PairwiseTurnGPT: a multi-stream turn prediction model for spoken dialogue

Sean Leishman and **Peter Bell** and **Sarenne Wallbridge**
University of Edinburgh, UK
{s2051283, peter.bell, s.wallbridge}@ed.ac.uk

Abstract

Spoken conversation is characterised by rapid turn transitions and frequent speaker overlaps. However, existing models of turn-taking treat dialogue as a series of incremental turns. We propose PairwiseTurnGPT, a language model that captures the temporal dynamics of lexical content by modelling dialogue as two aligned speaker streams. PairwiseTurnGPT provides a much more nuanced understanding of how lexical content contributes to predicting turn-taking behaviour in speech. By training the model with data configurations containing different turn-taking behaviours, we demonstrate the relative contributions of partial, complete, and backchannel overlaps for accurately predicting the variety of turn ends that occur in spoken dialogue. We also show that PairwiseTurnGPT improves on serialised models of dialogue for predicting turn ends and the more difficult task of predicting when a turn will start.

1 Introduction

Turn-taking—deciding who speaks at what point during an interaction—is a crucial component of successful spoken communication between humans. However, as the example in Figure 1 depicts, it is an intricate task. The organisation between conversational partners has fascinated psycholinguists for decades, particularly how interlocutors achieve such short transitions between turns: gaps between turns typically range from -100 to 500ms (Levinson and Torreira, 2015) (the negative end of the range indicating an overlap between turns). To explain the speed at which turn-taking occurs, Sacks et al. (1974) pioneered the *predictive* model of turn-taking, theorising that people engage in some form of "projection" to determine an appropriate point to begin their own turn while their partner is still speaking. Under this model, speakers construct their speech such that potential turn transition points are foreshadowed to their listener. This raises the question – what features of speech do listeners rely on to predict potential turn ends?

In spoken conversation, turn-end cues stem from both the lexical content and its prosodic realisation. (Ford and Thompson, 1996; Bögels and Torreira, 2015; De Ruiter et al., 2006; Ward, 2019). However, their relative contributions are unknown. While models that leverage lexical and prosodic cues in isolation and combination can learn to predict some turn-taking behaviour, the simplicity of linguistic representations used in such models may obscure the true contributions of lexical content to turn-end prediction (Ward et al., 2018; Roddy et al., 2018). More recently, Ekstedt and Skantze (2020) proposed *TurnGPT*, a model for turn-end prediction that harnesses the power of pre-trained GPT-2 (Radford et al., 2019). TurnGPT achieves a high accuracy in predicting turn-endings, demonstrating the value of lexical information for this task.

TurnGPT has, however, been predominately trained and designed for *written* conversation. Like the GPT-2 model it is based on, the model is limited to a single stream of input. Although written dialogue can be neatly collapsed into a single stream of interleaved turns, compressing spoken conversation in this way disregards much of the nuance of realistic turn-taking behaviour. Whilst TurnGPT has been applied to spoken dialogue, it does so by serialising overlapping utterances into a single stream, sometimes requiring significant reordering or removal of lexical content; Figures 2a and 2b give examples of the TurnGPT formatting.

This paper seeks to better understand the contribution of lexical content to turn-taking in spoken communication by more accurately modelling its temporal dynamics. For this, we propose modelling transcripts as a dual-stream system that resembles their original production much more closely. We present *PairwiseTurnGPT*, a novel architecture capable of modelling these synchronous input streams. Doing so provides insights into how

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

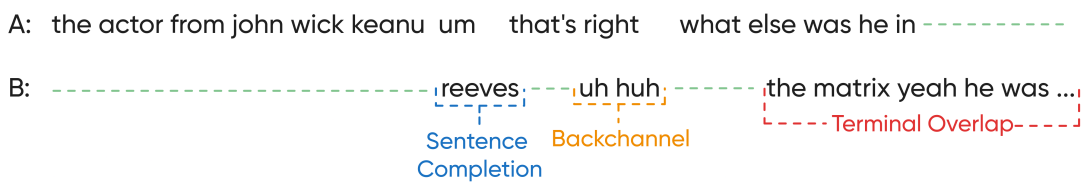


Figure 1: A (synthetic) example of overlaps in dialogue.

well lexical content can predict different types of turn-taking behaviour, including how speakers both end and begin their turns. By comparing training data configurations containing varying degrees of turn-taking complexity (i.e., partial, complete, and backchannel overlaps), we analyse their relative contributions to turn-taking prediction.

