio TurnGPT over Pairwise TurnGPT. In the example, we note that there are false positive pause predictions in yellow, and an end-to-turn token is falsely predicted at the end of the example in Figure 7.8.

We show a further example in Figure 7.9. We first compare the Pairwise TurnGPT and Audio TurnGPT predictions in Figures 7.9a and 7.9b. Both models predict the next token of both speakers in the dyad, but during this excerpt, only one speaker is active so only this speaker is shown for simplicity. We note that Pairwise TurnGPT and Audio TurnGPT predictions are broadly similar. Both models correctly predict the end-of-turn token, and do not make false-positive end-of-turn predictions (in Audio TurnGPT the probability is close to the threshold, but just below it). The VAP predictions are shown in Figure 7.9c. The end of turn is not predicted by the VAP model, however, the probability of speech from speaker 1 remains correctly low throughout the excerpt, indicating a speaker change is not predicted.

Table 7.8 Performance metrics for special token prediction with Pairwise TurnGPT (Weight init + Mask C/a modifications) and the proposed Audio TurnGPT model. A 5-fold average test set performance is shown. Thresholds chosen to maximise validation set balanced accuracy. TP=True positive, FN=False negative.

Model+Token	TN	FP	FN	TP	Sens.	Spec.	Prec.	F1	Bal. Acc. (%)
<EOT>									
Audio-TurnGPT	347,216	34,787	715	5,428	0.88	0.91	0.14	0.94	90
Pairwise TurnGPT (modified)	341,689	39,090	1,206	4,913	0.80	0.90	0.12	0.93	85
<PAUSE>									
Audio-TurnGPT	267,072	111,404	697	8,973	0.93	0.70	0.08	0.81	82
Pairwise TurnGPT (modified)	266,548	110,708	1,202	8,440	0.88	0.71	0.07	0.81	79
<YIELD>									
Audio-TurnGPT	369,585	16,540	397	1,624	0.80	0.96	0.09	0.97	88
Pairwise TurnGPT (modified)	367,904	16,977	569	1,448	0.72	0.95	0.08	0.97	84
<EINT>									
Audio-TurnGPT	378,853	7,497	78	1,718	0.96	0.98	0.19	0.99	97
Pairwise TurnGPT (modified)	377,120	7,983	289	1,506	0.84	0.98	0.16	0.99	91

Table 7.9 Performance metrics for special token prediction with Pairwise TurnGPT (Weight init + Mask C/a modifications) and the proposed Audio-TurnGPT model. A 5-fold average test set performance is shown. Thresholds chosen to maximise validation set F1 score. TP=True positive, FN=False negative.

Model+Token	TN	FP	FN	TP	Sens.	Spec.	Prec.	F1	Bal. Acc. (%)
<EOT>									
Audio-TurnGPT	377,361	4,642	3,169	2,974	0.48	0.99	0.39	0.98	74
Pairwise TurnGPT (modified)	375,611	5,168	3,509	2,610	0.42	0.99	0.34	0.98	71
<PAUSE>									
Audio-TurnGPT	369,114	9,362	6,609	3,061	0.32	0.98	0.25	0.96	65
Pairwise TurnGPT (modified)	367,108	10,148	6,806	2,836	0.30	0.97	0.22	0.96	63
<YIELD>									
Audio-TurnGPT	384,441	1,684	1,267	754	0.37	1.00	0.31	0.99	68
Pairwise TurnGPT (modified)	383,132	1,749	1,387	630	0.31	1.00	0.27	0.99	65
<EINT>									
Audio-TurnGPT	385,373	977	637	1,159	0.64	1.00	0.54	1.00	82
Pairwise TurnGPT (modified)	383,954	1,149	819	976	0.54	1.00	0.46	1.00	77

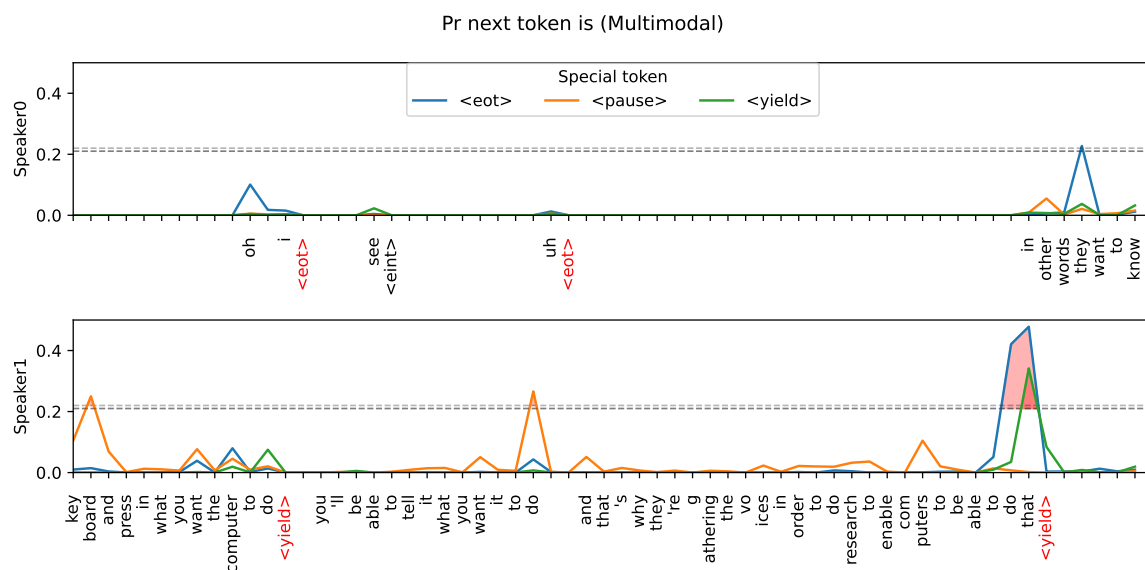
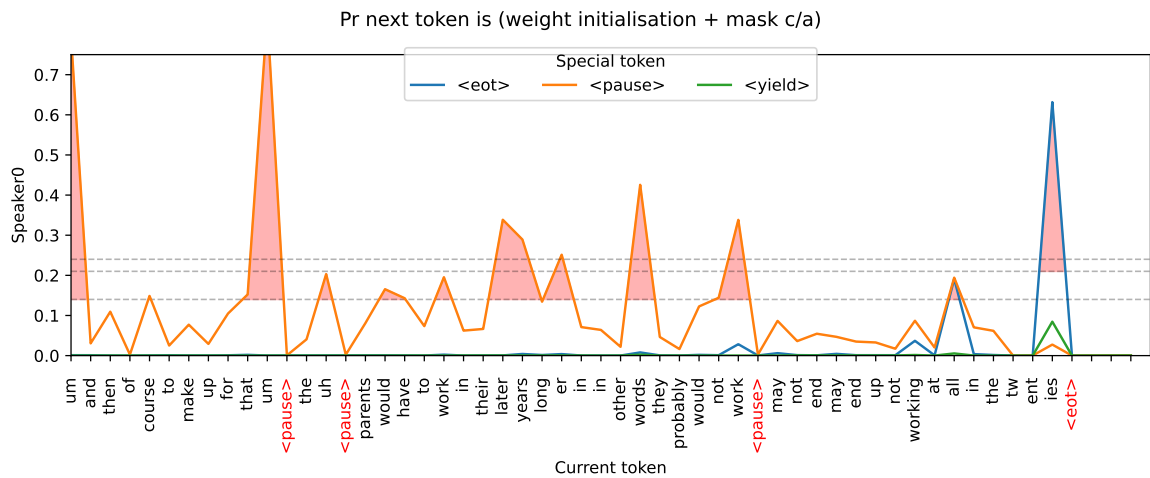
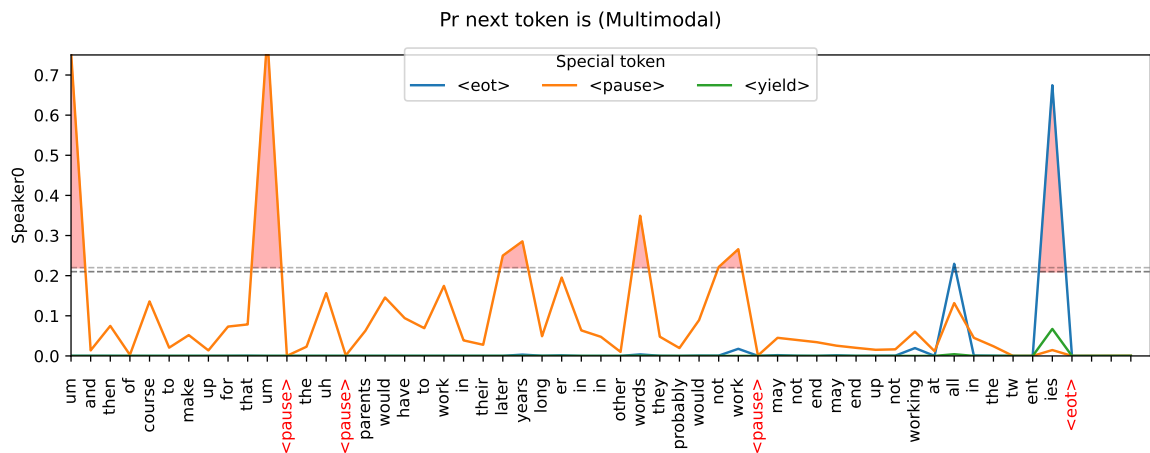


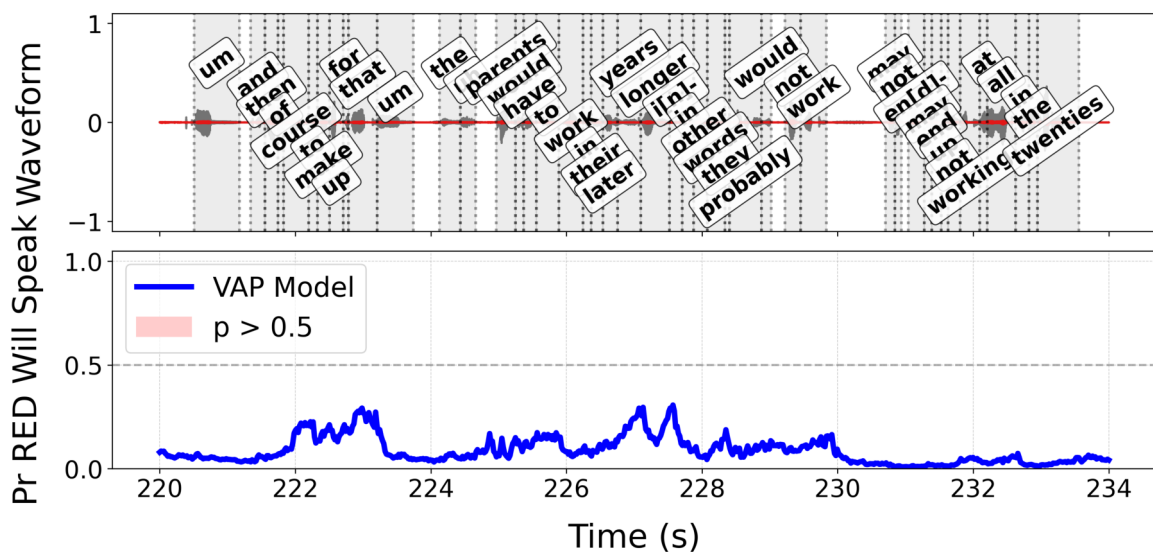
Figure 7.8 Predictions of Audio-TurnGPT on the Switchboard corpus. The excerpt is from session *sw2238* from the held-out test set. Thresholds for positive prediction are optimised on the validation set, with token-specific thresholds applied. The regions where the probability exceeds the threshold are highlighted in red.



(a) Predictions for the text-only Pairwise TurnGPT model (Leishman et al. 2024), with our proposed modifications, explored in Chapter 5



(b) Predictions from our audio + text model, Audio-GPT



(c) Predictions from the audio-only VAP model (Ekstedt and Skantze 2022b), explored in Chapter 4

Figure 7.9 Comparing predictions of the text-only PairwiseGPT model, with our proposed text+audio Audio-TurnGPT model, and the audio-only VAP model. The example is from session sw4311 of the Switchboard corpus, unseen by all models during training.

7.4.7 Discussion

In this section, we have explored incorporating audio features into a LLM-based turn-taking model. This was based on our earlier analysis of LLM-based turn-taking models from the literature in Chapter 5, which found low precision originating in a large number of false positive end-turn predictions. We also noted that audio turn-taking models did not suffer the same errors. We therefore proposed to add an audio front-end into an LLM-based turn-taking model, which we called Audio TurnGPT.

At the corpus level, we found that the inclusion of audio features in our proposed model, Audio TurnGPT, led to a modest but statistically significant increase in performance in all prediction metrics. A possible explanation for the small increase is that, because the overall precision of Audio TurnGPT remained low, improvement was on a small minority of turn-taking events in the corpus. Alternatively, the low precision of the model may indicate it is not appropriately leveraging information from the audio channel.

In Chapter 5, we concluded that audio-only turn-taking models did not suffer the same syntactic ambiguity as text-only models. We noted this addresses Research Question 7a, as audio overcomes the lexical ambiguity. However, we found here that including audio cues into a text-based turn-taking model did not lead to qualitative improvements in performance. This could be for a number of reasons. Firstly, the use of multimodality in conversation is highly complex: sometimes one modality is sufficient / conveys the same information as the other modalities, other times both modalities are required for comprehension (Holler and Levinson 2019). This has been noted in human behavioural studies, which reached different conclusions on the necessity of prosody in turn-taking (J.-P. De Ruiter et al. 2006; Bögels and Torreira 2015). Future work should explore the role of prosody in the model further on specific examples from the corpora, perhaps deliberately constructed to contain ambiguity, as has been previously studied in human behavioural studies (Bögels and Torreira 2015).

Alternatively, it may be the case that limitations in the training process or architecture of our model mean that Audio TurnGPT does not fully exploit the audio channel. For instance, text may be the dominant modality in the model, with audio being useful for prediction on a minority of events. The loss function or multimodal fusion within the model may not adequately capture this. Future work should therefore explore better training methods and techniques for incorporating information from the audio modality. For instance, this could take the form of a learnable confidence score or weighting between audio and text modalities, which increases the importance of the audio modality where it is useful for prediction.

Our model has a causal alignment from text to audio, as it waits for entire words to be completed before alignment with audio. However, in practice, a further short delay between audio and text modalities is typically included to account for the latencies associated with real-time ASR (Roddy and Harte 2020). We do not include this delay for simplicity, assuming each word is available to the model as soon as it finishes. Our alignment is approximate as it

uses word boundaries, although we note there is a phoneme alignment with the Switchboard corpus. We could therefore make the cross-modal attention more granular at the phoneme level. However, such an alignment would be very impractical in a deployed system as real-time ASR does not provide a phonetic alignment, which is often completed after transcription, e.g. with forced alignment (McAuliffe et al. 2017).

7.5 Deployment of turn-taking models to real-world interaction

This thesis explores a range of predictive turn-taking models. We developed the models on pre-recorded, transcribed corpora of real, human-human interaction. In this section, we offer a discussion on the feasibility of the deployment models explored in this thesis to real-world human-robot interaction. We highlight the necessary steps and outstanding issues surrounding model development. Finally, we conclude with a discussion on how the work of this thesis brings turn-taking closer to successful deployment. This section motivates our discussion of future work in Chapter 8.

7.5.1 Deployment of models from the literature to human-robot interaction

In this thesis, we explored turn-taking models from the literature, including the VAP model (Ekstedt and Skantze 2022b) and the TurnGPT model (Ekstedt and Skantze 2020). As we have noted, both models are causal. In theory this makes them ideally suited to a real-world deployment, as there is no dependency on information from future time steps for prediction. However, in a real-world deployment, speech recognition and dialogue management add considerable latency, in addition to the latency of the models themselves.

These issues were explored in a recent work by Skantze and Irfan (2025). The interaction was a conversation between a human interlocutor and the Furhat robot, which achieved turn-taking using the VAP and TurnGPT models. The models were trained on Switchboard and deployed on an Nvidia 2080 GPU. In their study, Skantze and Irfan noted that the VAP model latency is very low, at 20 ms. However, the TurnGPT model requires transcription. In their study, Google speech-to-text was used in streaming mode. As the latency of this cloud service is variable, no specific latency figure was provided but we assume it was in the hundreds of milliseconds based on the recommended frame size of 100 ms⁸. Finally, there is the latency of the dialogue manager, which produces upcoming utterances from the Furhat robot. Adding up the latencies of all the components, this easily reaches several hundred milliseconds; problematic as human turn-taking very rapid (Chapter 2).

To reduce effective latency, Skantze and Irfan leveraged the model's tendency to predict turn endings early, which we also observed in Chapter 4. When the TurnGPT model prediction

⁸<https://docs.cloud.google.com/speech-to-text/docs/best-practices>

exceeded 0.2, the dialogue manager formed a candidate next utterance. The VAP model then predicted the end-of-turn. Their approach was compared to a baseline which waited for silences. In a qualitative study of user experience, they found participants felt the approach was more natural, and handled interruptions and overlapping speech more smoothly than end-pointing. The end-pointing model was described as laggy and awkward around pauses, making it feel more like a voice assistant than a conversational partner.

Their work therefore highlights the benefit of existing PTTM models and the feasibility of deployment. However, there are outstanding issues with latency, which is variable in a cloud-based ASR system (Skantze and Irfan 2025). A major limitation is a GPU is required to run modern PTTMs. GPU are costly and not energy efficient and require a powerful CPU and cooling, which makes their deployment in real-world consumer devices for instance in the home or workplace, not feasible from a device cost and energy consumption perspective.

7.5.2 Deployment of MM-VAP to human-robot interaction

Next, we consider the deployment of our MM-VAP model to real-world interactions. The model presents similar limitations to the VAP model in a deployment, most notably, the model would need to be deployed on a GPU similarly to the work of (Skantze and Irfan 2025), presenting cost and hardware issues.

The model is causal, making it suited to a real-world deployment. The visual feature processing is, however, not causal. This is because we normalised certain features at the participant level. This was due to the variability in the camera angles across participants in the Candor corpus, discussed in Sections 6.1.5 and 6.2.1. As we zero-mean certain features, the entire recording of a participant is required to compute the mean before inference. The need for normalisation could be avoided if a fixed camera setup was used in the deployed device, however, it is not practical for participants to always be in the same position. The model would therefore need to be robust to different camera angles and the positioning of participants relative to the deployed PTTM. A further barrier to deployment of the MM-VAP model is the use of OpenFace (Baltrušaitis, Robinson, et al. 2016) to extract visual features. Although the model can run in real-time, it is not GPU-optimised.

Furthermore, we found that the tracking was not robust to participants entering or leaving the frame, multiple participants appearing, etc. as noted in Chapter 6. We overcame this by ignoring these instances, but this is not practical in a real-world setting and would need to be handled appropriately. For instance, the model would need to be able to rely on audio alone to make its predictions if a participant is momentarily not visible or obscured. Finally, a further limitation which should be overcome is that the extracted audio features have a different sample rate than the video features (30 vs 50 Hz). We upsampled by interpolation, which in a real-time deployment would lead to a slight delay as future frames are needed.

This alignment issue would need to be solved in a deployment to a real-world human-robot interaction scenario.

7.5.3 Remaining issues to be overcome in a real-world deployment

A major issue which needs to be addressed is the way in which participants behave in a human-robot versus a human-human interaction. This is a key concern in human-robot interaction. Frijns et al. (2023) note that in studying human-robot interaction, models of human-human communication are often transposed directly into human-robot interaction. This is insufficient as in a conversational interaction between human and a robot, the robot is often not viewed as an equal conversation partner by the human, which impacts the way in which the interaction unfolds. Therefore, they argue that human-robot interaction should not be treated as a human-human interaction with a robot substituted in (Frijns et al. 2023).

This has implications for predictive turn-taking models, which are trained on human-human interactions and then deployed in human-robot interactions. Although evidence shows humans use similar turn-taking cues in interactions with robots, such as filled pauses and gaze aversion (Bilac et al. 2017), speech directed at a robot has clear differences in F_0 , timing and rhythm (Cohn and Zellou 2021). As PTTMs are trained speech with the prosody, rhythm and turn-taking dynamics of human-human conversation, care must be taken to ensure models work well in these conditions. This has been studied with the VAP model in the aforementioned study by Skantze and Irfan (2025), but it is unclear whether speakers use similar facial expressions and non-verbal cues that our MM-VAP model is trained on. Future work should therefore carefully note that speakers may exhibit markedly different non-verbal turn-taking behaviours interacting with a robot than interacting with a human. As the MM-VAP model was trained on human-human interaction, generalisation to embodied human-robot interaction should be carefully explored.

7.5.4 The role this thesis plays in achieving predictive turn-taking in human-robot interaction

Our contributions to knowledge are summarised in Section 1.6 and further discussed our closing chapter. However, this thesis has made a number of advancements in our understanding of model deployment to real-world human-robot interaction, discussed here.

Our first contribution in this area is our evaluation of PTTM performance in noise, which is to our knowledge, the first evaluation of turn-taking models in noise. We showed that audio only models are not robust to acoustic interference, even if speech remains largely intelligible, an important finding as PTTMs are expected to be deployed across diverse environments, and they will invariably encounter noise. Our second contribution relates to our understanding of multimodality in PTTMs. In speech, we presented a thorough analysis of the roles of

prosodic and lexical cues in the models, showing how both make an important contribution to turn-taking. We also showed how the incorporation of non-verbal cues improves model performance. Returning to noise, we found visual cues supported prediction in noise, however this was only the case when prior examples of noise were included during training. This illustrates the importance of the training process in multimodal models, to ensure that models are able to exploit the extra information offered by multimodal cues to the fullest extent.

This shows that multimodality makes models robust: where intelligibility is removed, prosody can be relied upon, and where noise masked speech, non-verbal cues can be exploited for turn-taking prediction. Given the vast majority of PTM are currently trained on speech alone, our findings therefore highlight the importance that must be placed upon non-verbal cues in the development of these models. A further contribution we have made is to highlight the importance of considering the generalisation of the models across different datasets and types of interaction. We noted a model trained on Switchboard telephone speech did not perform well on Candor videoconferencing speech and vice-versa. We discuss how this could be because in telephone speech, there are no non-verbal cues, and alternatively discussed how the types of interaction differ in topic, length, and turn-taking dynamics. As corpora are extensively used in PTM training, this highlighted the importance of ensuring trained models will generalise across a variety of interaction style, formality level and topic. To our knowledge, this has not been considered explicitly in the turn-taking literature, though we note that multilingual generalisation across corpora has been considered (Inoue, Bing'er Jiang, et al. 2024).

Chapter 8

Conclusion and future work

The development of robust, accurate predictive turn-taking models (PTTMs) is a vital step in achieving fluid interaction between humans and machines. In this thesis, we proposed new models and methods to advance the state of the art in predictive turn-taking. We demonstrated that a range of prosodic, lexical and visual cues can be exploited for prediction in PTTMs, yielding insights into the role of multimodality in turn-taking prediction. This reflects the highly multimodal nature of human–human interaction, and underscores the need for turn-taking, and conversational models more broadly, to incorporate multimodal features.

This chapter brings together our findings from our work chapters. We align our findings with our research questions and discuss related work that was conducted independently of our own. Finally, the chapter concludes the thesis by outlining several promising directions for future work based on our findings.

8.1 Summarising the key findings and insights of this thesis

In this section, we review our major findings and align them with our research questions (Section 1.4, pg. 6).

8.1.1 Exploring machine transcription of conversational speech and demonstrating its utility in training data generation

The broad interest in the use of automatic speech recognition (ASR) for conversational speech transcription — along with the lack of discussion on ASR use for a general audience — led us to propose **Research Question 1**, which is explored in **Chapter 3**. We explored manual and ASR transcription of conversational speech, finding commercial ASR mostly captured an accurate representation of what was said, whereas open-source ASR contained considerably more errors. A notable gap emerged in the transcription of non-lexical tokens (NLTs) - short words such as *uhm* and *ehm* that are important in conversation. We also

discussed the benefits and disadvantages of both types of ASR platform from a practical standpoint. This discussion was aimed at a general audience looking to use ASR in their work.

Thus, whether or not the findings of Chapter 3 demonstrate ASR is a suitable tool for automated transcription largely depends on a researcher's expectation of what a transcription should contain. If a researcher wants to capture purely what was said (an *orthographic* transcription), commercial systems perform quite well with few errors. However, important aspects are missing, such as the tagging of linguistic phenomena, and non-verbal behaviours that are often noted in transcription (head nodding to signal 'yes') are similarly absent. Therefore, in many research applications, ASR is unlikely to be useful without hand correction. However, as noted, using ASR to generate an initial first-pass transcription for manual correction may be faster than transcribing from scratch. The discussions and insight offered in Chapter 3 fill an important niche given the widespread interest in ASR as a transcription tool.

As our review of the literature noted, the development of PTTMs exclusively depended upon manual transcription. We therefore explored the use of ASR to transcribe conversations for training PTTMs. We proposed **Research Question 2** to tackle this, addressed in **Chapter 4**. We found no statistically significant difference in performance of the audio-only voice activity projection (VAP) turn-taking model (Ekstedt and Skantze 2022b) when trained on the ASR and manual word alignments. To our knowledge, this is the first analysis of its kind, and the lack of any significant difference in performance answers Research Question 2. Along with our benchmarking of ASR in Chapter 3, this finding enables us to vastly expand the amount of data available for training using ASR. This is particularly important for multimodal data, where manually transcribed corpora are scarce.

8.1.2 Understanding the role of visual cues in predictive turn-taking

Our literature review highlighted the role visual cues play in supporting turn-taking in human-human interaction. Despite this, PTTMs typically rely upon speech and text to make their predictions, with only a single analysis of multimodal cues reported at the time of thesis planning (Roddy, Skantze, et al. 2018b). We therefore proposed **Research Question 4**, which we addressed in **Chapter 4**. We trained the VAP turn-taking model on the Candor corpus of videoconferencing interaction (Reece et al. 2023), a new domain. Despite not having access to the same non-verbal information as the interlocutors, the VAP model performed well. Thus, these findings demonstrate that visual cues are not strictly necessary to include for accurate turn-taking prediction on this corpus, answering Research Question 4.

Despite this, our review of the turn-taking literature noted that the vital role of visual cues in turn-taking is well-established. We therefore proposed **Research Question 5**, which we addressed in **Chapter 6** by proposing our new **multimodal voice activity projection (MM-**

VAP) model. The model included a variety of facial features such as gaze, head pose and facial expression, in addition to audio cues. We demonstrated MM-VAP outperformed VAP on the Candor corpus across all turn-taking metrics considered. The difference in performance was slight, but statistically significant. The inclusion of visual cues therefore improves turn-taking prediction, answering Research Question 5. This is an important contribution as prior work has typically ignored visual cues. Although the performance increase is relatively small, this is in line with the turn-taking literature, which details how subtle non-verbal cues such as eyebrow movement play a supporting role in turn-taking in human-human interaction (Section 2.2).

8.1.3 Turn-taking model evaluation in real-world deployments

Our literature review highlighted an important gap: PTTMs are exclusively trained on corpora of human-human interaction that were recorded in clean, noise-free conditions. However, in a real-world deployment, they will be expected to operate across a range of conditions, which may include noise. We addressed this gap by first proposing **Research Question 6**, addressed in **Chapter 7**. We generated examples of noise which mimic the types of noise that a deployed PTTM may experience, adding them to the Candor corpus at different signal-to-noise ratio (SNR). In light (+10 dB SNR) noise, the accuracy of both VAP and MM-VAP dramatically dropped, indicating brittle performance. PTTMs are therefore not inherently robust to the types of noise that they will encounter in real-world settings, answering Research Question 6.

As studies of human interaction demonstrate that humans exploit visual cues to overcome acoustic noise in conversation (MacLeod and Summerfield 1987), we proposed **Research Question 7b** to explore whether MM-VAP could similarly exploit visual information. We found that when we included examples of noise during training, a large gap in the performance of MM-VAP and VAP models emerged. As the noise strength increased, the VAP model performance reduced to chance, whereas the MM-VAP model performance remained. This demonstrates the MM-VAP model exploits visual cues when acoustic noise disrupts speech-based turn-taking cues, addressing Research Question 7b. There are, however, important caveats: noise must be included during training, and the performance increase does not necessarily extend to types of noise unseen during training.

We also conducted a cross-corpus analysis of the VAP model. We found that a VAP model trained on the Switchboard corpus (telephone speech) did not achieve similar levels of performance in Candor (videoconferencing speech), and vice versa. We discussed several potential reasons for this, including how the availability of non-verbal cues may impact the way in which interlocutors use verbal turn-taking cues. However, there are many confounding factors here, including large differences in corpus recording setup, conversation topics and style, and the session length. Yet in the present literature, models are trained and tested on a

single corpus and then expected to generalise to new unseen and unconstrained interaction (Skantze and Irfan 2025; Skantze 2021; Inoue, Okafuji, et al. 2025). Our findings therefore highlight the importance of assessing generalisation, which is vital in real-world deployments.

8.1.4 Uncovering the role of prosodic and lexical cues in predictive turn-taking

In addition to the role of visual cues in conversation, speech itself is richly multimodal. Our literature review outlined studies which reveal how prosodic and lexical cues are leveraged by humans during turn-taking (Chapter 2). We therefore proposed **Research Question 3** to explore the relative contributions of prosodic and lexical cues to prediction. In **Chapter 4**, we analysed turn-taking cues from speech in the Candor corpus, finding significant differences in fundamental frequency ($F0$) and intensity prior to shifts and holds: relationships that a trained PTM may exploit.

However, analysis of PTM is complicated by the use of self-supervised speech representation (S3R). We therefore proposed a **vocoder-based methodology** to selectively control for the presence and absence of $F0$, intensity and lexical cues in speech, supporting a controlled analysis of turn-taking cues in the VAP turn-taking model. This revealed that both prosodic and lexical cues support predictive turn-taking, as the model performed best when both cues were intact. However, we found that speech can be replaced by noise, matching the original $F0$ and intensity contours, and model performance remains above chance. Similarly, when prosodic cues were removed by flattening and lexical cues remained intact, performance remained above chance. This analysis therefore addresses Research Question 3, as both prosodic and lexical cues are exploited for prediction in a trained PTM, but either can be used where one cue becomes unavailable, without further training. However, a number of cues related to turn-taking, such as syllable and vowel duration, are preserved by the vocoder and may be exploited instead of lexical cues.

We therefore continued Chapter 4 with an analysis of large language model (LLM) based turn-taking models. As these models use only text, they more fully isolate the role of lexical cues. We found that the previously reported high accuracy of the models masked the high number of false positive predictions. A qualitative analysis revealed that these originate in large part due to syntactic ambiguity, as the utterance was grammatically and syntactically complete, but the speaker continued. We therefore proposed **Research Question 7a**, asking whether cues from speech could overcome this vulnerability. We compared with audio-only turn-taking models and demonstrated they do not share this vulnerability, showing the role of both prosodic and lexical cues in turn-taking prediction and addressing this research question. In **Chapter 7**, we proposed our **Audio TurnGPT** model which is an LLM-based turn-taking model. We found the model reduced the number of false positive predictions, further addressing Research Question 7a.

8.2 Contemporary related work

We now highlight related works developed independently of our own since the time of thesis planning. A multimodal PTTM has been developed by Onishi et al. (Onishi et al. 2024; Onishi et al. 2023), demonstrating the ongoing research activity in this field. The model shares similarities with our MM-VAP introduced in Chapter 6, however, there are key differences in the modelling approaches used and analysis conducted.

Like our MM-VAP model, Onishi et al.'s model is based on the VAP model and also utilises visual cues extracted from OpenFace. The inclusion of visual cues leads to an improved performance over the VAP model. An ablation study revealed that facial action units (FAUs) contribute the most to the increased performance; the same conclusion as our ablation study in Section 6.2.3. A key difference is that their model depends on the ground-truth alignment for both training and inference. Our MM-VAP model only requires audio and video for inference, which is a more realistic setting as there are all that would be available in a real-world deployment. Their model also utilised a simpler concatenation of features rather than our use of cross attention to integrate audio and non-verbal turn-taking features (Section 6.2.2; pg.154). A further difference is that their model is trained on the Noxi corpus, which is very limited in size (11 hours of English data).

It is difficult to directly compare their model with our MM-VAP model, as their model is not publicly available. A further issue is that their work uses a different corpus, training split and so on, complicating a direct comparison between the works. This challenge reflects the lack of a benchmark dataset for comparison of turn-taking models, noted in a recent survey of turn-taking prediction (Castillo-López et al. 2025). Whilst there would be many challenges in creating such a benchmark for the community, languages, corpora, task based corpora, and so on, as highlighted at various stages of this thesis (e.g. Sections 2.3 and 4.5). To ensure comparison of our model with future work, we have made our training parameters, model and train/test/validation splits of the Candor corpus publicly available (O'Connor Russell and Harte 2025a).

Other related contemporary work includes a recent study by Inoue, Okafuji, et al., who looked at the impact of noise on turn-taking prediction performance (Inoue, Okafuji, et al. 2025), similarly to our work in Chapter 7. They train a VAP turn-taking model on corpora of Japanese interactions, with noise added in from mixed sources including the MUSAN corpus (Snyder et al. 2015), which we also use in our experiment. The VAP model was then deployed to a small robot and placed in a shopping centre where real-world ambient noise was present. The model trained with noise outperformed the model with no noise added during training. Their findings complement our work in Chapter 7. Though they do not quantify the disruption caused to the model by noise, and how this varies with the strength of the noise e.g. accuracy vs SNR, as done in our work (Figures 7.1, 7.2 and 7.1). It would be

interesting to test whether our MM-VAP could further improve this performance through the inclusion of visual cues in a real-world deployment.

Finally, several corpora suitable for turn-taking model development have been released since thesis planning. As noted in Section 2.3 of Chapter 2 the 4,000 hr *Seamless Interaction Dataset* corpus from Meta (Agrawal et al. 2025) has recently been released. It is freely available and contains full-frontal recordings of participants engaged in dyadic English interaction, making it ideally suited to the development of multimodal PTTMs. Furthermore, as this corpus is recorded in-person, it removes the limitations of our work which originate from the use of videoconferencing. Our use of the Candor corpus was necessary at the time of thesis planning due to a lack of suitable corpora for model development. Another noteworthy corpus is the AVCocktail corpus (Nguyen et al. 2025). Participants sit around a table in small groups. Single channel audio is available, and video of each participant is also available. Participants engage in multiple, simultaneous English language conversations. This makes AVCocktail ideally suited for an analysis of PTTM performance in noise. As the conversations are recorded in noise by speakers directly experiencing the noise, a key limitation of Chapter 7 is addressed, as we added noise to speech recorded in clean, noise-free conditions.

8.3 Directions for future work

We identify a number of interesting directions for future work that have arisen throughout the thesis, before making some final concluding remarks in the following section.

8.3.1 Develop predictive turn-taking models for multiparty interaction

A natural direction for further work is the development of PTTMs for multiparty interaction. Multiparty interaction is considerably more complex than the dyadic interactions considered in this thesis, as the next speaker must be identified (Section 2.6.10, pg. 53). Some models have been developed for multiparty interaction, but they generally rely on gaze being annotated at a very granular level and leverage corpora of limited size (Malik et al. 2020).

The development of multimodal models has been complicated by the lack of suitable corpora. Future work may therefore explore whether dyadic models can be used as a useful starting point for developing multiparty models. For instance, the VAP model might be trained on a large two-party corpus such as Candor and then used to predict when speaker changes occur. This could then be followed by a classifier trained on a smaller dataset of multiparty interaction, which predicts who the next speaker should be, bypassing the complexity of training a full multiparty turn-taking model from scratch.

In a recent study, Elmers et al. trained an audio-only variant of the VAP model for Japanese three-party conversational interaction (Elmers et al. 2025). However, they found low accuracy

in determining the next speaker from audio alone (62%). Future work should establish whether this was due to the absence of non-verbal cues in the model. As multiparty next-speaker detection is generally agreed to rely on visual cues (Negroponte 1995; A. H. Anderson 2006; Petukhova and Bunt 2009), future work might consider this in the context of extending the MM-VAP model to multiparty interaction.

8.3.2 Robust multimodal architectures

There are a number of improvements that could be made to our MM-VAP model presented in Chapter 6 in future work. The model should be able to handle different frame rates of the audio and video modalities. In the current model, this is handled by upsampling of the video modalities, but future versions of the model should be able to handle this mismatch natively, perhaps using a variant of the cross-attention method utilised in our Audio TurnGPT model outlined in Section 7.4.2. The model should be made more robust to different camera angles. In the current model, different camera angles of participants in the Candor corpus is handled by normalisation of certain features to have a zero mean (Section 6.1). This requires the entire video to estimate the mean, a dependency which should be removed in future work.