2 Background

2.1 Theories of turn-taking

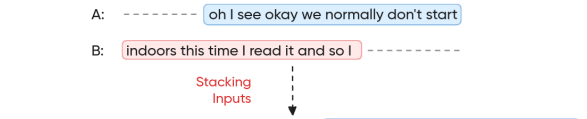
Human turn-taking behaviour has generally been characterised by two processes within the literature: the *reactionary* and the *predictive* approach. The former assumes that participants understand end-of-turn signals and react to them accordingly while the predictive approach posits that listeners predict the end of the turn in advance to time their response. The reactionary approach was pioneered by (Duncan, 1972, 1973, 1974; Duncan and Fiske, 2015) who argued for a precise set of context-free turn-yielding 'signals' which include both vocal and gestural signals (Yngve, 1970).

Others have argued against the general model of a reactionary approach because turn-transitions occur too quickly and turn-yielding signals occur too late within a speaker's utterance for the listener to simply react to an end-of-turn signal (Levinson and Torreira, 2015; Riest et al., 2015). Under the predictive account of turn-taking, the speed of turn transitions is possible because speakers predict appropriate points at which to start their turn (Sacks et al., 1974). This model views turns as combinations of *Turn Construction Units (TCUs)*. TCUs are separated by *Transition Relevance Places (TRPs)* that mark where a turn-transition (turn-shift) can but does not have to occur.

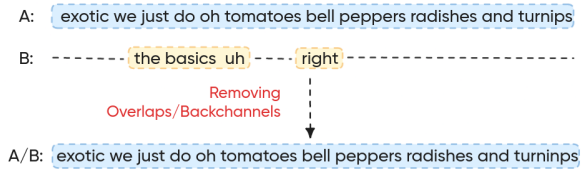
2.2 Behavioural evidence

Early research into turn-yielding signals identified prosodic, syntactic and gestural features that coincide with turn-completions (Duncan, 1972), however defining their contributions has proved complicated. For example, gestural features (Duncan, 1972) and gaze (Kendon, 1967) have shown to be useful cues for turn-taking, but they are action-dependent and more context-sensitive than other features (Clayman, 2012). Numerous works have demonstrated the importance of lexical information for this task. De Ruiter et al. (2006) found that end-of-turn prediction was unaffected by the removal of intonational contours but it was affected by the removal of lexicosyntactic information. Similarly, Magyari and De Ruiter (2012) found that when participants predicted the remaining part of a sentence, this prediction was more accurate if their end-of-turn prediction was also accurate. This suggests that listeners use predicted utterances to determine turn-completion. Pickering and Garrod (2013) also found that listeners imitate the speaker to determine their intention, which they use alongside the speaker's speaking rate to correctly time their own prepared utterance. Findings on importance of lexicosyntactic information align well with the predictive account of Sacks et al. (1974).

Although Ford and Thompson (1996) show that most TRPs occur at syntactic completion points, they theorised that multiple factors are used to determine the completion of TCUs. This theory was tested by Bögels and Torreira (2015) who also sought to refute the claim that intonation had no



(a) Overlapping portions of turns between speaker A and B are ignored by stacking turns in the combined stream. <eot> tokens mark the end of a speaker turn.



(b) Backchannels and completely overlapped turns from speaker B are removed in the combined stream.

Figure 2: The difficulties of serialising spoken dialogue transcripts into a single combined stream.

effect on turn-taking prediction by De Ruiter et al. (2006). This was done by performing the same experiment but with instances of questions with equal syntactic completion points but different turn-shift locations They found that in cases of syntactical ambiguity, lexicosyntactic information is not sufficient for turn-end projection and as such they claim intonation plays a role in disambiguation.

2.3 Computational models for End-of-Turn Detection and Prediction

Models trained to predict turn-taking behaviour are another method for investigating the relative contributions of lexical and acoustic cues. Skantze (2017) show that training with POS tags improves on a purely prosodic model, which supports the importance of syntactic completeness as a cue for turn-end prediction. However, Ward et al. (2018) outperforms Skantze (2017) using only prosodic features. Maier et al. (2017) and Roddy et al. (2018) both employed LSTM RNN models to investigate prosodic and linguistic features in conjunction; Roddy et al. (2018) found that acoustic features are more beneficial and Maier et al. (2017) found that linguistic features performed worse than in their baseline condition. However, linguistic features used in these studies have been simplistic and are unlikely to capture *pragmatic* completeness, a feature deemed crucial by Ford and Thompson (1996).