There are a number of issues with the OpenFace feature extraction method utilised by the model, which also should be addressed. The model fails for various reasons, such as when participants enter and leave the frame. We overcame this by simply excluding these sessions, but future models should be more robust to these phenomena, e.g. a temporary occlusion of a participant. Many of these issues would likely be resolved if OpenFace features were replaced with a learned representation such as AVHubert (Shi et al. 2022). However, as noted in Chapters 5 and 7, there is an inevitable trade-off between representation power and model interpretability when using self-supervised representations.

8.3.3 Real-world deployment of models to human-robot interaction

As noted in Section 7.5, current PTTMs are designed and tested in the lab, with limited consideration of real-world deployments (Skantze and Irfan 2025). Future work should therefore focus on tackling the issues we highlighted in Section 7.5, such as the development of a lightweight PTTMs that can run on small devices such as smart speakers without the need for a GPU. This work may be informed by our finding of a VAP model trained on noise matching the prosody of speech, achieving 87-91% of the performance of a model trained on unmodified speech in Chapter 5 (Figure 5.6, pg. 129). Future models may therefore only need the prosodic contour of speech, possibly reducing model size and hardware requirements.

Future work should consider other languages and dialects, as our work considers only US English. Turn-taking models should also consider a diverse range of interlocutors, including children, the elderly and neurodiverse interlocutors. Our finding of relatively limited

generalisation across corpora in Section 4.5 also warrants further investigation. Corpora can differ in their communication styles and register, which may hinder the ability of models to generalise across diverse corpora. As noted, another issue which should be tackled further is noise, as our method in Chapter 7 does not account for how interlocutors adjust their articulation in response to noise.

Finally, another remaining issue is to what extent models generalise when they are deployed to real-world interactions between humans and robots. Whilst turn-taking models trained on human-human corpora have been shown to work well in human-robot interaction (Skantze and Irfan 2025; Inoue, Okafuji, et al. 2025), multimodal models such as our MM-VAP model have not been studied. Hence, future work should consider the deployment of our MM-VAP model to real human-robot interaction. This raises a number of important points. Firstly, humans are likely to address a robot differently than they address one another, and they may alter their use of gesture, gaze and facial expression. However, there has been relatively limited work in this area. It is therefore unclear whether there would be similar non-verbal cues on which MM-VAP has been trained (human-human interaction) that might be able to exploit once deployed to a human-robot interaction. Furthermore, the embodied robot should not only respond to human non-verbal turn-taking cues, but it should also be able to produce them e.g. gaze aversion. Listener sensitivity of turn-taking cues in synthesised speech has been demonstrated (O'Mahony et al. 2023). Listener sensitivity has also been demonstrated for non-verbal turn-taking cues such as gaze aversion in a human-robot interaction (Skantze, Hjalmarsson, et al. 2014). A future study might therefore explore the simultaneous prediction of upcoming turn-taking events with the generation of turn-taking cues in a human-robot interaction deployment.

8.4 Concluding remarks

This thesis has presented a comprehensive evaluation and analysis of multimodality in predictive turn-taking models. We have introduced new models and methodologies to explore the contributions of visual, prosodic and lexical cues to turn-taking prediction. Our findings have outlined the vital contributions of each modality to prediction, echoing the multimodal nature of human-human interaction itself. It is hoped that these findings reinforce the vital role of multimodality in conversation. Future work should therefore consider language as a process that takes place in a situated, multimodal environment: all cues must be analysed by the model, across all the communication modalities available to a human interlocutor.

Appendices

Appendix A

Baseline turn-taking model performance

A.1 Derivation of baseline model performance

A.1.1 Baseline hold/shift prediction model

We compute a baseline model for the shift prediction metric, defined in Section 4.2.2, pg. 88. There is an imbalance in the number of holds and shifts in each corpus, as shown in Table 4.5, pg. 88. In the baseline model, the model always predicts the majority class (hold). This means the number of true and false positives and negatives outputted by the baseline model are:

$$TP = 0; \quad FP = 0; \quad FN = N_{shift}; \quad TN = N_{hold} \quad (A.1)$$

Where N_{shift} and N_{hold} are the numbers of shifts and holds in each corpus, as listed in Table 4.5. To derive the baseline, we compute the precision (P) and recall (R) of the model using Equation 2.14, pg. 52. For the shift class, precision and recall are:

$$P_{shift} = 0 \quad R_{shift} = 0 \quad (A.2)$$

For the hold class, precision and recall are:

$$P_{hold} = \frac{N_{hold}}{N_{hold} + N_{shift}} \quad R_{hold} = \frac{N_{hold}}{N_{hold} + 0} = 1 \quad (A.3)$$

Following Equation 2.13, we then compute the $F1$ scores for both classes:

$$F_1(\text{HOLD}) = \frac{2P_{hold}R_{hold}}{P_{hold} + R_{hold}} = \frac{2P_{hold}}{1 + P_{hold}}; \quad F_1(\text{SHIFT}) = 0 \quad (A.4)$$

Then, following Equation 2.17 we compute weighted F_1 score of both classes:

$$F_1(\text{WEIGHTED}) = \frac{N_{\text{hold}}}{N_{\text{hold}} + N_{\text{shift}}} F_1(\text{HOLD}) \quad (\text{A.5})$$

Finally, the balanced accuracy is given by:

$$\text{Balanced Accuracy} = \frac{1}{2} (R_{\text{hold}} + R_{\text{shift}}) = 0.5 \quad (\text{A.6})$$

A.1.2 Baseline shift prediction model

We also compute a baseline model for the shift prediction metrics defined in Section 4.2.3, pg. 89 following the same procedure. As before, the baseline model always predicts a hold. However, as described in Section 4.2.3 we sample an equal number of holds as there are shifts, meaning is no class imbalance as $N_{\text{hold}} = N_{\text{shift}}$.

For the shift class, precision and recall are:

$$P_{\text{shift}} = 0 \quad R_{\text{shift}} = 0 \quad (\text{A.7})$$

For the hold class, precision and recall are:

$$P_{\text{hold}} = \frac{N_{\text{hold}}}{N_{\text{hold}} + N_{\text{shift}}} = 0.5 \quad R_{\text{hold}} = \frac{N_{\text{hold}}}{N_{\text{hold}} + 0} = 1 \quad (\text{A.8})$$

We then compute the F_1 scores for both classes:

$$F_1(\text{HOLD}) = \frac{2P_{\text{hold}}R_{\text{hold}}}{P_{\text{hold}} + R_{\text{hold}}} = \frac{2 \times 0.5}{1 + 0.5} = 0.67; \quad F_1(\text{SHIFT}) = 0 \quad (\text{A.9})$$

Then, following Equation 2.17 we compute weighted F_1 score of both classes:

$$F_1(\text{WEIGHTED}) = \frac{N_{\text{hold}}}{N_{\text{hold}} + N_{\text{shift}}} F_1(\text{HOLD}) = \frac{1}{3} \approx 0.33 \quad (\text{A.10})$$

$$\text{Balanced Accuracy} = \frac{R_{\text{hold}} + R_{\text{shift}}}{2} = \frac{1 + 0}{2} = 0.5 \quad (\text{A.11})$$

A.2 Baseline model performance

Using the derivations from Section A.1 and the number of holds and shifts in the corpora, we compute the baseline model performances in Table A.1.

Table A.1 Baseline (always-hold) model F_1 scores and balanced accuracy for shifts and holds at different minimum floor-transfer offset (FTO) thresholds in the switchboard (SWB) and candor corpora.

Corpus	FTO	F_1 (HOLD)	F_1 (WEIGHTED)	Accuracy (Balanced, %)
SWB (ASR)	< -250	0.73	0.42	50
SWB (GT)	< -250	0.58	0.24	50
CND	< -250	0.84	0.61	50
SWB (ASR)	> +250	0.22	0.03	50
SWB (GT)	> +250	0.42	0.11	50
CND	> +250	0.45	0.13	50
SWB (ASR)	> 0	0.25	0.04	50
SWB (GT)	> 0	0.26	0.04	50
CND	> 0	0.36	0.08	50

Appendix B

Statistical significance tests

B.1 VAP model versus baseline model performance

This section presents one-sample t-tests comparing the performance of VAP models against baseline model performances from Appendix A. The results from these tests are referenced in Chapter 4.

B.1.1 Switchboard VAP model (ground-truth alignment) vs. baseline

Table B.1 One-sample t-tests comparing the VAP model trained on Switchboard (SWB) (ground-truth alignment) against an always-hold baseline model. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Corpus	Baseline	Mean	CI	Effect size	<i>p</i> / <i>t</i>
HOLD/SHIFT Prediction							
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ground-truth)	0.11	0.82	(0.81, 0.82)	94.86	<.01 / 212.11
	F_1 (hold)		0.42	0.90	(0.89, 0.90)	163.03	<.01 / 364.54
	F_1 (shift)		0.00	0.47	(0.44, 0.51)	16.79	<.01 / 37.53
	Accuracy (balanced)		0.50	0.66	(0.64, 0.68)	10.49	<.01 / 23.46
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ground-truth)	0.11	0.80	(0.79, 0.80)	132.66	<.01 / 296.64
	F_1 (hold)		0.42	0.89	(0.89, 0.89)	198.66	<.01 / 444.22
	F_1 (shift)		0.00	0.41	(0.37, 0.45)	13.87	<.01 / 31.00
	Accuracy (balanced)		0.50	0.63	(0.61, 0.65)	8.44	<.01 / 18.88
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ground-truth)	0.04	0.84	(0.83, 0.84)	122.81	<.01 / 274.61
	F_1 (hold)		0.26	0.91	(0.91, 0.92)	194.12	<.01 / 434.06
	F_1 (shift)		0.00	0.43	(0.40, 0.47)	14.97	<.01 / 33.46
	Accuracy (balanced)		0.50	0.65	(0.63, 0.67)	9.26	<.01 / 20.70
Before overlap (FTO < -250 ms)	F_1 (weighted)	SWB (ground-truth)	0.24	0.64	(0.62, 0.65)	39.00	<.01 / 87.21
	F_1 (hold)		0.58	0.58	(0.55, 0.60)	-0.08	.87 / -0.17
	F_1 (shift)		0.00	0.67	(0.66, 0.69)	62.47	<.01 / 139.69
	Accuracy (balanced)		0.50	0.63	(0.61, 0.65)	9.88	<.01 / 22.09
SHIFT Prediction							
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ASR)	0.33	0.74	(0.70, 0.76)	17.97	<.01 / 40.17
	F_1 (hold)		0.67	0.76	(0.74, 0.78)	5.40	<.01 / 12.08
	F_1 (shift)		0.00	0.70	(0.67, 0.74)	24.07	<.01 / 53.83
	Accuracy (balanced)		0.50	0.73	(0.70, 0.76)	10.51	<.01 / 23.50
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ASR)	0.33	0.75	(0.71, 0.79)	14.37	<.01 / 32.13
	F_1 (hold)		0.67	0.77	(0.73, 0.80)	3.20	.002 / 7.15
	F_1 (shift)		0.00	0.74	(0.70, 0.77)	25.47	<.01 / 56.95
	Accuracy (balanced)		0.50	0.76	(0.72, 0.79)	8.56	<.01 / 19.14

B.1.2 Switchboard VAP model (ASR alignment) vs. baseline

Table B.2 One-sample t-tests comparing the VAP model trained on Switchboard (SWB) (ASR alignment) against an always-hold baseline model. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Corpus	Baseline	Mean	CI	Effect size	p / t
HOLD/SHIFT Prediction							
During mutual silence (FTO > +250 ms)	F_1 (weighted)	SWB (ASR)	0.03	0.81	(0.80, 0.82)	91.75	<.01 / 205.15
	F_1 (hold)		0.22	0.90	(0.89, 0.90)	160.67	<.01 / 359.27
	F_1 (shift)		0.00	0.45	(0.41, 0.48)	16.02	<.01 / 35.83
	Accuracy (balanced)		0.50	0.65	(0.63, 0.67)	10.24	<.01 / 22.89
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ASR)	0.03	0.79	(0.78, 0.80)	104.48	<.01 / 233.63
	F_1 (hold)		0.22	0.89	(0.88, 0.89)	543.46	<.01 / 1215.22
	F_1 (shift)		0.00	0.40	(0.35, 0.44)	10.82	<.01 / 24.18
	Accuracy (balanced)		0.50	0.63	(0.60, 0.65)	6.56	<.01 / 14.67
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ASR)	0.04	0.83	(0.82, 0.84)	118.25	<.01 / 264.40
	F_1 (hold)		0.25	0.91	(0.90, 0.91)	195.24	<.01 / 436.56
	F_1 (shift)		0.00	0.42	(0.38, 0.46)	14.20	<.01 / 31.74
	Accuracy (balanced)		0.50	0.65	(0.63, 0.67)	8.98	<.01 / 20.08
Before overlap (FTO < -250 ms)	F_1 (weighted)	SWB (ASR)	0.42	0.62	(0.61, 0.63)	20.44	<.01 / 45.71
	F_1 (hold)		0.73	0.56	(0.53, 0.59)	-6.96	<.01 / -15.56
	F_1 (shift)		0.00	0.66	(0.63, 0.68)	31.36	<.01 / 70.12
	Accuracy (balanced)		0.50	0.61	(0.60, 0.63)	9.88	<.01 / 22.10
SHIFT Prediction							
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ASR)	0.33	0.73	(0.70, 0.76)	17.97	<.01 / 40.17
	F_1 (hold)		0.67	0.76	(0.74, 0.78)	5.40	<.01 / 12.08
	F_1 (shift)		0.00	0.70	(0.67, 0.74)	24.07	<.01 / 53.83
	Accuracy (balanced)		0.50	0.73	(0.70, 0.76)	10.51	<.01 / 23.50
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ASR)	0.33	0.75	(0.71, 0.79)	14.37	<.01 / 32.13
	F_1 (hold)		0.67	0.77	(0.73, 0.80)	3.20	.002 / 7.15
	F_1 (shift)		0.00	0.74	(0.70, 0.77)	25.47	<.01 / 56.95
	Accuracy (balanced)		0.50	0.76	(0.72, 0.79)	8.56	<.01 / 19.14

B.1.3 Candor VAP model vs. baseline

Table B.3 One-sample t-tests comparing the VAP model trained on Candor (CND) against an always-hold baseline model. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Corpus	Baseline	Mean	CI	Effect size	<i>p</i> / <i>t</i>
HOLD/SHIFT Prediction							
During mutual silence (FTO > +250 ms)	<i>F</i> ₁ (weighted)	CND	0.13	0.83	(0.83, 0.84)	184.61	<.01 / 412.79
	<i>F</i> ₁ (hold)		0.45	0.88	(0.88, 0.89)	246.02	<.01 / 550.11
	<i>F</i> ₁ (shift)		0.00	0.70	(0.69, 0.71)	72.04	<.01 / 161.09
	Accuracy (balanced)		0.50	0.79	(0.78, 0.80)	36.11	<.01 / 80.75
Before end of turn (FTO > +250 ms)	<i>F</i> ₁ (weighted)	CND	0.13	0.81	(0.81, 0.82)	126.79	<.01 / 283.52
	<i>F</i> ₁ (hold)		0.45	0.87	(0.87, 0.88)	97.86	<.01 / 218.81
	<i>F</i> ₁ (shift)		0.00	0.67	(0.65, 0.68)	75.16	<.01 / 168.06
	Accuracy (balanced)		0.50	0.77	(0.76, 0.77)	43.73	<.01 / 97.77
Before end of turn (FTO > 0 ms)	<i>F</i> ₁ (weighted)	CND	0.08	0.86	(0.85, 0.86)	196.20	<.01 / 438.71
	<i>F</i> ₁ (hold)		0.36	0.91	(0.91, 0.92)	234.31	<.01 / 523.92
	<i>F</i> ₁ (shift)		0.00	0.65	(0.64, 0.67)	56.45	<.01 / 126.22
	Accuracy (balanced)		0.50	0.78	(0.77, 0.79)	31.03	<.01 / 69.38
Before overlap (FTO < -250 ms)	<i>F</i> ₁ (weighted)	CND	0.61	0.78	(0.77, 0.78)	22.29	<.01 / 49.84
	<i>F</i> ₁ (hold)		0.84	0.85	(0.84, 0.86)	1.95	.012 / 4.35
	<i>F</i> ₁ (shift)		0.00	0.57	(0.54, 0.59)	26.81	<.01 / 59.96
	Accuracy (balanced)		0.50	0.70	(0.68, 0.72)	14.50	<.01 / 32.41
SHIFT Prediction							
Before end of turn (FTO > +250 ms)	<i>F</i> ₁ (weighted)	CND	0.33	0.86	(0.85, 0.86)	143.91	<.01 / 321.80
	<i>F</i> ₁ (hold)		0.67	0.84	(0.83, 0.85)	36.26	<.01 / 81.08
	<i>F</i> ₁ (shift)		0.00	0.87	(0.86, 0.87)	202.53	<.01 / 452.87
	Accuracy (balanced)		0.50	0.85	(0.85, 0.86)	93.99	<.01 / 210.18
Before end of turn (FTO > 0 ms)	<i>F</i> ₁ (weighted)	CND	0.33	0.86	(0.85, 0.86)	132.52	<.01 / 296.33
	<i>F</i> ₁ (hold)		0.67	0.84	(0.84, 0.85)	38.66	<.01 / 86.44
	<i>F</i> ₁ (shift)		0.00	0.87	(0.87, 0.88)	184.52	<.01 / 412.60
	Accuracy (balanced)		0.50	0.86	(0.85, 0.86)	93.13	<.01 / 208.25

B.2 Switchboard VAP model (ground-truth alignment) vs. Switchboard VAP model (ASR alignment)

This section compares VAP models trained on Switchboard ground-truth and ASR alignments with paired t-tests.