Ekstedt and Skantze (2020) proposed TurnGPT to harness the strong language modelling of GPT-2. TurnGPT finetunes GPT-2 with a modified objective for dialogue by adding speakers tokens and turn-shift tokens in the model input. The use of a pre-trained language model allows for greater pragmatic and semantic feature representation: TurnGPT is shown to rely not only on syntax but also on the overall pragmatic context of an utterance for turn-end prediction. Jiang et al. (2023) extended the model to condition its predictions on a generated response and found further improvements in end-of-turn prediction performance.

However, these models are not designed specifically for spoken dialogue with much more complex turn-taking behaviour than written dialogue. Transcripts of spontaneous spoken conversations only make up 4% of the training set for TurnGPT. More importantly, the dialogue transcripts are collapsed into a single stream of input for GPT2. To do this, dialogue transcripts are serialised based on turn units; turn units which are fully overlapped or are classified as backchannels, are removed from the transcription, whilst consecutive turn units from the same speaker are concatenated to form each speaker's full turns. The process is depicted in Figure 2a. As well as removing important information about a conversation, the process might also be viewed as fundamentally altering the task of turn-taking prediction in spoken dialogue.

Recently, TurnGPT representations have been used by Wang et al. (2024) in conjunction with an acoustic model to predict backchannel events in spontaneous dialogue transcripts. Like TurnGPT, this model is trained using data serialised into sequential speaker turns; however, backchannels are reintroduced using word-level time stamps. Whilst backchannels are known to be strongly linked to their prosodic realisation (Gravano and Hirschberg, 2011), Wang et al. (2024) found good performance for their prediction using only a language model.

The studies described above provide evidence that lexical information contributes to turn-end prediction, but do not provide a complete picture of its contribution in spoken conversation. We apply powerful modern language models to more realistic representations of turn-taking.

2.4 Characterising overlaps

Overlaps are a frequent and important component of spoken dialogue. Overlaps can occur when speakers mispredict the end of a TCU; however, they can also serve interactional purposes that are lost when serialising spoken dialogue transcripts into a single stream.

Overlaps can be categorised as competitive or cooperative, where speakers are either vying for the floor or aiding one another in the construction of a turn Schegloff (2000). As depicted in the example in Figure 1, cooperative overlap can consist of: *terminal overlap*, where the listener predicts the end of a turn and begins speaking prior to the other speaker finishing their turn; *turn completion* where the listener helps the speaker complete their turn but doesn't intend to take the floor; and *backchannels* such as "uh huh" and "hmm". These typically occur where the speaker requests affirmation from the listener and have their own set of cues, as defined by Clark (1996).

3 PairwiseTurnGPT: A New Approach

Our proposed approach – which we call PairwiseTurnGPT – models each speaker in a conversation

as an independent stream of tokens. We pair tokens across the two streams based on word timing information, enabling effective modelling of the complex interactions between speakers. This avoids the deficiencies inherent in the standard serialised approach (Ekstedt and Skantze, 2020), where turns are interleaved in a manner that erases turn-taking phenomena potentially conveying important information. Though Wang et al. (2024) only incorporate a limited aspect of temporal dynamics, their results demonstrate the value of such information. By aligning streams at the word level, we encode this structure much more explicitly.

3.1 Model Architecture

PairwiseTurnGPT is composed of a GPT-2 stream for each speaker in the dialogue. A diagram is included in Figure 4). Similarly to the spoken dialogue model proposed by Nguyen et al. (2023), GPT-2 weights are shared between the streams. Through a multi-head cross-attention layer in each transformer block, predictions in each stream are conditioned on the conversational history of both speakers. The training objective is the sum of the cross-entropy loss for each speaker streams[1].

To incorporate spoken turn-taking phenomena in PairwiseTurnGPT, we use word-level timings to align the speaker streams. GPT-2's BPE tokenisation functions at the sub-word level, therefore we obtain token timings by uniformly splitting word timings across tokens (Figure 5 depicts an example). Tokens are then aligned in a pairwise manner. For tokens with no significant overlap (defined as an overlapping duration no greater than 50% of the shorter-duration word), an empty `<emp>` token is used to make up the token pair. An example of this alignment is shown in Figure 3a.