Table B.4 Paired t-tests comparing VAP models trained on Switchboard (SWB) (ground-truth) and SWB (ASR) alignments. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Model A	Model B	Mean A	Mean B	CI A	CI B	Effect size	p / t
HOLD/SHIFT Prediction									
During mutual silence (FTO > +250 ms)	F_1 (weighted)	SWB (ground-truth)	SWB (ASR)	0.82	0.81	(0.81, 0.82)	(0.80, 0.82)	0.91	.11 / 2.04
	F_1 (hold)			0.90	0.90	(0.89, 0.90)	(0.89, 0.90)	0.56	.28 / 1.26
	F_1 (shift)			0.47	0.45	(0.44, 0.51)	(0.41, 0.48)	0.91	.11 / 2.03
	Accuracy (balanced)			0.66	0.65	(0.64, 0.68)	(0.63, 0.67)	0.87	.12 / 1.94
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ground-truth)	SWB (ASR)	0.80	0.79	(0.79, 0.80)	(0.78, 0.80)	1.22	.05 / 2.72
	F_1 (hold)			0.89	0.89	(0.89, 0.89)	(0.88, 0.89)	1.57	.02 / 3.50
	F_1 (shift)			0.41	0.40	(0.37, 0.45)	(0.35, 0.44)	0.50	.32 / 1.13
	Accuracy (balanced)			0.63	0.63	(0.61, 0.65)	(0.60, 0.65)	0.44	.38 / 0.99
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ground-truth)	SWB (ASR)	0.84	0.83	(0.83, 0.84)	(0.82, 0.84)	0.81	.14 / 1.82
	F_1 (hold)			0.91	0.91	(0.91, 0.92)	(0.90, 0.91)	1.43	.03 / 3.19
	F_1 (shift)			0.43	0.42	(0.40, 0.47)	(0.38, 0.46)	0.38	.45 / 0.85
	Accuracy (balanced)			0.65	0.65	(0.63, 0.67)	(0.63, 0.67)	0.28	.56 / 0.63
Before overlap (FTO < -250 ms)	F_1 (weighted)	SWB (ground-truth)	SWB (ASR)	0.63	0.62	(0.62, 0.65)	(0.61, 0.63)	1.07	.08 / 2.39
	F_1 (hold)			0.58	0.56	(0.55, 0.60)	(0.53, 0.59)	0.76	.17 / 1.69
	F_1 (shift)			0.67	0.66	(0.66, 0.69)	(0.63, 0.68)	0.51	.32 / 1.14
	Accuracy (balanced)			0.63	0.61	(0.61, 0.65)	(0.60, 0.63)	1.23	.05 / 2.75
SHIFT Prediction									
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ground-truth)	SWB (ASR)	0.73	0.74	(0.71, 0.76)	(0.70, 0.76)	0.18	.70 / 0.41
	F_1 (hold)			0.76	0.76	(0.73, 0.79)	(0.74, 0.78)	-0.25	.61 / -0.56
	F_1 (shift)			0.71	0.70	(0.69, 0.73)	(0.67, 0.74)	0.68	.20 / 1.52
	Accuracy (balanced)			0.73	0.73	(0.71, 0.76)	(0.70, 0.76)	0.16	.74 / 0.36
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ground-truth)	SWB (ASR)	0.75	0.75	(0.73, 0.78)	(0.71, 0.79)	0.04	.94 / 0.08
	F_1 (hold)			0.76	0.77	(0.74, 0.79)	(0.73, 0.80)	-0.06	.91 / -0.12
	F_1 (shift)			0.74	0.74	(0.71, 0.77)	(0.70, 0.77)	0.17	.72 / 0.39
	Accuracy (balanced)			0.76	0.75	(0.73, 0.78)	(0.72, 0.79)	0.01	.98 / 0.03

B.3 Cross-corpus analysis of VAP model performance

This section compares the performance of a VAP model trained a corpus (corpus A) and deployed and another (corpus B) with a VAP model tested on the same corpus it was trained on (corpus A). Paired t-tests are used for comparison. Notation: training corpus $\rightarrow$ testing corpus. The ASR alignment is used for all Switchboard (SWB) models. The results from these tests are referenced in Chapter 4.

B.3.1 Comparing Candor $\rightarrow$ Switchboard VAP model against a Switchboard VAP model

Table B.5 Paired t-tests comparing CND $\rightarrow$ SWB and SWB VAP models cross-corpus performance. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Model A	Model B	Mean A	Mean B	CI A	CI B	Effect size	p / t
HOLD/SHIFT Prediction									
During mutual silence (FTO > +250 ms)	F_1 (weighted)	SWB (ASR)	CND $\rightarrow$ SWB (ASR)	0.81	0.76	(0.80, 0.82)	(0.76, 0.77)	5.90	<.01 / 13.19
	F_1 (hold)			0.90	0.88	(0.89, 0.90)	(0.87, 0.89)	1.99	.01 / 4.45
	F_1 (shift)			0.45	0.27	(0.41, 0.48)	(0.24, 0.31)	10.17	<.01 / 22.73
	Accuracy (balanced)			0.65	0.57	(0.63, 0.67)	(0.56, 0.58)	9.35	<.01 / 20.91
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ASR)	CND $\rightarrow$ SWB (ASR)	0.79	0.75	(0.78, 0.80)	(0.75, 0.76)	11.70	<.01 / 26.16
	F_1 (hold)			0.89	0.88	(0.88, 0.89)	(0.87, 0.88)	1.39	.04 / 3.10
	F_1 (shift)			0.40	0.23	(0.35, 0.44)	(0.18, 0.29)	8.61	<.01 / 19.25
	Accuracy (balanced)			0.63	0.55	(0.60, 0.65)	(0.53, 0.57)	9.10	<.01 / 20.34
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ASR)	CND $\rightarrow$ SWB (ASR)	0.83	0.79	(0.82, 0.84)	(0.79, 0.80)	6.90	<.01 / 15.42
	F_1 (hold)			0.91	0.90	(0.90, 0.91)	(0.90, 0.91)	2.20	.01 / 4.92
	F_1 (shift)			0.42	0.23	(0.38, 0.46)	(0.19, 0.26)	7.33	<.01 / 16.40
	Accuracy (balanced)			0.65	0.56	(0.63, 0.67)	(0.54, 0.57)	6.59	<.01 / 14.74
Before overlap (FTO < -250 ms)	F_1 (weighted)	SWB (ASR)	CND $\rightarrow$ SWB (ASR)	0.62	0.53	(0.61, 0.63)	(0.51, 0.54)	6.02	<.01 / 13.45
	F_1 (hold)			0.56	0.45	(0.53, 0.59)	(0.34, 0.55)	1.74	.02 / 3.88
	F_1 (shift)			0.66	0.58	(0.63, 0.68)	(0.50, 0.66)	1.34	.04 / 2.99
	Accuracy (balanced)			0.61	0.53	(0.60, 0.63)	(0.50, 0.55)	6.12	<.01 / 13.68
SHIFT Prediction									
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB (ASR)	CND $\rightarrow$ SWB (ASR)	0.73	0.61	(0.70, 0.76)	(0.55, 0.66)	3.07	<.01 / 6.86
	F_1 (hold)			0.76	0.67	(0.74, 0.78)	(0.62, 0.72)	2.27	.01 / 5.08
	F_1 (shift)			0.70	0.54	(0.67, 0.74)	(0.44, 0.63)	2.36	.01 / 5.27
	Accuracy (balanced)			0.73	0.61	(0.70, 0.76)	(0.56, 0.67)	3.16	<.01 / 7.06
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB (ASR)	CND $\rightarrow$ SWB (ASR)	0.75	0.61	(0.71, 0.79)	(0.57, 0.66)	4.91	<.01 / 10.97
	F_1 (hold)			0.77	0.61	(0.73, 0.80)	(0.53, 0.70)	1.99	.01 / 4.45
	F_1 (shift)			0.74	0.62	(0.70, 0.77)	(0.54, 0.70)	3.22	<.01 / 7.21
	Accuracy (balanced)			0.76	0.62	(0.72, 0.79)	(0.57, 0.67)	4.03	<.01 / 9.02

B.3.2 Comparing Switchboard → Candor VAP model against a Candor VAP model

Table B.6 Paired t-tests comparing SWB → CND cross-corpus performance. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Model A	Model B	Mean A	Mean B	CI A	CI B	Effect size	p / t
HOLD/SHIFT Prediction									
During mutual silence (FTO > +250 ms)	F_1 (weighted)	SWB → CND	CND	0.71	0.83	(0.69, 0.73)	(0.83, 0.84)	-9.80	<.01 / -21.90
	F_1 (hold)			0.81	0.88	(0.80, 0.82)	(0.88, 0.89)	-9.67	<.01 / -21.62
	F_1 (shift)			0.46	0.70	(0.43, 0.49)	(0.69, 0.71)	-8.55	<.01 / -19.12
	Accuracy (balanced)			0.63	0.79	(0.61, 0.65)	(0.78, 0.80)	-8.96	<.01 / -20.04
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB → CND	CND	0.71	0.81	(0.69, 0.72)	(0.81, 0.82)	-12.41	<.01 / -27.74
	F_1 (hold)			0.80	0.87	(0.79, 0.81)	(0.87, 0.88)	-10.14	<.01 / -22.66
	F_1 (shift)			0.45	0.67	(0.43, 0.48)	(0.65, 0.68)	-13.55	<.01 / -30.29
	Accuracy (balanced)			0.62	0.77	(0.61, 0.64)	(0.76, 0.77)	-15.86	<.01 / -35.46
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB → CND	CND	0.77	0.86	(0.76, 0.77)	(0.85, 0.86)	-21.77	<.01 / -48.69
	F_1 (hold)			0.87	0.91	(0.86, 0.87)	(0.91, 0.92)	-12.38	<.01 / -27.69
	F_1 (shift)			0.40	0.66	(0.37, 0.42)	(0.64, 0.67)	-14.57	<.01 / -32.58
	Accuracy (balanced)			0.62	0.78	(0.61, 0.63)	(0.77, 0.79)	-13.28	<.01 / -29.70
Before overlap (FTO < -250 ms)	F_1 (weighted)	SWB → CND	CND	0.68	0.78	(0.66, 0.70)	(0.77, 0.78)	-5.59	<.01 / -12.49
	F_1 (hold)			0.81	0.85	(0.80, 0.81)	(0.84, 0.86)	-3.94	<.01 / -8.80
	F_1 (shift)			0.33	0.57	(0.28, 0.39)	(0.54, 0.59)	-6.83	<.01 / -15.26
	Accuracy (balanced)			0.57	0.70	(0.54, 0.59)	(0.68, 0.72)	-7.30	<.01 / -16.32
SHIFT Prediction									
Before end of turn (FTO > +250 ms)	F_1 (weighted)	SWB → CND	CND	0.70	0.86	(0.68, 0.72)	(0.85, 0.86)	-9.80	<.01 / -21.90
	F_1 (hold)			0.68	0.84	(0.67, 0.70)	(0.83, 0.84)	-15.00	<.01 / -33.54
	F_1 (shift)			0.71	0.87	(0.68, 0.75)	(0.86, 0.87)	-6.29	<.01 / -14.07
	Accuracy (balanced)			0.70	0.85	(0.68, 0.72)	(0.85, 0.86)	-11.78	<.01 / -26.33
Before end of turn (FTO > 0 ms)	F_1 (weighted)	SWB → CND	CND	0.70	0.86	(0.68, 0.72)	(0.85, 0.86)	-8.76	<.01 / -19.59
	F_1 (hold)			0.69	0.84	(0.67, 0.70)	(0.84, 0.85)	-10.52	<.01 / -23.53
	F_1 (shift)			0.72	0.87	(0.69, 0.75)	(0.87, 0.88)	-6.73	<.01 / -15.06
	Accuracy (balanced)			0.71	0.86	(0.69, 0.72)	(0.85, 0.86)	-9.32	<.01 / -20.83

B.4 Comparison of VAP, MM-VAP, and video-only turn-taking models

This section compares VAP (audio-only), MM-VAP (audio + video, early and late fusion versions) and a video-only turn-taking model on the Candor corpus. The results from these tests are referenced in Chapter 6.

B.4.1 VAP vs. MM-VAP (early fusion)

Table B.7 Paired t-tests comparing VAP and MM-VAP (early fusion) models on the Candor corpus. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Model A	Model B	Mean A	Mean B	CI A	CI B	Effect size	p / t
HOLD/SHIFT Prediction									
During mutual silence (FTO > +250 ms)	F_1 (weighted)	VAP	MM-VAP (early fusion)	0.83	0.84	(0.83, 0.84)	(0.84, 0.86)	-1.96	.01 / -4.38
	F_1 (hold)			0.88	0.90	(0.88, 0.89)	(0.89, 0.90)	-1.94	.01 / -4.34
	F_1 (shift)			0.70	0.74	(0.69, 0.71)	(0.72, 0.75)	-1.92	.01 / -4.30
	Accuracy (balanced)			0.79	0.82	(0.78, 0.80)	(0.81, 0.83)	-1.92	.01 / -4.29
Before end of turn (FTO > +250 ms)	F_1 (weighted)	VAP	MM-VAP (early fusion)	0.81	0.83	(0.81, 0.82)	(0.82, 0.84)	-1.74	.02 / -3.90
	F_1 (hold)			0.87	0.88	(0.87, 0.88)	(0.88, 0.89)	-1.60	.02 / -3.58
	F_1 (shift)			0.67	0.69	(0.65, 0.68)	(0.68, 0.71)	-1.91	.01 / -4.27
	Accuracy (balanced)			0.77	0.79	(0.76, 0.77)	(0.77, 0.80)	-1.98	.01 / -4.42
Before end of turn (FTO > 0 ms)	F_1 (weighted)	VAP	MM-VAP (early fusion)	0.86	0.87	(0.85, 0.86)	(0.87, 0.88)	-2.21	.01 / -4.93
	F_1 (hold)			0.91	0.92	(0.91, 0.92)	(0.92, 0.92)	-2.08	.01 / -4.65
	F_1 (shift)			0.66	0.69	(0.64, 0.67)	(0.68, 0.71)	-2.18	.01 / -4.88
	Accuracy (balanced)			0.78	0.80	(0.77, 0.79)	(0.79, 0.81)	-2.09	.01 / -4.67
Before overlap (FTO < -250 ms)	F_1 (weighted)	VAP	MM-VAP (early fusion)	0.78	0.79	(0.77, 0.78)	(0.78, 0.81)	-1.23	.05 / -2.76
	F_1 (hold)			0.85	0.87	(0.84, 0.86)	(0.86, 0.88)	-1.03	.08 / -2.29
	F_1 (shift)			0.57	0.60	(0.54, 0.59)	(0.58, 0.62)	-1.30	.04 / -2.90
	Accuracy (balanced)			0.70	0.72	(0.68, 0.72)	(0.71, 0.73)	-1.30	.04 / -2.91
SHIFT Prediction									
Before end of turn (FTO > +250 ms)	F_1 (weighted)	VAP	MM-VAP (early fusion)	0.86	0.87	(0.85, 0.86)	(0.86, 0.88)	-1.76	.02 / -3.93
	F_1 (hold)			0.84	0.86	(0.83, 0.85)	(0.85, 0.86)	-1.55	.03 / -3.48
	F_1 (shift)			0.87	0.88	(0.86, 0.87)	(0.88, 0.89)	-1.80	.02 / -4.02
	Accuracy (balanced)			0.85	0.87	(0.85, 0.86)	(0.86, 0.88)	-1.66	.02 / -3.71
Before end of turn (FTO > 0 ms)	F_1 (weighted)	VAP	MM-VAP (early fusion)	0.86	0.88	(0.85, 0.86)	(0.87, 0.89)	-2.54	<.01 / -5.68
	F_1 (hold)			0.84	0.86	(0.84, 0.85)	(0.85, 0.87)	-2.46	.01 / -5.51
	F_1 (shift)			0.87	0.89	(0.87, 0.88)	(0.88, 0.90)	-2.23	.01 / -5.00
	Accuracy (balanced)			0.86	0.88	(0.85, 0.86)	(0.87, 0.88)	-2.51	<.01 / -5.61

B.4.2 VAP vs. MM-VAP (late fusion)

Table B.8 Paired t-tests comparing VAP and MM-VAP (late fusion) models on the Candor corpus. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

Metric	Statistic	Corpus A	Corpus B	Mean A	Mean B	CI A	CI B	Effect size	p / t
HOLD/SHIFT Prediction									
During mutual silence (FTO > +250 ms)	F_1 (weighted)	VAP	MM-VAP (late fusion)	0.83	0.86	(0.83, 0.84)	(0.85, 0.86)	-3.73	<.01 / -8.34
	F_1 (hold)			0.88	0.90	(0.88, 0.89)	(0.89, 0.90)	-3.23	<.01 / -7.22
	F_1 (shift)			0.70	0.74	(0.69, 0.71)	(0.73, 0.75)	-3.77	<.01 / -8.44
	Accuracy (balanced)			0.79	0.83	(0.78, 0.80)	(0.82, 0.83)	-3.67	<.01 / -8.21
Before end of turn (FTO > +250 ms)	F_1 (weighted)	VAP	MM-VAP (late fusion)	0.81	0.83	(0.81, 0.82)	(0.83, 0.84)	-3.57	<.01 / -7.98
	F_1 (hold)			0.87	0.89	(0.87, 0.88)	(0.88, 0.89)	-3.46	<.01 / -7.74
	F_1 (shift)			0.66	0.70	(0.65, 0.68)	(0.69, 0.71)	-3.60	<.01 / -8.05
	Accuracy (balanced)			0.77	0.80	(0.76, 0.77)	(0.79, 0.80)	-3.42	<.01 / -7.64
Before end of turn (FTO > 0 ms)	F_1 (weighted)	VAP	MM-VAP (late fusion)	0.86	0.87	(0.85, 0.86)	(0.87, 0.88)	-2.91	<.01 / -6.50
	F_1 (hold)			0.91	0.92	(0.91, 0.92)	(0.92, 0.92)	-2.36	.01 / -5.27
	F_1 (shift)			0.66	0.71	(0.64, 0.67)	(0.69, 0.71)	-2.97	<.01 / -6.64
	Accuracy (balanced)			0.78	0.83	(0.77, 0.79)	(0.81, 0.83)	-2.67	<.01 / -5.96
Before overlap (FTO < -250 ms)	F_1 (weighted)	VAP	MM-VAP (late fusion)	0.78	0.80	(0.77, 0.78)	(0.80, 0.81)	-4.12	<.01 / -9.21
	F_1 (hold)			0.85	0.87	(0.84, 0.86)	(0.87, 0.88)	-2.76	<.01 / -6.18
	F_1 (shift)			0.57	0.62	(0.54, 0.59)	(0.61, 0.63)	-3.18	<.01 / -7.11
	Accuracy (balanced)			0.70	0.74	(0.68, 0.72)	(0.73, 0.74)	-3.31	<.01 / -7.40
SHIFT Prediction									
Before end of turn (FTO > +250 ms)	F_1 (weighted)	VAP	MM-VAP (late fusion)	0.86	0.87	(0.85, 0.86)	(0.86, 0.88)	-2.18	.01 / -4.88
	F_1 (hold)			0.84	0.86	(0.83, 0.85)	(0.85, 0.87)	-1.98	.01 / -4.43
	F_1 (shift)			0.87	0.88	(0.86, 0.87)	(0.87, 0.89)	-2.28	.01 / -5.09
	Accuracy (balanced)			0.85	0.87	(0.85, 0.86)	(0.86, 0.88)	-2.05	.01 / -4.59
Before end of turn (FTO > 0 ms)	F_1 (weighted)	VAP	MM-VAP (late fusion)	0.86	0.87	(0.85, 0.86)	(0.86, 0.88)	-2.58	<.01 / -5.77
	F_1 (hold)			0.84	0.86	(0.84, 0.85)	(0.85, 0.87)	-1.71	.02 / -3.82
	F_1 (shift)			0.87	0.89	(0.87, 0.88)	(0.88, 0.90)	-4.25	<.01 / -9.51
	Accuracy (balanced)			0.86	0.87	(0.85, 0.86)	(0.86, 0.88)	-1.91	.01 / -4.27

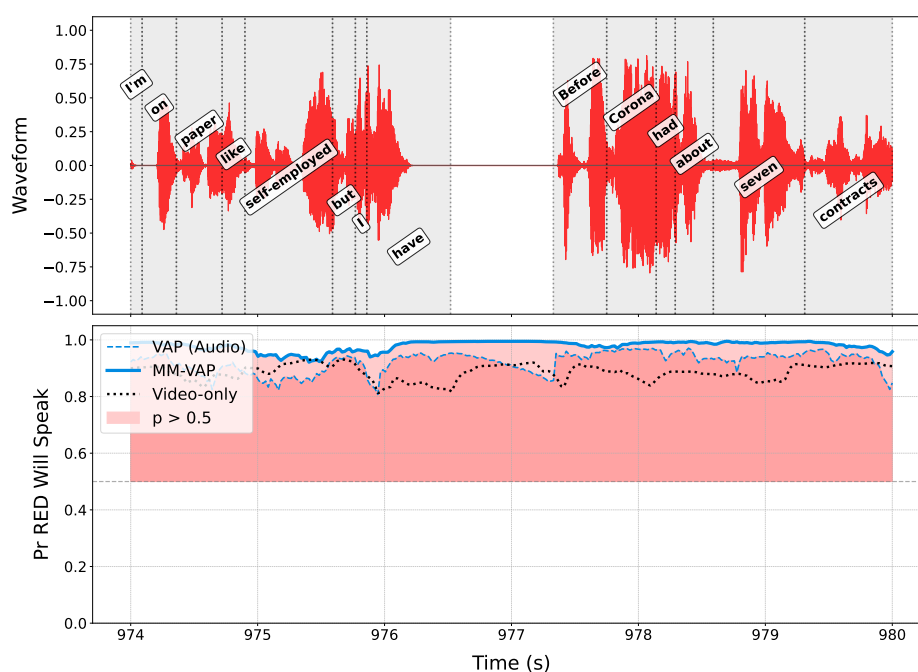
B.4.3 MM-VAP (late fusion) vs MM-VAP (early fusion)

Table B.9 Paired t-tests comparing MM-VAP (early fusion) and MM-VAP (late fusion) models. The mean test-set performance, baseline performance, 95% confidence intervals, effect sizes (Cohen's d), p -values, and t statistics are reported. Significant ($p < .05$) values reported in bold.

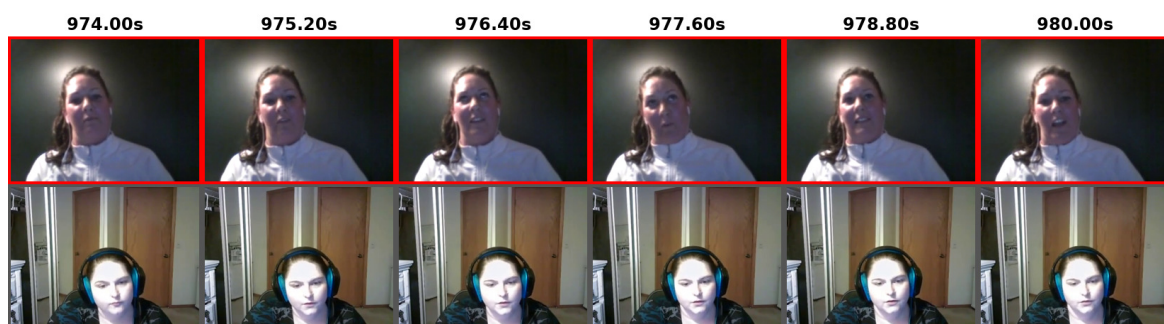
Metric	Statistic	Corpus A	Corpus B	Mean A	Mean B	CI A	CI B	Effect size	p / t
HOLD/SHIFT Prediction									
During mutual silence (FTO > +250 ms)	F_1 (weighted)	MM-VAP (early fusion)	MM-VAP (late fusion)	0.84	0.86	(0.84, 0.86)	(0.85, 0.86)	-0.50	.33 / -1.11
	F_1 (hold)			0.90	0.90	(0.89, 0.90)	(0.89, 0.90)	-0.51	.31 / -1.15
	F_1 (shift)			0.74	0.74	(0.72, 0.75)	(0.73, 0.75)	-0.48	.35 / -1.07
	Accuracy (balanced)			0.82	0.83	(0.81, 0.83)	(0.82, 0.83)	-0.45	.37 / -1.00
Before end of turn (FTO > +250 ms)	F_1 (weighted)	MM-VAP (early fusion)	MM-VAP (late fusion)	0.83	0.83	(0.82, 0.84)	(0.83, 0.84)	-0.50	.33 / -1.12
	F_1 (hold)			0.88	0.89	(0.88, 0.89)	(0.88, 0.89)	-0.28	.56 / -0.64
	F_1 (shift)			0.69	0.70	(0.68, 0.71)	(0.69, 0.71)	-0.76	.16 / -1.71
	Accuracy (balanced)			0.79	0.80	(0.78, 0.80)	(0.79, 0.80)	-0.95	.10 / -2.12
Before end of turn (FTO > 0 ms)	F_1 (weighted)	MM-VAP (early fusion)	MM-VAP (late fusion)	0.87	0.87	(0.87, 0.88)	(0.87, 0.88)	-0.21	.66 / -0.48
	F_1 (hold)			0.92	0.92	(0.92, 0.92)	(0.92, 0.92)	-0.15	.76 / -0.33
	F_1 (shift)			0.69	0.70	(0.68, 0.71)	(0.69, 0.71)	-0.26	.59 / -0.59
	Accuracy (balanced)			0.79	0.83	(0.79, 0.81)	(0.80, 0.83)	-0.31	.53 / -0.69
Before overlap (FTO < -250 ms)	F_1 (weighted)	MM-VAP (early fusion)	MM-VAP (late fusion)	0.79	0.80	(0.78, 0.81)	(0.80, 0.81)	-0.78	.16 / -1.75
	F_1 (hold)			0.87	0.87	(0.86, 0.88)	(0.87, 0.88)	-0.67	.21 / -1.51
	F_1 (shift)			0.60	0.62	(0.58, 0.62)	(0.61, 0.63)	-0.88	.12 / -1.97
	Accuracy (balanced)			0.72	0.74	(0.71, 0.73)	(0.73, 0.74)	-0.93	.11 / -2.08
SHIFT Prediction									
Before end of turn (FTO > +250 ms)	F_1 (weighted)	MM-VAP (early fusion)	MM-VAP (late fusion)	0.87	0.87	(0.86, 0.88)	(0.86, 0.88)	0.08	.86 / 0.18
	F_1 (hold)			0.86	0.86	(0.85, 0.86)	(0.85, 0.87)	-0.04	.93 / -0.09
	F_1 (shift)			0.88	0.88	(0.88, 0.89)	(0.87, 0.89)	0.19	.69 / 0.43
	Accuracy (balanced)			0.87	0.87	(0.86, 0.88)	(0.86, 0.88)	0.00	.99 / 0.01
Before end of turn (FTO > 0 ms)	F_1 (weighted)	MM-VAP (early fusion)	MM-VAP (late fusion)	0.88	0.87	(0.87, 0.89)	(0.86, 0.88)	0.37	.46 / 0.82
	F_1 (hold)			0.86	0.86	(0.85, 0.87)	(0.85, 0.87)	0.39	.43 / 0.87
	F_1 (shift)			0.89	0.89	(0.88, 0.90)	(0.88, 0.90)	0.35	.48 / 0.78
	Accuracy (balanced)			0.88	0.87	(0.87, 0.88)	(0.86, 0.88)	0.38	.44 / 0.86

Appendix C

Extra examples



(a) Waveform, transcription and probability trace of model outputs showing the probability that the speaker depicted in red will speak in the next 1.2-2.0 sections (p_{future}).



(b) Frames illustrating participant non-verbal behaviour during the mutual silence between speaking turns, with timestamps above in seconds.

Figure C.1 Transcription and qualitative comparison MM-VAP, VAP and video-only turn-taking models on session *1f7e582c-c6bc-46b6-b5a4-e5d78e8a46ac* of the Candor corpus, along with an illustration of participant non-verbal behaviour.

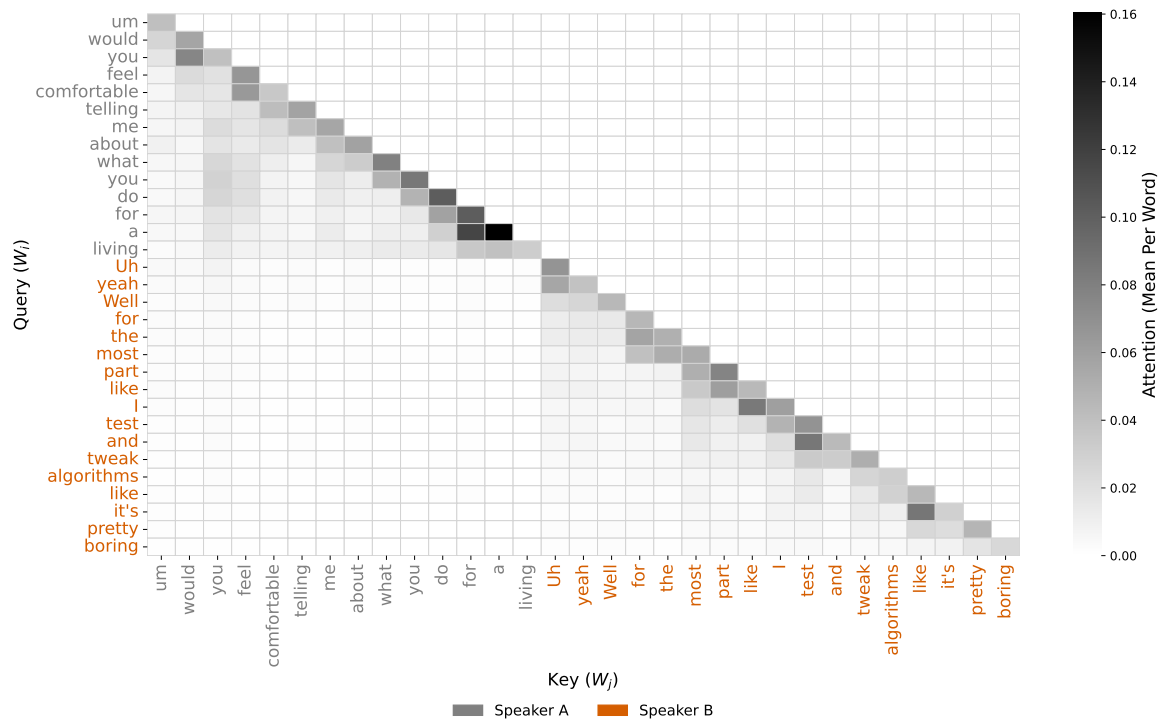


Figure C.2 MM-VAP (early fusion) attention map depicting Speaker B audio self-attention, aggregated by word

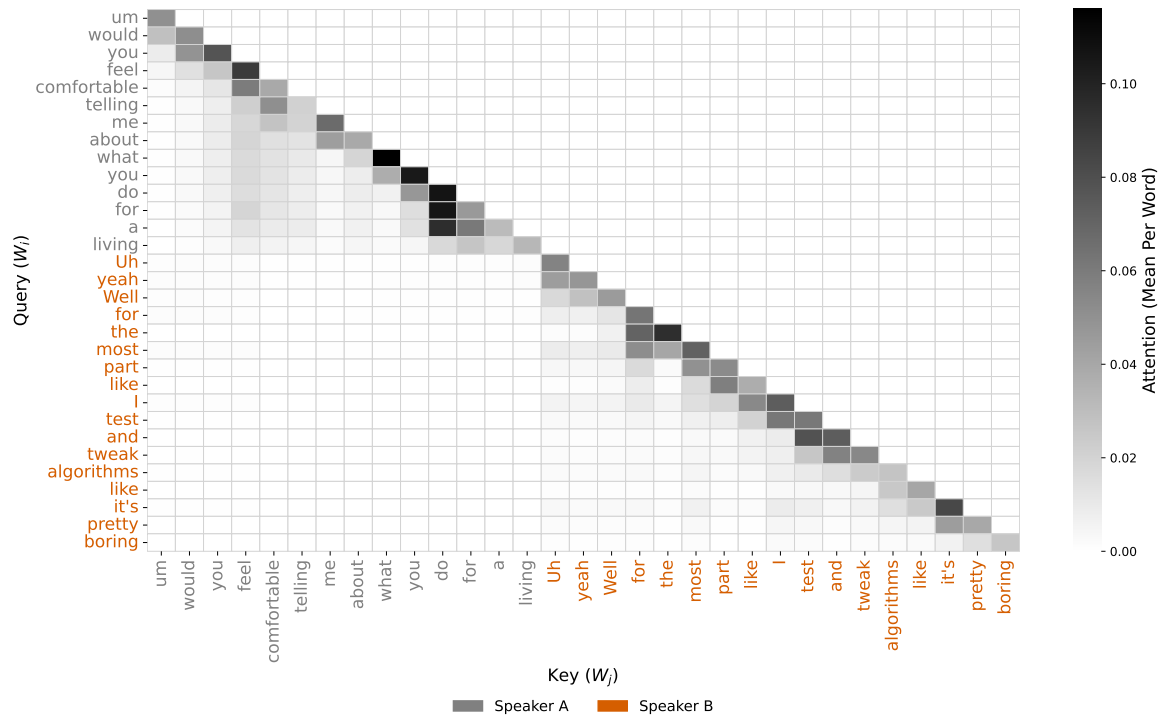


Figure C.3 MM-VAP (early fusion) attention map depicting Speaker B video self-attention, aggregated by word

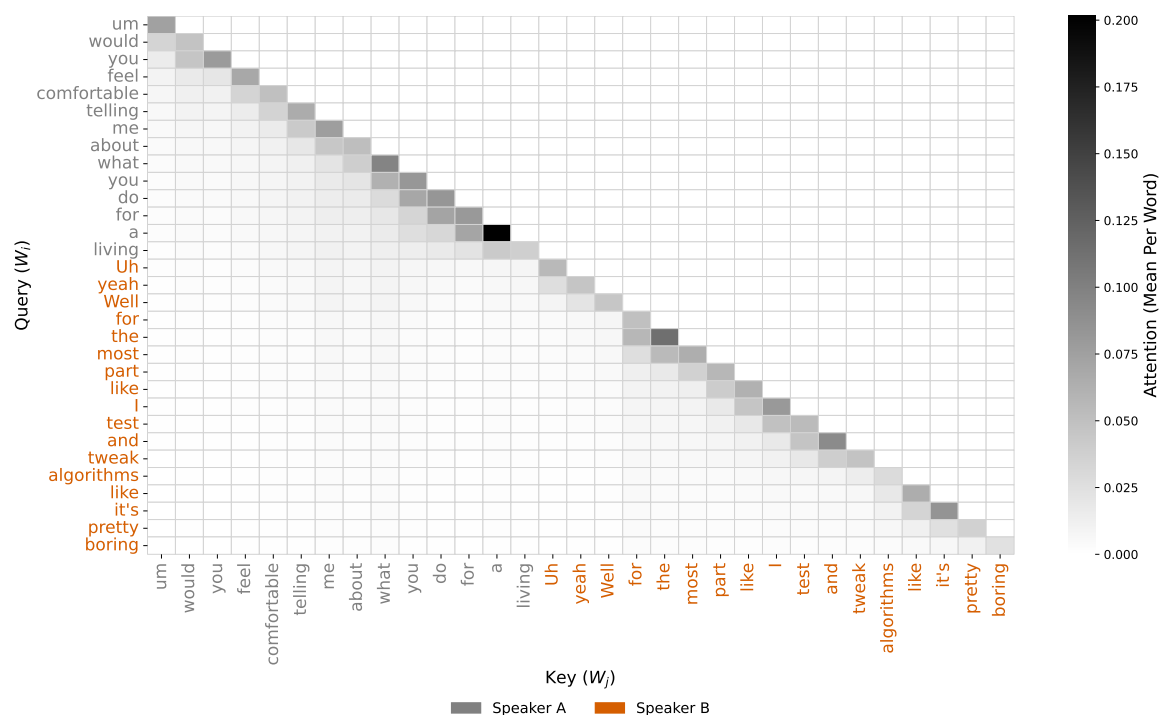


Figure C.4 MM-VAP (late fusion) attention map depicting Speaker B audio self-attention, aggregated by word