3.1.1 Turn-Level Annotation

Pairwise alignment enables our models to learn taking behaviours that involve fine-grained overlap between conversational partners. From the aligned data we identify categories of such turn ending strategies to better understand which behaviours are captured by PairwiseTurnGPT. **Backchannels** involve one speaker interjecting a short utterance such as "hmm", "uh huh" or "yeah" to provide feedback to the speaker (Ward, 2004). We follow (Ekstedt and Skantze, 2020) and define these based on their lexical content[2] and a pause of at least $1s$ between surrounding turns from the same speaker. **Complete Overlap** occurs where one speaker begins and ends their turn before the other speaker finishes theirs, as depicted in Figure 2b. **Yield Turn-Shift** are when one speaker begins their turn before the other speaker finishes theirs (i.e. a partial overlap). Yield turns are those that contain an overlap of $> 0.1s$, or where the other speaker produces an overlap within 2s of the turn ending. **Normal Turn-Shift** turns involve one speaker finishing their turn and the other speaker beginning theirs after a pause. The difference between normal and yield turns is shown in Figure 3b. Appendix C shows how turn types are distributed in Switchboard.

The full alignment process is demonstrated in Figure 3a which includes turn annotation: the determination of the type for each utterance; turn alignment: ensuring each token is aligned appropriately and turn token addition: adding in the end-of-turn token corresponding to the determined turn type.

4 Experimental Setup

4.1 Model

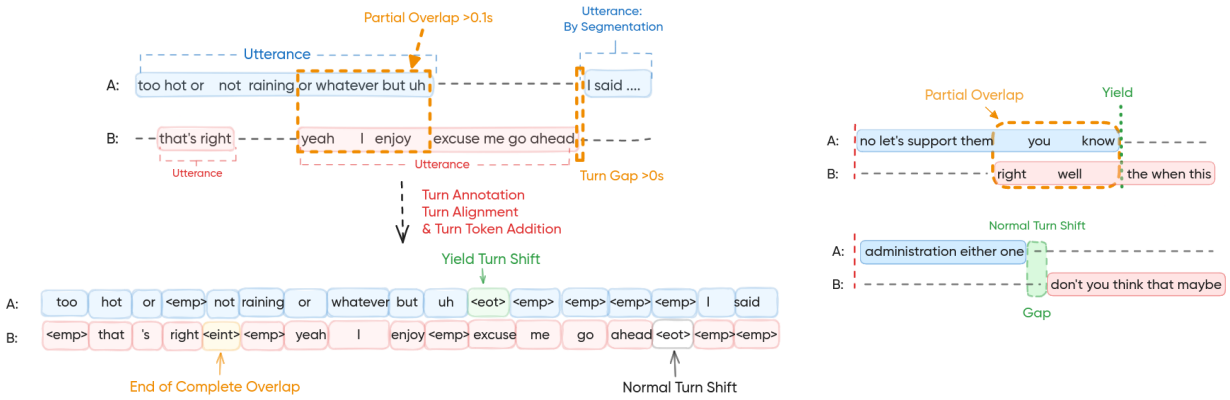
To allow comparison with TurnGPT, we initialise our model with *GPT2-base* which consists of 124M parameters, 12 layers, 12 heads and 768 hidden units. The pre-trained weights were obtained from the OpenAI GPT2 model from the Hugging Face Transformers library (Wolf et al., 2020). The cross-attention weights are initialised using the default Hugging Face method by sampling a unit normal distribution. We fine-tune using the AdamW optimizer with a learning rate of $6.25e^{-5}$ and a weight decay of 0.01. All models are trained for 5 epochs or until the validation loss does not decrease for two consecutive epochs with batches of size 4.

4.2 Data

We train and evaluate PairwiseTurnGPT on the Telephone Speech Switchboard Corpus which consists of 2430 conversations between 542 participants (Godfrey et al., 1992; Deshmukh et al., 1998). Although the dataset is not large, it includes manual transcriptions and manually corrected word timings, making it an ideal base for our study. We remove all annotations of non-verbal vocalisations from the transcripts while partial words, mispronunciations and coinages are replaced with the full

[1]We release our code at https://github.com/Sean-Leishman/PairwiseTurnGPT. This includes code for model training and data preprocessing.

[2]We use the list of candidate backchannel responses defined in Ekstedt and Skantze (2020)



(a) Pairwise data preprocessing: tokens are aligned based on token timing information. B's first utterance is labelled as a complete overlap; the end of A's first utterance is labelled as a "yield" turn shift on account of the partial overlap with B's second utterance, whilst B's second utterance is labelled as a normal turn shift.

(b) Yield & normal turn-shift; labels are based on the amount of overlap with the other speaker's turn.