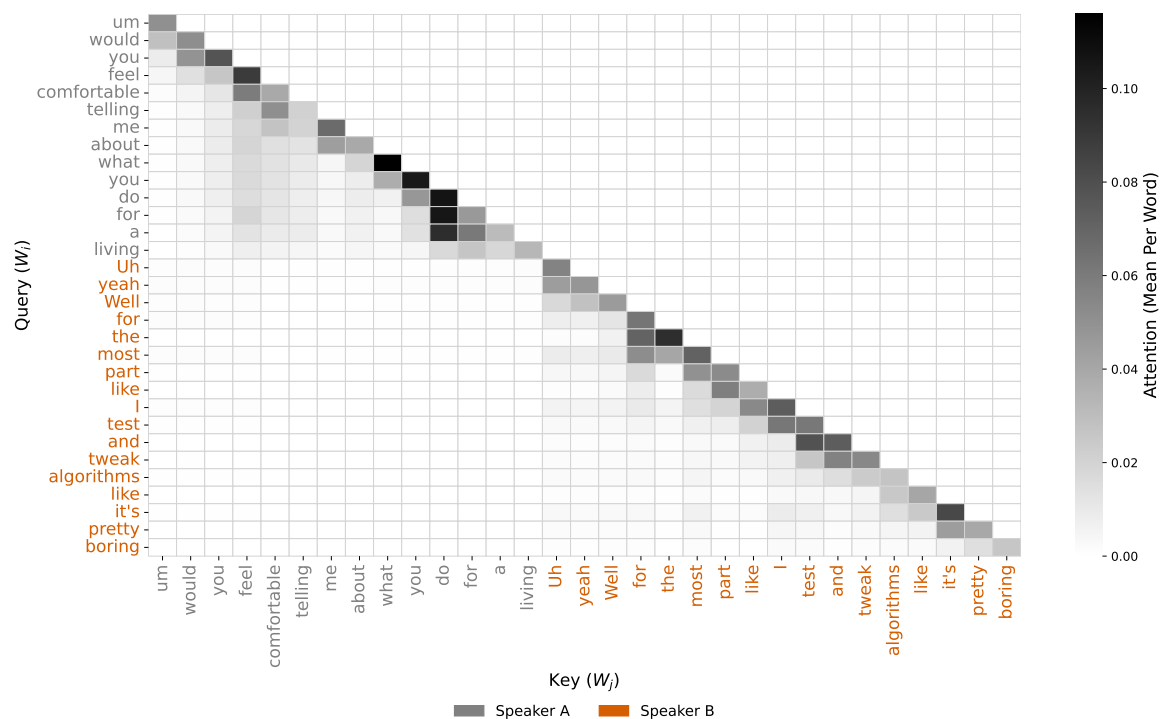


Figure C.5 MM-VAP (late fusion) attention map depicting Speaker B video self-attention, aggregated by word

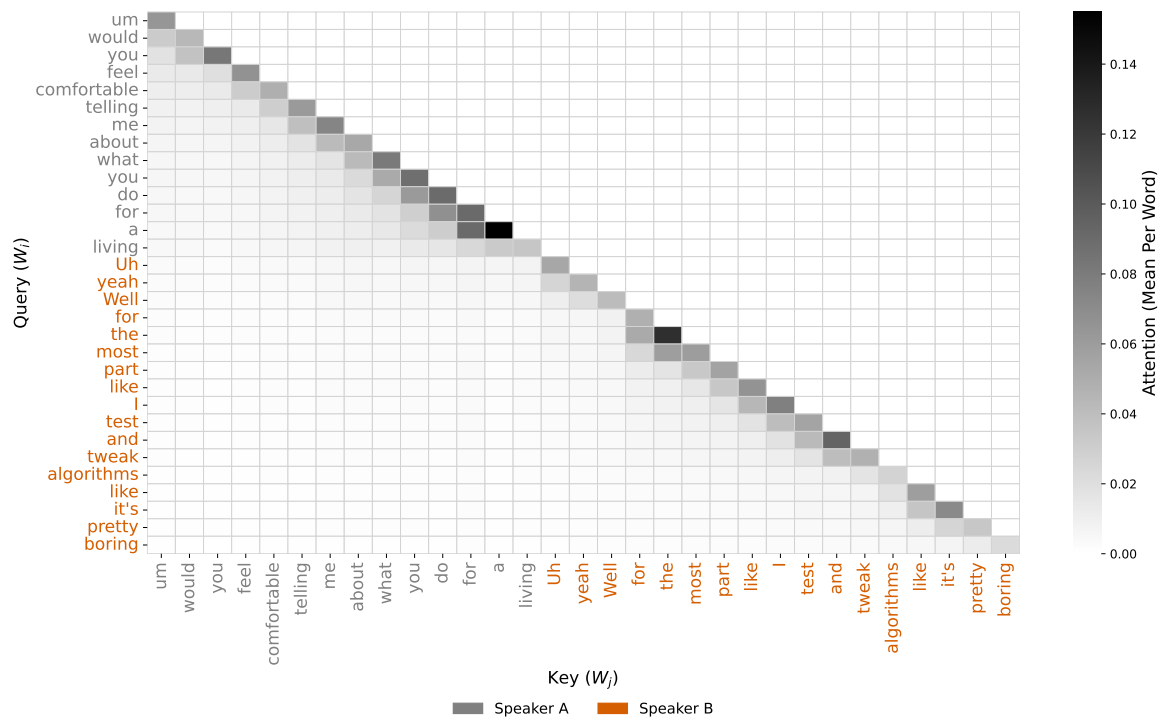


Figure C.6 MM-VAP (late fusion) attention map depicting Speaker B audio→ Speaker A audio cross-attention (layer 1, aggregated by word)

Bibliography

- Afouras, Triantafyllos et al. (2018). "LRS3-TED: a large-scale dataset for visual speech recognition". In: *arXiv preprint arXiv:1809.00496* (cited on pp. 184, 185).
- Agrawal, Purvi and Sriram Ganapathy (2019). "Modulation filter learning using deep variational networks for robust speech recognition". In: *IEEE journal of selected topics in signal processing* 13.2, pp. 244–253 (cited on p. 75).
- Agrawal et al., Vasu (2025). "Seamless Interaction: Dyadic Audiovisual Motion Modeling and Large-Scale Dataset". In: URL: <https://ai.meta.com/research/publications/seamless-interaction-dyadic-audiovisual-motion-modeling-and-large-scale-dataset/> (cited on pp. 26, 166, 218).
- Akagi, Masato and Taro Ienaga (1997). "Speaker individuality in fundamental frequency contours and its control". In: *Journal of the Acoustical Society of Japan (E)* 18.2, pp. 73–80 (cited on p. 131).
- Akhthamov, Oleg et al. (May 2020). "Using Complexity-Identical Human- and Machine-Directed Utterances to Investigate Addressee Detection for Spoken Dialogue Systems". In: *Sensors* 20.9, p. 2740. ISSN: 1424-8220. DOI: [10.3390/s20092740](https://doi.org/10.3390/s20092740). (Visited on 01/20/2023) (cited on pp. 2, 3).
- Alharbi, Sadeen et al. (2021). "Automatic Speech Recognition: Systematic Literature Review". In: *IEEE Access* 9, pp. 131858–131876. DOI: [10.1109/ACCESS.2021.3112535](https://doi.org/10.1109/ACCESS.2021.3112535) (cited on pp. 31, 61, 75, 76).
- Alnefaie, Ahlam et al. (2021). "An overview of conversational agent: applications, challenges and future directions". In: *17th International Conference on Web Information Systems and Technologies*. SCITEPRESS-Science and Technology Publications (cited on p. 3).
- Amazon, Inc. (2024). *Amazon ASR*. [Online; accessed April 2024]. URL: <https://aws.amazon.com/transcribe/> (cited on pp. 58–60).
- Anderson, Anne H (2006). "Achieving understanding in face-to-face and video-mediated multiparty interactions". In: *Discourse processes* 41.3, pp. 251–287 (cited on pp. 53, 219).
- Anderson, Anne H et al. (1991). "The HCRC map task corpus". In: *Language and speech* 34.4, pp. 351–366 (cited on pp. 23, 24, 27).
- Aubrey, Andrew J et al. (2013). "Cardiff conversation database (ccdb): A database of natural dyadic conversations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 277–282 (cited on p. 27).
- Baevski, Alexei et al. (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33, pp. 12449–12460 (cited on pp. 39, 58, 60, 76, 128).
- Baltrušaitis, Tadas, Chaitanya Ahuja, et al. (2018). "Multimodal machine learning: A survey and taxonomy". In: *IEEE transactions on pattern analysis and machine intelligence* 41.2, pp. 423–443 (cited on pp. 43, 155, 182).
- Baltrušaitis, Tadas, Peter Robinson, et al. (2016). "OpenFace: an open source facial behavior analysis toolkit". In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 1–10 (cited on pp. 35–37, 148, 210).

- Baumann, Timo et al. (2017). "Recognising Conversational Speech: What an Incremental ASR Should Do for a Dialogue System and How to Get There". In: *Dialogues with Social Robots*. Vol. 427. Singapore: Springer, pp. 421–432. doi: [10.1007/978-981-10-2585-3_35](https://doi.org/10.1007/978-981-10-2585-3_35) (cited on pp. 31, 78).
- Benus, Stefan (2009). "Variability and stability in collaborative dialogues: turn-taking and filled pauses". In: *Interspeech 2009*, pp. 796–799. doi: [10.21437/Interspeech.2009-180](https://doi.org/10.21437/Interspeech.2009-180) (cited on p. 73).
- Bilac, Miriam et al. (2017). "Gaze and filled pause detection for smooth human-robot conversations". In: *International Conference on Humanoid Robotics (Humanoids)*. IEEE, pp. 297–304. doi: [10.1109/humanoids.2017.8246889](https://doi.org/10.1109/humanoids.2017.8246889) (cited on p. 211).
- Bilakhia, Sanjay et al. (2015). "The MAHNOB Mimicry Database: A database of naturalistic human interactions". In: *Pattern recognition letters* 66, pp. 52–61 (cited on pp. 5, 25, 27).
- Biron, Tirza et al. (2025). "Disentanglement of prosodic meaning: Toward a framework for the analysis of nonverbal information in speech". In: *Proceedings of the National Academy of Sciences* 122.37, e2500510122 (cited on p. 21).
- Boersma, Paul and Vincent Van Heuven (2001). "Speak and unSpeak with PRAAT". In: *Glott International* 5.9/10, pp. 341–347 (cited on p. 62).
- Bögels, Sara and Francisco Torreira (2015). "Listeners use intonational phrase boundaries to project turn ends in spoken interaction". In: *Journal of Phonetics* 52, pp. 46–57 (cited on pp. 5, 7, 14, 18, 19, 21, 117, 119, 143, 208).
- Bokhove, Christian and Christopher Downey (2018). "Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data". In: *Methodological Innovations* 11.2, p. 2059799118790743. doi: [10.31219/osf.io/sn7w9](https://doi.org/10.31219/osf.io/sn7w9) (cited on p. 73).
- Bolker, Benjamin M (2015). "Linear and generalized linear mixed models". In: *Ecological statistics: contemporary theory and application* 2015, pp. 309–333 (cited on p. 111).
- Brown, Erik C et al. (2014). "Evaluating signal-correlated noise as a control task with language-related gamma activity on electrocorticography". In: *Clinical Neurophysiology* 125.7, pp. 1312–1323 (cited on p. 119).
- Brown, Gillian et al. (1983). *Teaching Talk*. Cambridge, U.K.: Cambridge University Press (cited on p. 23).
- Cafaro, Angelo et al. (2017). "The NoXi database: multimodal recordings of mediated novice-expert interactions". In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 350–359 (cited on pp. 25, 27).
- Calhoun, Sasha et al. (2010). "The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue". In: *Language Resources and Evaluation* 44, pp. 387–419. doi: [10.1007/s10579-010-9120-1](https://doi.org/10.1007/s10579-010-9120-1) (cited on pp. 74, 77).
- Castillo-López, Galo et al. (2025). "A Survey of Recent Advances on Turn-taking Modeling in Spoken Dialogue Systems". In: *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pp. 254–271 (cited on p. 217).
- Cavalcanti, Julio Cesar and Gabriel Skantze (2025). "'Dyadosyncrasy', Idiosyncrasy and Demographic Factors in Turn-Taking". In: *Interspeech 2025*, pp. 1093–1097. doi: [10.21437/Interspeech.2025-697](https://doi.org/10.21437/Interspeech.2025-697) (cited on pp. 15, 24).
- Cavé, Christian et al. (1996). "About the relationship between eyebrow movements and Fo variations". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Vol. 4. IEEE, pp. 2175–2178 (cited on p. 21).
- Cavé, Christian et al. (2002). "Eyebrow movements and voice variations in dialogue situations: an experimental investigation." In: *INTER_SPEECH*, pp. 2353–2356 (cited on p. 18).
- Choi, Jeongsoo et al. (2023). "Intelligible Lip-to-Speech Synthesis with Speech Units". In: *Interspeech 2023*, pp. 4349–4353. doi: [10.21437/Interspeech.2023-194](https://doi.org/10.21437/Interspeech.2023-194) (cited on pp. 122, 194).

- Christensen, Heidi et al. (2010). "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments." In: *Interspeech*, pp. 1918–1921 (cited on p. 27).
- Ciampelli, S et al. (2023). "Combining automatic speech recognition with semantic natural language processing in schizophrenia". In: *Psychiatry research* 325, p. 115252 (cited on p. 30).
- Cieri, Christopher et al. (2004). "The Fisher corpus: A resource for the next generations of speech-to-text." In: *LREC*. Vol. 4, pp. 69–71 (cited on p. 24).
- Clark, Herbert H. and Jean E. Fox Tree (2002). "Using uh and um in spontaneous speaking". In: *Cognition* 84.1, pp. 73–111. ISSN: 0010-0277. DOI: [https://doi.org/10.1016/S0010-0277\(02\)00017-3](https://doi.org/10.1016/S0010-0277(02)00017-3). URL: <https://www.sciencedirect.com/science/article/pii/S0010027702000173> (cited on p. 73).
- Clark, Leigh, Philip Doyle, et al. (2019). "The state of speech in HCI: Trends, themes and challenges". In: *Interacting with computers* 31.4, pp. 349–371 (cited on pp. 2, 3).
- Clark, Leigh, Nadia Pantidi, et al. (2019). "What makes a good conversation? Challenges in designing truly conversational agents". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–12 (cited on pp. 1, 2).
- Clift, Rebecca (2016). *Conversation analysis*. Cambridge University Press (cited on pp. 1, 59, 165).
- Cohn, Michelle and Georgia Zellou (2021). "Prosodic differences in human-and Alexa-directed speech, but similar local intelligibility adjustments". In: *Frontiers in Communication* 6, p. 675704 (cited on p. 211).
- Cole, Jennifer (2015). "Prosody in context: A review". In: *Language, Cognition and Neuroscience* 30.1-2, pp. 1–31 (cited on pp. 118, 131).
- Cutler, Anne and Mark Pearson (1985). "On the analysis of prosodic turn-taking cues". In: *Intonation in discourse*. Routledge, pp. 139–156 (cited on p. 18).
- (2018). "On the analysis of prosodic turn-taking cues". In: *Intonation in discourse*. Routledge, pp. 139–156 (cited on p. 117).
- De Ruiter, Jan-Peter et al. (2006). "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation". In: *Language* 82.3, pp. 515–535 (cited on pp. 15, 18, 19, 119, 143, 208).
- Devine, Eric G et al. (2000). "Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports". In: *Journal of the American Medical Informatics Association* 7.5, pp. 462–468. DOI: [10.1136/jamia.2000.0070462](https://doi.org/10.1136/jamia.2000.0070462) (cited on p. 72).
- Duncan, Starkey (1972). "Some signals and rules for taking speaking turns in conversations." In: *Journal of personality and social psychology* 23.2, p. 283 (cited on pp. 16, 17, 21, 49, 109, 117, 118, 131).
- Ekman, Paul and Wallace V Friesen (1978). "Facial action coding system". In: *Environmental Psychology & Nonverbal Behavior* (cited on p. 35).
- Ekstedt, Erik and Gabriel Skantze (2020). "TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2981–2990 (cited on pp. 29, 40, 51, 54, 56, 110, 132–135, 143, 171, 209).
- (2022a). "How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models". In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 541–551 (cited on pp. 25, 49, 50, 89, 119).
- (2022b). "Voice Activity Projection: Self-supervised Learning of Turn-taking Events". In: *Interspeech 2022*, pp. 5190–5194. DOI: [10.21437/Interspeech.2022-10955](https://doi.org/10.21437/Interspeech.2022-10955) (cited on pp. 5, 6, 9, 24–26, 29, 31–33, 38, 40, 44–46, 49, 54–56, 85, 87–89, 91, 93, 95, 118, 119, 190, 203, 207, 209, 214).

- Elmers, Mikey et al. (2025). "Triadic Multi-party Voice Activity Projection for Turn-taking in Spoken Dialogue Systems". In: *Interspeech 2025*, pp. 3015–3019. doi: [10.21437/Interspeech.2025-2660](https://doi.org/10.21437/Interspeech.2025-2660) (cited on p. 218).
- Emmendorfer, Alexandra K and Judith Holler (2025). "Facial signals shape predictions about the nature of upcoming conversational responses". In: *Scientific Reports* 15.1, p. 1381 (cited on p. 22).
- Eyben, Florian et al. (2015). "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE transactions on affective computing* 7.2, pp. 190–202 (cited on pp. 5, 42).
- Farrús, Mireia et al. (2007). "Jitter and shimmer measurements for speaker recognition". In: *Interspeech 2007*, pp. 778–781. doi: [10.21437/Interspeech.2007-147](https://doi.org/10.21437/Interspeech.2007-147) (cited on pp. 34, 35).
- Feng, Siyuan et al. (2024). "Towards inclusive automatic speech recognition". In: *Computer Speech & Language* 84, p. 101567 (cited on p. 75).
- Field, Anjalie et al. (2023). "Developing Speech Processing Pipelines for Police Accountability". In: pp. 1229–1233. ISSN: 2958-1796. doi: [10.21437/Interspeech.2023-2109](https://doi.org/10.21437/Interspeech.2023-2109) (cited on p. 30).
- Flecha-García, María (June 2010). "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English". In: *Speech Communication* 52, pp. 542–554. doi: [10.1016/j.specom.2009.12.003](https://doi.org/10.1016/j.specom.2009.12.003) (cited on p. 18).
- Fletcher, Janet and Jonathan Harrington (2001). "High-rising terminals and fall-rise tunes in Australian English". In: *Phonetica* 58.4, pp. 215–229 (cited on p. 118).
- Frijns, Helena Anna et al. (2023). "Communication models in human–robot interaction: an asymmetric MODEL of ALterity in human–robot interaction (AMODAL-HRI)". In: *International Journal of Social Robotics* 15.3, pp. 473–500 (cited on p. 211).
- García Molina, José Tomás et al. (2024). "Automatic Speech Recognition in Psychiatric Interviews: A Rocket to Diagnostic Support in Psychosis". In: *Revista Colombiana de Psiquiatría*. doi: [10.1016/j.rcp.2023.12.002](https://doi.org/10.1016/j.rcp.2023.12.002) (cited on pp. 57, 73).
- Gaur, Yashesh et al. (2016). "The effects of automatic speech recognition quality on human transcription latency". In: *International Web for All Conference*, pp. 1–8. doi: [10.1145/2899475.2899478](https://doi.org/10.1145/2899475.2899478) (cited on p. 73).
- Georgila, Kallirroi et al. (2020). "Evaluation of off-the-shelf speech recognizers across diverse dialogue domains". In: *International Conference on Language Resources and Evaluation (LREC)*, pp. 6469–6476 (cited on p. 57).
- Gimeno-Gómez, David et al. (2025). "Unveiling Interpretability in Self-Supervised Speech Representations for Parkinson's Diagnosis". In: *IEEE Journal of Selected Topics in Signal Processing* (cited on pp. 6, 46, 50, 55, 109, 130).
- Gobl, Christer and Ailbhe Ní Chasaide (2003). "The role of voice quality in communicating emotion, mood and attitude". In: *Speech communication* 40.1-2, pp. 189–212 (cited on p. 34).
- Godfrey, John J et al. (1992). "SWITCHBOARD: Telephone speech corpus for research and development". In: *Acoustics, speech, and signal processing, ieee international conference on*. Vol. 1. IEEE Computer Society, pp. 517–520 (cited on pp. 5, 16, 24, 27, 28, 58, 60, 61, 74, 81, 132).
- Goldman, Jean-Philippe (2011). "EasyAlign: An Automatic Phonetic Alignment Tool Under Praat". In: *Interspeech*, pp. 3233–3236. doi: [10.21437/Interspeech.2011-815](https://doi.org/10.21437/Interspeech.2011-815) (cited on pp. 59, 74).
- Gollasch, David and Gerhard Weber (2021). "Age-related differences in preferences for using voice assistants". In: *Proceedings of Mensch Und Computer 2021*, pp. 156–167 (cited on p. 2).

- Goodwin, Charles (1986). "Between and within: Alternative sequential treatments of continuers and assessments". In: *Human Studies* 9, pp. 205–217. doi: [10.1007/bf00148127](https://doi.org/10.1007/bf00148127) (cited on p. 73).
- Goodwin, Charles and John Heritage (1990). "Conversation analysis". In: *Annual review of anthropology* 19, pp. 283–307 (cited on p. 23).
- Google, Inc. (2024). *Google ASR*. [Online; accessed April 2024]. URL: <https://cloud.google.com/speech-to-text?hl=en> (cited on pp. 58–60).
- Gorisch, Jan and Thomas Schmidt (2024). "Evaluating Workflows for Creating Orthographic Transcripts for Oral Corpora by Transcribing from Scratch or Correcting ASR-Output". In: *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 6564–6574 (cited on pp. 72, 77).
- Graham, Calbert and Nathan Roll (2024). "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits". In: *JASA Express Letters* 4.2. doi: [10.1121/10.0024876](https://doi.org/10.1121/10.0024876) (cited on pp. 71, 78).
- Gravano, Agustín and Julia Hirschberg (2009). "Turn-yielding cues in task-oriented dialogue". In: *Proceedings of the SIGDIAL 2009 Conference*, pp. 253–261 (cited on pp. 5, 17, 34, 49, 111, 112, 117, 127, 131).
- (2011). "Turn-taking cues in task-oriented dialogue". In: *Computer Speech & Language* 25.3, pp. 601–634 (cited on p. 17).
- Greatbatch, David (1988). "A turn-taking system for British news interviews". In: *Language in society* 17.3, pp. 401–430 (cited on p. 25).
- Han, Jiyeon et al. (2024). "AscleAI: A LLM-based Clinical Note Management System for Enhancing Clinician Productivity". In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–7 (cited on p. 30).
- Heldner, Mattias and Jens Edlund (2010). "Pauses, gaps and overlaps in conversations". In: *Journal of Phonetics* 38.4, pp. 555–568 (cited on pp. 3, 14, 15, 17, 24, 32, 89).
- Hjalmarsson, Anna (2011). "The additive effect of turn-taking cues in human and synthetic voice". In: *Speech Communication* 53.1, pp. 23–35 (cited on pp. 17, 19, 22, 117).
- Holler, Judith, Robin H Kendrick, et al. (2016). *Turn-taking in human communicative interaction*. Frontiers Media SA (cited on pp. 2, 3, 16, 166).
- Holler, Judith and Stephen C Levinson (2019). "Multimodal language processing in human communication". In: *Trends in Cognitive Sciences* 23.8, pp. 639–652 (cited on pp. 1, 2, 166, 208).
- Horn, Berthold KP and Brian G Schunck (1981). "Determining optical flow". In: *Artificial intelligence* 17.1-3, pp. 185–203 (cited on p. 35).
- Hsu, Wei-Ning et al. (2021). "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 3451–3460. doi: [10.1109/TASLP.2021.3122291](https://doi.org/10.1109/TASLP.2021.3122291) (cited on pp. 58, 60).
- Huggingface (2025). *Huggingface*. <https://huggingface.co/>. Accessed: October 2025 (cited on pp. 128, 132).
- Huggingface (2024). https://huggingface.co/spaces/hf-audio/open_asr_leaderboard. [Online; accessed April 2024] (cited on p. 60).
- Hwang, Jeff et al. (2023). *TorchAudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for PyTorch*. arXiv: [2310.17864](https://arxiv.org/abs/2310.17864) [eess.AS] (cited on p. 84).
- Indefrey, Peter (2011). "The spatial and temporal signatures of word production components: a critical update". In: *Frontiers in psychology* 2, p. 255 (cited on pp. 3, 15).
- Indefrey, Peter and Willem JM Levelt (2004). "The spatial and temporal signatures of word production components". In: *Cognition* 92.1-2, pp. 101–144 (cited on p. 15).

- Inoue, Koji, Bing'er Jiang, et al. (May 2024). "Multilingual Turn-taking Prediction Using Voice Activity Projection". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, pp. 11873–11883. URL: <https://aclanthology.org/2024.lrec-main.1036> (cited on pp. 25, 31, 39, 46–50, 52–56, 85, 89, 155, 212).
- Inoue, Koji, Yuki Okafuji, et al. (2025). "A Noise-Robust Turn-Taking System for Real-World Dialogue Robots: A Field Experiment". In: *arXiv preprint arXiv:2503.06241* (cited on pp. 216, 217, 220).
- Isaacs, Ellen A and John C Tang (1993). "What video can and can't do for collaboration: a case study". In: *ACM International Conference on Multimedia*, pp. 199–206. doi: [10.1016/b978-155860651-7/50134-0](https://doi.org/10.1016/b978-155860651-7/50134-0) (cited on p. 74).
- Ishii, Ryo et al. (2025). "Predicting End-of-turn and Backchannel Based on Multimodal Voice Activity Prediction Model". In: *Proceedings of the 27th International Conference on Multimodal Interaction*, pp. 446–455 (cited on p. 167).
- Jadoul, Yannick et al. (2018). "Introducing Parselmouth: A Python interface to Praat". In: *Journal of Phonetics* 71, pp. 1–15. doi: <https://doi.org/10.1016/j.wocn.2018.07.001> (cited on pp. 34, 111).
- Jiang, Bing'er et al. (2023). "What makes a good pause? Investigating the turn-holding effects of fillers". In: *20th International Congress of Phonetic Sciences (ICPhS). August 7-11 2023, Prague, Czech Republic*. International Phonetic Association, pp. 3512–3516 (cited on p. 49).
- Jiang, Bing'er et al. (2023). "Response-conditioned Turn-taking Prediction". In: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12241–12248 (cited on pp. 51, 56, 143).
- Jongman, Allard et al. (2000). "Acoustic characteristics of English fricatives". In: *The Journal of the Acoustical Society of America* 108.3, pp. 1252–1263 (cited on p. 121).
- Jurafsky, Daniel and James H. Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 1st. USA: Prentice Hall PTR. ISBN: 0130950696 (cited on p. 62).
- Kashefi, Arian et al. (2024). "Unlocking the Potential of mHealth for Smoking Cessation: An Expert Viewpoint". In: *International Conference on Human-Computer Interaction*. Springer, pp. 59–79 (cited on p. 30).
- Kendon, Adam (1967). "Some functions of gaze-direction in social interaction". In: *Acta psychologica* 26, pp. 22–63 (cited on pp. 5, 17, 101, 164).
- Kendrick, Robin H and Francisco Torreira (2015). "The timing and construction of preference: A quantitative study". In: *Discourse Processes* 52.4, pp. 255–289 (cited on pp. 1, 43).
- Klis, Anika van der et al. (2023). "Using Open-Source Automatic Speech Recognition Tools for the Annotation of Dutch Infant-Directed Speech". In: *Multimodal Technologies and Interaction* 7.7, p. 68. doi: [10.3390/mti7070068](https://doi.org/10.3390/mti7070068) (cited on p. 78).
- Knapp, Mark L et al. (1978). *Nonverbal communication in human interaction*. Vol. 1. Holt, Rinehart and Winston New York (cited on pp. 1, 74).
- Koutsombogera, Maria and Carl Vogel (2018). "Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (cited on pp. 24, 27).
- Kraaij, Wessel et al. (2005). "The AMI meeting corpus". In: *Proc. International Conference on Methods and Techniques in Behavioral Research*, pp. 1–4 (cited on pp. 24, 27).
- Kuznetsova, Alexandra et al. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13, pp. 1–26. doi: [10.18637/jss.v082.i13](https://doi.org/10.18637/jss.v082.i13) (cited on p. 112).

- Lai, Catherine (2014). "Final Rises in task-oriented and conversational dialogue". In: *Speech Prosody 2014*. ISCA, pp. 520–524 (cited on p. 118).
- Lameris, Harm et al. (2025). "VoiceQualityVC: A voice conversion system for studying the perceptual effects of voice quality in speech". In: *Proc. Interspeech*, pp. 2295–2299 (cited on p. 34).
- Latif, Kareem et al. (2023). "'Mm-hm,' 'Uh-uh': Are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology?" In: *Journal of the American Medical Informatics Association* 30.4, pp. 703–711. doi: [10.1093/jamia/ocad001](https://doi.org/10.1093/jamia/ocad001) (cited on pp. 29–31, 57, 66, 72, 76).
- Latif, Nida et al. (2018). "Knowing when to respond: The role of visual information in conversational turn exchanges". In: *Attention, Perception, & Psychophysics* 80.1, pp. 27–41 (cited on pp. 14, 19).
- Lee, Minhyeok (2023). "Mathematical analysis and performance evaluation of the gelu activation function in deep learning". In: *Journal of Mathematics* 2023.1, p. 4229924 (cited on p. 157).
- Leishman, Sean et al. (2024). "Pairwiseturngpt: a multi-stream turn prediction model for spoken dialogue". In: *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue* (cited on pp. 5, 29, 38–40, 51, 52, 54, 56, 110, 132–135, 139, 140, 143, 182, 200–203, 207).
- Lennon, Daire et al. (2006). "A hmm framework for motion based parsing for video from observational psychology". In: *IEEE Irish Machine Vision and Image Processing Conference* (cited on p. 35).
- Levinson, Stephen C and Francisco Torreira (2015). "Timing in turn-taking and its implications for processing models of language". In: *Frontiers in psychology* 6, p. 731 (cited on pp. 2, 14, 16, 89, 101).
- Li, Siyan et al. (2022). "When can I speak? Predicting initiation points for spoken dialogue agents". In: *arXiv preprint arXiv:2208.03812* (cited on pp. 3, 26, 41–43, 50, 51, 54–56).
- Liesenfeld, Andreas et al. (2023). "The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems". In: *Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 482–495. doi: [10.18653/v1/2023.sigdial-1.45](https://doi.org/10.18653/v1/2023.sigdial-1.45) (cited on pp. 31, 65, 71).
- Lin, Zhaofeng and Naomi Harte (2025). "Uncovering the Visual Contribution in Audio-Visual Speech Recognition". In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (cited on p. 193).
- Lippmann, Richard P. (1997). "Speech recognition by machines and humans". In: *Speech Communication* 22.1, pp. 1–15. doi: [10.1016/S0167-6393\(97\)00021-6](https://doi.org/10.1016/S0167-6393(97)00021-6) (cited on p. 31).
- Litman, Diane et al. (2016). "The teams corpus and entrainment in multi-party spoken dialogues". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1421–1431 (cited on p. 27).
- Loakes, Debbie (2022). "Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes?" In: *Frontiers in Communication* 7, p. 803452 (cited on pp. 31, 57).
- (2024). "Automatic speech recognition and the transcription of indistinct forensic audio: How do the new generation of systems fare?" In: *Frontiers in Communication* 9, p. 1281407 (cited on pp. 29, 30, 76).
- Lopez, Alianda et al. (2022). "Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English and German: What Goes Missing?" In: *Conference on Natural Language Processing (KONVENS)*, pp. 135–143 (cited on pp. 31, 65, 71).
- Loshchilov, Ilya and Frank Hutter (2017). "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (cited on p. 84).

- Lugaresi, Camillo et al. (2019). "Mediapipe: A framework for building perception pipelines". In: *arXiv preprint arXiv:1906.08172* (cited on p. 35).
- Ma, Marcus et al. (2024). "Introducing Bed Word: a new automated speech recognition tool for sociolinguistic interview transcription". In: *Linguistics Vanguard*. doi: [10.1515/lingvan-2023-0073](https://doi.org/10.1515/lingvan-2023-0073) (cited on p. 73).
- MacLeod, Alison and Quentin Summerfield (1987). "Quantifying the contribution of vision to speech perception in noise". In: *British journal of audiology* 21.2, pp. 131–141 (cited on pp. 7, 215).
- Malik, Usman et al. (Nov. 2020). "Who Speaks Next? Turn Change and Next Speaker Prediction in Multimodal Multiparty Interaction". en. In: *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. Baltimore, MD, USA: IEEE, pp. 349–354. ISBN: 978-1-72819-228-4. doi: [10.1109/ICTAI50040.2020.00062](https://doi.org/10.1109/ICTAI50040.2020.00062). URL: <https://ieeexplore.ieee.org/document/9288220/> (visited on 07/07/2022) (cited on pp. 53, 218).
- Mann, Henry B and Donald R Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other". In: *The annals of mathematical statistics*, pp. 50–60 (cited on p. 149).
- Marxer, Ricard et al. (2018). "The impact of the Lombard effect on audio and visual speech recognition systems". In: *Speech communication* 100, pp. 58–68 (cited on p. 193).
- Mateer, Marie and Ann Taylor (1995). *Dysfluency Annotation Stylebook for the Switchboard Corpus*. Department of Computer and Information Science, University of Pennsylvania (cited on p. 74).
- McAuliffe, Michael et al. (2017). "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi". In: *Interspeech*, pp. 498–502. doi: [10.21437/Interspeech.2017-1386](https://doi.org/10.21437/Interspeech.2017-1386) (cited on pp. 74, 209).
- Microsoft, Inc. (2024). *Mircorsoft Azure*. [Online; accessed April 2024]. URL: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio&trk=article-ssr-frontend-pulse_little-text-block/ (cited on pp. 31, 62, 70).
- Morbini, Fabrizio et al. (2013). "Which ASR should I choose for my dialogue system?" In: *Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 394–403 (cited on pp. 31, 57, 70, 72).
- Morise, Masanori et al. (2016). "World: a vocoder-based high-quality speech synthesis system for real-time applications". In: *IEICE TRANSACTIONS on Information and Systems* 99.7, pp. 1877–1884 (cited on pp. 110, 120).
- Müller, Philipp, Michael Xuelin Huang, et al. (2018). "Detecting low rapport during natural interactions in small groups from non-verbal behaviour". In: *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pp. 153–164 (cited on p. 27).
- Müller, Philipp, Ekta Sood, et al. (2020). "Anticipating averted gaze in dyadic interactions". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–10 (cited on pp. 25, 27).
- Muradás-Taylor, Becky (2022). "Accuracy and stability in English speakers' production of Japanese pitch accent". In: *Language and Speech* 65.2, pp. 377–403 (cited on p. 118).
- Mushtaq, Faizah et al. (2019). "Evaluating time-reversed speech and signal-correlated noise as auditory baselines for isolating speech-specific processing using fNIRS". In: *PLoS One* 14.7 (cited on p. 119).
- Negroponte, Nicholas (1995). *Being Digital*. 1. publ. London: Hodder & Stoughton. ISBN: 978-0-340-64525-3 (cited on p. 219).
- Nguyen, Thai-Binh et al. (2025). "Cocktail-Party Audio-Visual Speech Recognition". In: *Interspeech 2025*, pp. 1828–1832. doi: [10.21437/Interspeech.2025-676](https://doi.org/10.21437/Interspeech.2025-676) (cited on p. 218).
- Nomikou, Iris (2023). "Joining actions through effort sounds: Mothers and infants in routine activities". In: *Language & Communication* 91, pp. 32–45 (cited on p. 23).