Figure 3: Data Labelling & Preprocessing

intended word. Switchboard doesn't have a standard evaluation set for this task, so we randomly split the corpus into train, validation and test sets ($[90/5/5]$ proportions, respectively).

4.3 Training data configurations

We train with pairwise data configurations that include varying degrees of turn-taking behaviour to understand their respective contributions to turn-end predictions.

- **Single stream**: As a baseline, we examine the performance of modelling isolated conversation streams. For this, we remove cross-attention and concatenate turns in each stream so no alignment between speakers takes place.
- **Serialised**: We simulate TurnGPT dialogue representations by aligning streams to turns rather than tokens, thus removing all overlap. Content tokens in one stream are always aligned with `<emp>` tokens in the other.
- **Aligned**: Partial overlaps, where a speaker interrupts prior to a turn ending, are included but not complete overlaps or backchannels.
- **Aligned + Overlaps**: Both partial and complete overlaps are included.
- **Aligned + Backchannels**: Partial overlaps and backchannels are included.
- **Aligned + Backchannels & Overlaps**: The fully aligned condition contains all turn-taking phenomena.

4.4 Evaluation Metrics

The end-of-turn prediction task involves mapping next-token prediction to a binary turn end prediction. We discretise the probability of end-of-turn tokens into a binary label using a threshold tuned on the validation set. Turn prediction is evaluated using Balanced Accuracy (bAcc), the mean of the true positive and true negative rates for turn end prediction; it is robust to the unbalanced nature of turn-end tokens and allows more direct comparison to the TurnGPT results. We also evaluate language modelling through token-level perplexity (PPL).

5 Results

5.1 Validating the pairwise architecture

We begin by establishing that the PairwiseTurnGPT architecture performs at a similar level to the original TurnGPT model. We also investigate the effect of the second speaker stream using different training configurations, where no cross-attention layer is introduced. We consider conditions where each stream consists of turns concatenated together (i.e. the single stream condition) to simulate no knowledge of the other speaker; and where each stream includes empty tokens (i.e. serialised without cross-attention), which simulates the temporal aspect of the other speaker but without any lexical content.

Each PairwiseTurnGPT configuration is evaluated using the serialised data configuration in Table 1. TurnGPT and serialised PairwiseTurnGPT achieve nearly identical turn-end prediction performance; however, PairwiseTurnGPT has higher PPL, indicating a weaker internal representation

Model	bAcc ↑	PPL ↓
TurnGPT	0.828	29.3
PairwiseTurnGPT		
Single Stream	0.805	39.3
Serialised w/o CA	0.825	32.9
Serialised	0.828	31.3

Table 1: End-of-turn prediction accuracy and perplexity scores for serialised data across models. PairwiseTurnGPT contains <emp> tokens while TurnGPT does not, so <emp> tokens are not evaluated.

of language. This may be a result of the more demanding training procedure.

We find that a reasonable turn-end accuracy of 0.805 can be achieved using the single-stream configuration. Without knowledge of the other speaker, this model is reliant on the syntactic completeness of the speaker's current utterance and a partial history of the conversation. Removing cross-attention (CA) from the serialised TurnGPT setup demonstrates how model performance is influenced by the other speaker's lexical content. We find that much of the serialised model performance can be achieved without cross-attention (i.e., only using information about when the other speaker is active).

5.2 Training data configurations

We evaluate the effect of training using our data configurations that incrementally approach the original spoken realisation. Rather than evaluating over a serialised configuration, as in Table 1; we evaluate each model on the fully-aligned configuration of the test data; results are shown in Table 2.

Single stream model The turn-end accuracy scores confirm that a speaker's turn ending is at least somewhat predictable from their own conversational history, which contains information regarding the syntactic and pragmatic completeness of the current utterance. As expected, all models trained to condition their predictions on both speaker streams improve over the single stream set up for all turn end types; even a model trained on the serialised data configuration can achieve a 14-point increase in accuracy.

Aligned vs. serialised Table 2 shows that the lossy encoding of the serialised configuration ignores much of the complex turn-taking behaviour in spoken dialogue: training on the aligned configuration produces better overall accuracy than the serialised configuration. The overall improvement comes primarily from the model's ability to predict yielded turn ends; we find a slight decrease in the ability to predict standard turn endings. The partial overlaps in the aligned configurations are a common feature of spoken dialogue; yielded turns constitute 18.11% of turn endings in the dataset (see Table 4). While their inclusion may complicate the prediction of normal turn endings, they are extremely valuable for capturing the true variety of turn ends in spoken conversation. However, in regard to normal turn-ends, performance varies between the serialised and aligned configurations. This suggests that for simpler turn-ends the model could benefit from developing a simpler understanding of turn-taking, without considering additional phenomena.

Effect of backchannels & overlaps Table 2 shows that models trained on aligned configurations, which include backchannels, are the most accurate overall. Only these configurations outperform the serialised data configuration for predicting normal end-of-turns. Their influence may reflect their communicative functions in spoken dialogue: for example, listeners can employ them to inform the speaker of their intention to continue listening (Yngve, 1970). Even from their lexical content alone, our results demonstrate that backchannels are useful cues for turn-taking.

The inclusion of overlaps also improves turn-end prediction over the aligned configuration, however, to a lesser extent. Overlaps may be more difficult to leverage as they make up a smaller proportion of the turn-end tokens in our corpus and are far less constrained than backchannel responses. For example, overlaps can be cooperative or competing acts in dialogue (Schegloff, 2000). If an overlap is competing, it may be less likely that its resolution can be derived from its lexical content alone. However, the inclusion of overlaps in the aligned data achieves the highest yield turn-ending prediction accuracy as the model can better differentiate between complete and partial overlaps.

Although the combination of both features still provides clear improvement over the serialised condition, including all behaviours doesn't necessarily provide an additive benefit for predicting all turn ends. Notably, the accuracy for yield turns in the fully aligned configuration deteriorates compared to that aligned with only overlaps, suggesting

Configuration	bAcc ↑			PPL ↓		
	Normal	Yield	All	Overlap	Non-Overlap	All
Single stream	0.728	0.640	0.710	–	–	–
Serialised	0.868	0.807	0.852	206	**5.63**	7.67
Aligned	0.863	0.927	0.881	48.1	5.77	6.95
+ Backchannel	**0.872**	0.930	0.914	41.3	5.68	6.55
+ Overlap	0.866	**0.936**	0.890	40.2	5.64	6.69
+ Backchannel & Overlap	0.869	0.934	**0.915**	**36.9**	5.67	**6.49**

Table 2: End-of-turn prediction balanced accuracy over turn types. "All" consists of normal, yield, backchannel and complete overlap turn endings. Perplexity is computed on overlapping, non-overlapping, and all tokens of the fully-aligned test set.

that backchannels may blur the distinction between yields and normal turn endings.

Rule-based comparison To probe the necessity of language modelling for this task, we designed a rule-based classifier to predict an end-of-turn whenever two speakers speak simultaneously. This classifier predicts a turn-shift whenever the listener interrupts the current speaker's utterance. Using the setup in Table 2, the classifier achieves a bAcc of 0.890 over yielded turn ends. This is a strong improvement over the serialised model accuracy of 0.807, indicating that an interruption is a significant signal. However, the gap between the fully aligned model with 0.934 accuracy shows that lexical content provides additional predictive power.

Perplexity As expected, the lowest PPL is found for the model trained on the fully aligned data. Although the serialised configuration does not handle overlapping tokens well, it produces the lowest "non-overlap" PPL. This model may be better able to model lexical content as it does not need to learn temporal aspects of overlapping tokens. The introduction of partial overlaps in the aligned configuration allows the model to better represent overlaps; each subsequent data configuration yields further improvement. By representing overlapping portions of the dataset more effectively, the model may learn patterns regarding how overlaps are resolved and lexical features that prompt a listener to produce an overlap.

Similar to the trends in end-of-turn prediction, the inclusion of backchannels in the aligned training data produces a larger overall reduction in PPL than overlaps. However, these configurations produce similar PPL scores across all token sets. The "overlap" subset does not contain backchannels and so while we may expect the backchannel configuration to perform similarly to the aligned configuration, it achieves a PPL that is closer to the aligned with overlaps configuration. This suggests that by learning to represent backchannel turns, the model can extrapolate to overlaps relatively well.

As expected, the fully aligned configuration trained with all types of overlap performs best overall and in the overlap subset of the fully aligned test set. However, this is not the case for non-overlapping tokens where the result is essentially a weighted sum of the PPL resulting from the aligned with backchannels and aligned with overlaps configurations. This is reflected in end-of-turn prediction and suggests a degree of uncertainty when combining two types of features.

5.3 Prediction of turn starts

Thus far, we have examined turn-taking through the lens of turn ends. However, pairwise alignment also allows us to analyse the different strategies people use to initiate a turn. Here, we evaluate how useful lexical content is for determining interjection points for different types of turns.