- Nota, Naomi et al. (2021). "Facial signals and social actions in multimodal face-to-face interaction". In: *Brain Sciences* 11.8, p. 1017 (cited on pp. 18, 19, 153).
- (2023). "Conversational eyebrow frowns facilitate question identification: An online study using virtual avatars". In: *Cognitive Science* 47.12, e13392 (cited on pp. 1, 14, 19).
- O'Conaill, Brid et al. (Dec. 1993). "Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication". en. In: *Human-Computer Interaction* 8.4, pp. 389–428. ISSN: 0737-0024, 1532-7051. DOI: [10.1207/s15327051hci0804_4](https://doi.org/10.1207/s15327051hci0804_4). URL: http://www.tandfonline.com/doi/abs/10.1207/s15327051hci0804_4 (visited on 05/24/2022) (cited on pp. 53, 71, 101, 166).
- O'Connor Russell, Sam, Delphine Chaurau, et al. (2026). "The Role of Prosodic and Lexical Cues in Turn-Taking with Self-Supervised Speech Representations". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Accepted for publication. Barcelona, Spain (cited on p. 12).
- O'Connor Russell, Sam and Naomi Harte (2024). "Towards Multimodal Turn-taking for Naturalistic Human-Robot Interaction". In: *Second International Multimodal Communication Symposium (MMSYM), Frankfurt, Germany*, pp. 91–92. URL: <http://arxiv.org/abs/2505.21043> (cited on p. 11).
- (2025a). "Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction". In: *The 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025, Vienna, Austria*. URL: <http://arxiv.org/abs/2505.21043> (cited on pp. 11, 217).
- (2025b). "Visual Cues Support Robust Turn-taking Prediction in Noise". In: *Interspeech 2025*, pp. 1073–1077. DOI: [10.21437/Interspeech.2025-668](https://doi.org/10.21437/Interspeech.2025-668) (cited on p. 12).
- O'Mahony, Johannah et al. (2023). "Synthesising turn-taking cues using natural conversational data". In: *12th ISCA Speech Synthesis Workshop*. ISCA, pp. 75–80 (cited on p. 220).
- O'Connor Russell, Sam et al. (2024). "What automatic speech recognition can and cannot do for conversational speech transcription". In: *Research Methods in Applied Linguistics* 3.3, p. 100163 (cited on pp. 11, 62).
- Oliveira, Miguel and Tiago Freitas (2008). "Intonation as a cue to turn management in telephone and face-to-face interactions". In: *Speech Prosody 2008*, p. 485 (cited on p. 106).
- Onishi, Kazuyo et al. (2023). "Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation". In: *Proceedings of the 11th International Conference on Human-Agent Interaction*, pp. 13–21 (cited on p. 217).
- (2024). "Multimodal Voice Activity Projection for Turn-taking and Effects on Speaker Adaptation". In: *IEICE Transactions on Information and Systems* (cited on pp. 25, 52, 167, 181, 217).
- Oxenham, Andrew J (2012). "Pitch perception". In: *Journal of Neuroscience* 32.39, pp. 13335–13338 (cited on pp. 34, 111).
- Panayotov, Vassil et al. (2015a). "Librispeech: An ASR corpus based on public domain audio books". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5206–5210. DOI: [10.1109/ICASSP.2015.7178964](https://doi.org/10.1109/ICASSP.2015.7178964) (cited on pp. 58, 60, 61).
- (2015b). "Librispeech: an asr corpus based on public domain audio books". In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5206–5210 (cited on pp. 46, 61).
- Papadopoulou, Martha Maria et al. (2021). "Benchmarking ASR systems based on post-editing effort and error analysis". In: *Translation and Interpreting Technology Online Conference*, pp. 199–207 (cited on pp. 73, 77).
- Park, Tae Jin et al. (2022). "A review of speaker diarization: Recent advances with deep learning". In: *Computer Speech & Language* 72, p. 101317 (cited on pp. 71, 75, 76).
- Parks, Elizabeth S and Kiah E Bennett (2024). "A transcription system for nonverbal listening behavior". In: *International Journal of Listening* 38.1, pp. 41–57 (cited on p. 30).

- Parsons, Phoebe et al. (2025). "Effects of Prosodic Information on Dialect Classification Using Whisper Features". In: *Interspeech 2025*, 2785–2789. doi: {10.21437/Interspeech.2025-200} (cited on pp. 51, 119).
- Pasad, Ankita et al. (2021). "Layer-wise analysis of a self-supervised speech representation model". In: *ASRU 2021*. IEEE (cited on pp. 6, 46, 49–51, 55, 130, 143).
- Paszke, Adam et al. (2017). "Automatic differentiation in PyTorch". In: (cited on p. 83).
- Peterson, Gordon E and Harold L Barney (1952). "Control methods used in a study of the vowels". In: *The Journal of the acoustical society of America* 24.2, pp. 175–184 (cited on p. 121).
- Petukhova, Volha and Harry Bunt (2009). "Who's next? Speaker-selection mechanisms in multiparty dialogue". In: *Workshop on the Semantics and Pragmatics of Dialogue* (cited on pp. 23, 219).
- Press, Ofir et al. (2021). "Train short, test long: Attention with linear biases enables input length extrapolation". In: *arXiv preprint arXiv:2108.12409* (cited on pp. 48, 157).
- Psathas, George and Timothy Anderson (1990). "The 'practices' of transcription in conversation analysis". In: *Semiotica* 78.1-2, pp. 75–100. doi: 10.1515/semi.1990.78.1-2.75 (cited on pp. 30, 59).
- PyWorld (2025). *PyWorld*. <https://pypi.org/project/pyworld/>. Accessed: October 2025 (cited on p. 120).
- Qiao, Sen et al. (2017). "Real-time human gesture grading based on OpenPose". In: *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, pp. 1–6 (cited on p. 35).
- Radford, Alec, Jong Wook Kim, et al. (2023). "Robust speech recognition via large-scale weak supervision". In: *International conference on machine learning*. PMLR, pp. 28492–28518 (cited on pp. 30, 31, 58–60, 77, 122, 187).
- Radford, Alec, Jeffrey Wu, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9 (cited on pp. 40, 43, 51, 52, 132, 200).
- Raiaan, Mohaimenul Azam Khan et al. (2024). "A review on large language models: Architectures, applications, taxonomies, open issues and challenges". In: *IEEE access* 12, pp. 26839–26874 (cited on p. 38).
- Rayson, Paul and Roger Garside (2000). "Comparing corpora using frequency profiling". In: *Workshop on Comparing Corpora held in conjunction with the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1–6 (cited on p. 64).
- Reece, Andrew et al. (2023). "The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation". In: *Science Advances* 9.13, eadf3197 (cited on pp. 5, 25–27, 30, 55, 75, 81, 214).
- Reverdy, Justine et al. (2022). "Roomreader: A multimodal corpus of online multiparty conversational interactions". In: *International Conference on Language Resources and Evaluation (LREC)*, pp. 2517–2527 (cited on pp. 23, 24, 26, 27, 30, 58, 60, 61, 64, 71, 77).
- Riviere, Morgane et al. (2020). "Unsupervised pretraining transfers well across languages". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7414–7418 (cited on pp. 39, 45, 46, 50, 83, 119, 128, 154, 157, 196).
- Roberts, SeÅin G. et al. (May 2015). "The effects of processing and sequence organization on the timing of turn taking: a corpus study". en. In: *Frontiers in Psychology* 6. ISSN: 1664-1078. doi: 10.3389/fpsyg.2015.00509. URL: http://www.frontiersin.org/Language_Sciences/10.3389/fpsyg.2015.00509/abstract (visited on 12/05/2025) (cited on p. 15).
- Roddy, Matthew and Naomi Harte (July 2020). "Neural Generation of Dialogue Response Timings". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics,

- pp. 2442–2452. DOI: [10.18653/v1/2020.acl-main.221](https://doi.org/10.18653/v1/2020.acl-main.221). URL: <https://aclanthology.org/2020.acl-main.221> (cited on pp. 24, 25, 29, 31, 38, 41, 43, 44, 50, 51, 54, 56, 208).
- Roddy, Matthew, Gabriel Skantze, et al. (2018a). “Investigating speech features for continuous turn-taking prediction using lstms”. In: *arXiv preprint arXiv:1806.11461* (cited on pp. 5, 24, 26, 32, 33, 42, 43, 46, 50, 52, 55, 56, 87, 89).
- (2018b). “Multimodal continuous turn-taking prediction using multiscale rnns”. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 186–190 (cited on pp. 24, 25, 30, 43, 54, 56, 166, 214).
- Ruan, Sherry et al. (2016). “Speech is 3x faster than typing for english and mandarin text entry on mobile devices”. In: *arXiv preprint arXiv:1608.07323* (cited on p. 2).
- (2018). “Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4, pp. 1–23 (cited on p. 2).
- Ruhi, Şükriye and Thomas Schmidt (2014). *Best practices for spoken corpora in linguistic research*. Cambridge Scholars Publishing (cited on p. 23).
- Rzepka, Christine et al. (2022). “Voice assistant vs. Chatbot—examining the fit between conversational agents’ interaction modalities and information search tasks”. In: *Information Systems Frontiers* 24.3, pp. 839–856 (cited on p. 2).
- Sacks, Harvey et al. (1974). “A simplest systematics for the organization of turn-taking for conversation”. In: *Language* 50.4, pp. 696–735 (cited on pp. 1–3, 13, 14, 89, 91, 160).
- Saga, Takeshi and Catherine Pelachaud (2025). “Voice Activity Projection Model with Multimodal Encoders”. In: *arXiv e-prints*, arXiv–2506 (cited on p. 166).
- Sanabria, Ramon et al. (2023). “The Edinburgh international accents of English corpus: Towards the democratization of English ASR”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (cited on pp. 31, 75).
- Schaffer, Deborah (1983). “The role of intonation as a cue to turn taking in conversation”. In: *Journal of phonetics* 11.3, pp. 243–257 (cited on pp. 117, 119, 127, 131).
- Schegloff, Emanuel A (2000a). “Overlapping talk and the organization of turn-taking for conversation”. In: *Language in society* 29.1, pp. 1–63 (cited on p. 1).
- (2000b). “When ‘others’ initiate repair”. In: *Applied Linguistics* 21.2, pp. 205–243 (cited on pp. 30, 57).
- Schneider, Steffen et al. (2019). “wav2vec: Unsupervised pre-training for speech recognition”. In: *arXiv preprint arXiv:1904.05862* (cited on p. 43).
- Sellen, Abigail (Dec. 1995). “Remote Conversations: The Effects of Mediating Talk With Technology”. en. In: *Human-Computer Interaction* 10.4, pp. 401–444. ISSN: 0737-0024. DOI: [10.1207/s15327051hci1004_2](https://doi.org/10.1207/s15327051hci1004_2). URL: http://www.tandfonline.com/doi/abs/10.1207/s15327051hci1004_2 (visited on 05/24/2022) (cited on pp. 70, 101, 166).
- Sennrich, Rico et al. (2016). “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725 (cited on pp. 51, 132).
- Shaw, Sara E et al. (2020). “Video consultations between patients and clinicians in diabetes, cancer, and heart failure services: linguistic ethnographic study of video-mediated interaction”. In: *Journal of medical Internet research* 22.5, e18378 (cited on p. 23).
- Shi, Bowen et al. (2022). “Robust self-supervised audio-visual speech recognition”. In: *arXiv preprint arXiv:2201.01763* (cited on pp. 30, 35, 185, 186, 193, 219).
- Skantze, Gabriel (2017). “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*, pp. 220–230 (cited on pp. 5, 24, 32, 38, 41–43, 49, 50, 52, 55, 56, 88).

- Skantze, Gabriel (2021). "Turn-taking in conversational systems and human-robot interaction: a review". In: *Computer Speech and Language* 67, p. 101178 (cited on pp. 1–4, 8, 15, 17, 41, 42, 55, 95, 101, 106, 107, 181, 216).
- Skantze, Gabriel, Anna Hjalmarsson, et al. (2014). "Turn-taking, feedback and joint attention in situated human–robot interaction". In: *Speech Communication* 65, pp. 50–66. ISSN: 0167-6393. DOI: <https://doi.org/10.1016/j.specom.2014.05.005>. URL: <https://www.sciencedirect.com/science/article/pii/S016763931400051X> (cited on p. 220).
- Skantze, Gabriel and Bahar Irfan (2025). "Applying general turn-taking models to conversational human-robot interaction". In: *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 859–868 (cited on pp. 16, 55, 106, 143, 209–211, 216, 219, 220).
- Snyder, David et al. (2015). "Musan: A music, speech, and noise corpus". In: *arXiv preprint arXiv:1510.08484* (cited on pp. 184, 185, 217).
- Sørensen, A. Josefine Munch et al. (2020). "Effects of noise and L2 on the timing of turn taking in conversation". English. In: vol. 7. *Proceedings of the International Symposium on Audiological and Auditory Research*, pp. 86–92 (cited on p. 193).
- Speechmatics, Ltd. (2024). *Speechmatics ASR*. [Online; accessed Dec 2024]. URL: <https://www.speechmatics.com/company/articles-and-news/introducing-ursa-the-worlds-most-accurate-speech-to-text> (cited on pp. 30, 31, 57–60).
- Srivastava, Nitish et al. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html> (cited on p. 157).
- Sterpu, George and Naomi Harte (2017). "Towards Lipreading Sentences with Active Appearance Models". In: *The 14th International Conference on Auditory-Visual Speech Processing*, pp. 70–75. DOI: [10.21437/AVSP.2017-14](https://doi.org/10.21437/AVSP.2017-14) (cited on p. 35).
- Sterpu, George, Christian Saam, et al. (2020). "How to teach DNNs to pay attention to the visual modality in speech recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, pp. 1052–1064 (cited on p. 193).
- Stevens, Kenneth N and Sheila E Blumstein (1978). "Invariant cues for place of articulation in stop consonants". In: *The Journal of the Acoustical Society of America* 64.5, pp. 1358–1368 (cited on p. 121).
- Stivers, Tanya et al. (2009). "Universals and cultural variation in turn-taking in conversation". In: *Proceedings of the National Academy of Sciences* 106.26, pp. 10587–10592 (cited on pp. 2, 3, 15, 24).
- Stolcke, Andreas and Jasha Droppo (2017). "Comparing Human and Machine Errors in Conversational Speech Transcription". In: *Interspeech*, pp. 137–141. DOI: [10.21437/Interspeech.2017-1544](https://doi.org/10.21437/Interspeech.2017-1544) (cited on pp. 31, 57, 63, 71).
- Tabassum, Madiha et al. (2019). "Investigating users' preferences and expectations for always-listening voice assistants". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.4, pp. 1–23 (cited on p. 2).
- Tian, Ying-li et al. (2001). "Recognizing Action Units for Facial Expression Analysis". en. In: vol. 23. 2, p. 19. DOI: [10.1109/34.908962](https://doi.org/10.1109/34.908962) (cited on p. 37).
- Tong, Yan et al. (2007). "Facial action unit recognition by exploiting their dynamic and semantic relationships". In: *IEEE transactions on pattern analysis and machine intelligence* 29.10, pp. 1683–1699 (cited on p. 153).
- Trujillo, James et al. (2021). "Speakers exhibit a multimodal Lombard effect in noise". In: *Scientific Reports* 11.1, p. 16721 (cited on p. 193).
- Trujillo, James P, Stephen C Levinson, et al. (2022). "A multi-scale investigation of the human communication system's response to visual disruption". In: *Royal Society Open Science* 9.4, p. 211489 (cited on pp. 21, 22).

- Trujillo, James P, Julija Vaitonyte, et al. (2019). "Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research". In: *Behavior research methods* 51.2, pp. 769–777 (cited on p. 35).
- Turnbull, William (2003). *Language in action: Psychological models of conversation*. Routledge (cited on pp. 1, 2).
- Umair, Muhammad et al. (Apr. 2022). "GailBot: An automatic transcription system for Conversation Analysis". In: *Dialogue & Discourse* 13.1, pp. 63–95. doi: [10.5210/dad.2022.103](https://doi.org/10.5210/dad.2022.103) (cited on pp. 29, 30, 57, 73).
- Uppenkamp, Stefan and Markus Röhl (2014). "Human auditory neuroimaging of intensity and loudness". In: *Hearing Research* 307, pp. 65–73 (cited on p. 34).
- Valstar, Michel et al. (2017). "Facial Actions as Social Signals". en. In: *Social Signal Processing*. Ed. by Judee K. Burgoon et al. Cambridge: Cambridge University Press, pp. 123–154. ISBN: 978-1-316-67620-2. doi: [10.1017/9781316676202.011](https://doi.org/10.1017/9781316676202.011). URL: https://www.cambridge.org/core/product/identifier/CBO9781316676202A021/type/book_part (visited on 03/15/2022) (cited on p. 37).
- Vaswani, A (2017). "s". In: *Advances in Neural Information Processing Systems* (cited on pp. 38, 39, 60, 139, 200).
- Vipperla, Ravichander et al. (2008). "Longitudinal study of ASR performance on ageing voices". In: *Interspeech 2008*, pp. 2550–2553. doi: [10.21437/Interspeech.2008-632](https://doi.org/10.21437/Interspeech.2008-632) (cited on p. 75).
- Walker, Michael B and Carmelina Trimboli (1982). "Smooth transitions in conversational interactions". In: *The Journal of Social Psychology* 117.2, pp. 305–306 (cited on pp. 15, 32, 88).
- Ward, Nigel (2006). "Non-lexical conversational sounds in American English". In: *Pragmatics & Cognition* 14.1, pp. 129–182 (cited on pp. 30, 77).
- Wennerstrom, Ann and Andrew F Siegel (2003). "Keeping the floor in multiparty conversations: Intonation, syntax, and pause". In: *Discourse Processes* 36.2, pp. 77–107 (cited on pp. 17, 49).
- Widodo, Handoyo Puji (2014). "Methodological considerations in interview data transcription". In: *International Journal of Innovation in English Language Teaching and Research* 3.1, pp. 101–107 (cited on p. 30).
- Wollin-Giering, Susanne et al. (Jan. 2024). "Automatic Transcription of English and German Qualitative Interviews". In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 25.1. doi: [10.17169/fqs-25.1.4129](https://doi.org/10.17169/fqs-25.1.4129). URL: <https://www.qualitative-research.net/index.php/fqs/article/view/4129> (cited on p. 30).
- Wong, Deanna and Pam Peters (2007). "A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification". In: *International Journal of Corpus Linguistics* 12.4, pp. 479–510. doi: [10.1075/ijcl.12.4.03won](https://doi.org/10.1075/ijcl.12.4.03won) (cited on pp. 77, 78).
- Woodruff, Allison and Paul M Aoki (2003). "How push-to-talk makes talk less pushy". In: *Proceedings of the 2003 ACM International Conference on Supporting Group Work*, pp. 170–179 (cited on p. 3).
- Wynn, Camille J et al. (2022). "Rhythm perception, speaking rate entrainment, and conversational quality: A mediated model". In: *Journal of Speech, Language, and Hearing Research* 65.6, pp. 2187–2203 (cited on p. 118).
- Yang, Guojun and Jafar Saniie (2016). "Eye tracking using monocular camera for gaze estimation applications". In: *2016 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, pp. 0292–0296 (cited on p. 165).
- Yang, Yuchen (2023). "Write What You Want: Applying Text-to-Video Retrieval to Audiovisual Archives". In: *ACM Journal on Computing and Cultural Heritage* 16.4, pp. 1–16 (cited on p. 30).

- Yu, Dong and Li Deng (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Signals and Communication Technology. London: Springer London. doi: [10.1007/978-1-4471-5779-3](https://doi.org/10.1007/978-1-4471-5779-3) (cited on p. 76).