We predict the beginning of a turn by summing the probability of all non-`<emp>` tokens and producing a binary prediction as was done for the end-of-turn task. Using the fully time aligned test configuration, we evaluate predictions at points where the current token is the `<emp>` token. We consider several turn start strategies: "Normal" is the start of a non-overlapping turn; "Interruption" is the start of an overlap turn in which the interrupted speaker yields the floor; "Overlap" is the start of a completely overlapped utterance; "BC" is the start of a backchannel.

Configuration	bAcc ↑				
	Normal	Interruption	Overlap	BC	All
Serialised	0.702	0.640	0.581	0.592	0.640
Aligned	0.746	0.669	0.592	0.604	0.669
+ Backchannel	0.809	0.763	0.684	0.753	0.763
+ Overlaps	0.774	0.700	0.614	0.647	0.700
+ BC & Overlaps	**0.819**	**0.774**	**0.689**	**0.765**	**0.774**

Table 3: Predicting start points for different turn types.

As expected, Table 3 shows that overlap turn starts are the most difficult type of turn start to predict. The addition of turn-taking phenomena improves prediction across all turn types. Interestingly, the addition of backchannels is far more useful than the addition of overlaps, mirroring our findings from turn-end prediction performance.

Predicting overlap turn starts is worse across all configurations compared to normal turn starts, likely because overlapping turns do not align with a turn end. However, the fact that yielded turn endings can be accurately predicted suggests that lexical content provides an indication of suitable interjection points in conversation. Results are similar for backchannel turn predictions.

Interestingly, the fully time aligned configuration performs best across all turn start types. However, this was not the case for predicting the end of a turn where this configuration was not the most accurate for normal or yield turn ends. We posited that the result over end-of-turns is due to the overall complexity of the training data. However, it seems that by framing the task differently, the model is able to leverage this information.

6 Discussion & Conclusions

By modelling spoken dialogue transcripts as two separate streams of lexical content, our proposed PairwiseTurnGPT provides a much more nuanced understanding of how lexical content contributes to the predictability of turn-taking behaviour than was previously possible. It also improves the accuracy of predicting turn ends over models of dialogue serialised at the level of turns.

We find that both the timing and content of overlaps contribute jointly to increased predictive power. By comparing training data configurations containing different turning-taking behaviours, we demonstrate the relative contributions of partial, complete, and backchannel overlaps for accurately predicting the variety of turn ends and starts that occur in spoken dialogue. Each training data augmentation improves overall turn-end prediction but through different means. Though the underlying intent of backchannel responses is known to be mediated by their prosodic realisation, our results show that the lexical content alone is already a valuable cue for predicting turn ends (Lai, 2009). Overlaps are also useful but to a lesser extent. Though they complicate non-overlapped turn-end predictions, they are crucial for accurately modelling yielded turns. We find that the alignment configuration containing all forms of overlap can muddy the distinction between yields and normal turn endings. Interestingly, training with this data configuration consistently improves predictive performance across turn start types, suggesting that predicting turn starts and ends may benefit from different information. For example, Jiang et al. (2023) has shown that turn starting points are better predicted when conditioned on the content of the upcoming response. Though related, our results highlight the importance of investigating turn-ends and turn-starts as separate prediction tasks.

By allowing for synchronous streams of lexical content, PairwiseTurnGPT provides nuanced insight into how much lexical context contributes to the prediction of turn-taking behaviours in spoken dialogue. This model has the potential to be used in dialogue systems or for gaining deeper insights into human turn-taking behaviour.

Limitations We selected the Switchboard dataset as a representation of extremely natural spoken dialogue and for its manually annotated transcripts and timestamps. However, it is limited in size. Some of our results suggest that the training set may not be sufficiently large to capture the complexities of these interactions fully. In particular, the cross-attention layer that encodes interactions between the two speaker streams is trained from scratch. For example, the difference in accuracy between the serialised conditions with and without cross-attention in Table 1 is less significant than we might expect. Investigating larger training corpora may allow the model to better capture the interaction between both streams. Exploring the predictability of turn-taking in other types of spoken conversation, such as interviews or conversations between friends, could also further illuminate the role of lexical information in turn-taking prediction. We expect the inclusion of prosodic information to improve turn-taking behaviour prediction further and leave this for our future work.

References

Sara Bögels and Francisco Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Steven E Clayman. 2012. Turn-constructional units and the transition-relevance place. *The handbook of conversation analysis*, pages 151–166.

Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Johnathon Hamaker, and Joseph Picone. 1998. Resegmentation of switchboard. In *Internation Conference on Spoken Langauge Processing*.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

Starkey Duncan. 1973. Toward a grammar for dyadic conversation. *Semiotica*, 9(1).

Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns1. *Language in society*, 3(2):161–180.

Starkey Duncan and Donald W Fiske. 2015. *Face-to-face interaction: Research, methods, and theory*. Routledge.

Erik Ekstedt 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*, pages 2981–2990. Association for Computational Linguistics.

Cecilia E Ford and Sandra A Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184.

John J. Godfrey, Edward Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520 vol.1.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023. Response-conditioned turn-taking prediction. In *ACL Findings of the Association for Computational Linguistics*, pages 12241–12248. Association for Computational Linguistics.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Catherine Lai. 2009. Perceiving surprise on cue words: Prosody and semantics interact on right and really. In *Proceedings of Interspeech*, pages 1963–1966.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

Lilla Magyari and Jan P De Ruiter. 2012. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology*, 3:376.

Angelika Maier, Julian Hough, and David Schlangen. 2017. Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems. In *Proceedings of Interspeech*, pages 1676–1680.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative Spoken Dialogue Language Modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Carina Riest, Annett B Jorschick, and Jan P de Ruiter. 2015. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:89.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. In *Proceedings of Interspeech*, pages 586–590.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. volume 50, page 696 – 735.

Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63.

Gabriel Skantze. 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*, pages 220–230.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12121–12125.

Nigel Ward. 2004. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody International Conference*.

Nigel G Ward. 2019. *Prosodic patterns in English conversation*. Cambridge University Press.

Nigel G Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. 2018. Turn-taking predictions across languages and genres using an lstm recurrent neural network. In *SLT IEEE Spoken Language Technology Workshop*, pages 831–837. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistic Society*, pages 567–578.

A Model architecture

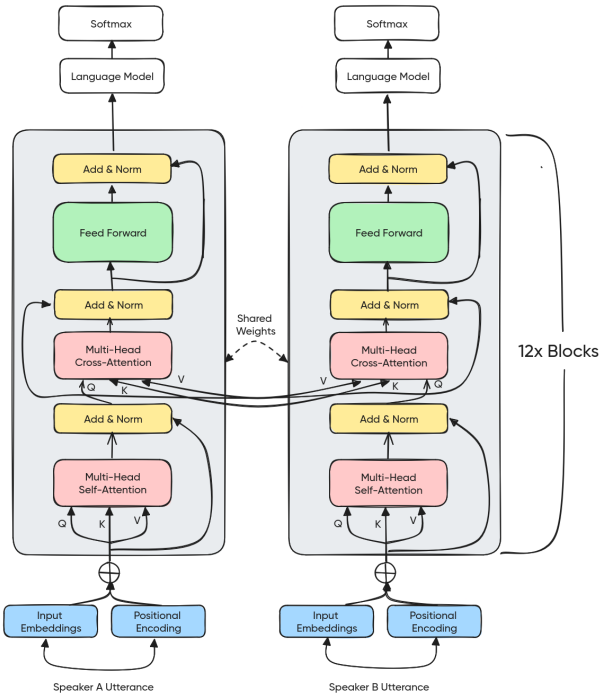


Figure 4: PairwiseTurnGPT Architecture

B Sub-word token alignment

Figure 5 shows how the word ‘uhhuh‘ is decomposed into three sub-word tokens ([“uh”,“h”,“uh”]) under the GPT-2 byte pair encoding tokenizer. The timing of each sub-word token is approximated from the original word-level timestamps by splitting the word duration uniformly across the constituent tokens.

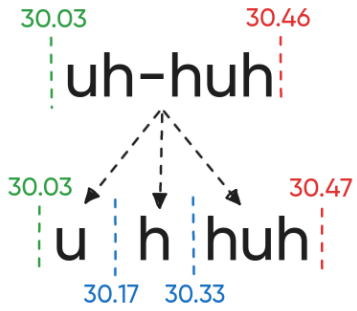


Figure 5: Deriving subtoken alignment.

C Turn type frequency

Turn Type	Token	Count	%
Normal	<eot>	**74522**	**45.19**
Partial Overlap	<yield>	29861	18.11
Overlap	<eint>	16250	9.85
Backchannel	<ebc>	44281	26.85
All	164914	-	

Table 4: The frequency of each turn type in Switchboard using our turn annotation procedure.