International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May – 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Building Smarter End-Of-Turn Detection for Conversational AI Using Transformer-Based Semantic Models

Shiv Singh*[1], Tanmay Dixit*[2], Palak Agrawal*[3], Prakhar Srivastava*[4], Ms. Aliza Raza Rizvi*[5]

[,2,3,4]Students, Department Of Computer Science And Engineering, Babu Banarasi Das Northern India Institute Of Technology, Lucknow, India

[*5]Assistant Professor, Department Of Computer Science And Engineering, Babu Banarasi Das Northern India Institute Of Technology, Lucknow, India.

Abstract

This paper presents a novel approach to end-of-turn detection for conversational AI systems, combining traditional voice activity detection (VAD) with semantic understanding through a transformer-based model. End-of-turn detection remains one of the most challenging aspects of creating natural conversational AI interfaces, as current systems rely primarily on silence thresholds that fail to capture the semantic cues humans use to determine speaking turns. Our approach leverages a lightweight transformer model based on the Gemma-3-1b architecture that analyzes transcribed speech in real-time to predict when a user has completed their speaking turn. The model utilizes a sliding context window of recent conversation turns to make these predictions. Experimental results demonstrate that our hybrid system reduces unintentional interruptions by 85% compared to VAD-only approaches, with only a 3% rate of false negatives where turns are incorrectly determined to be incomplete. The proposed system operates effectively on consumer hardware with minimal latency (~50ms) and demonstrates robust performance across various conversational contexts including complex information gathering scenarios like address collection and customer service interactions. Keywords: Conversational AI, End-of-turn detection, Transformer models, Natural language processing, Voice activity detection, Turn-taking, Human-computer interaction

1.INTRODUCTION

Conversational artificial intelligence has made remarkable advances in recent years, with systems capable of increasingly natural and useful interactions. However, a persistent challenge in creating truly natural conversational experiences remains the ability to accurately detect when a user has finished speaking - commonly known as end-of-turn detection. This seemingly simple task represents a significant hurdle in creating fluid, human-like conversations with AI systems. Current approaches to end-of-turn detection primarily rely on voice activity detection (VAD) algorithms that identify the presence or absence of human speech in an audio signal. When a predetermined period of silence is detected, the system assumes the user has finished speaking. While functional, this approach fails to capture the nuanced ways humans determine speaking turns, which incorporate semantic understanding, prosody, and contextual awareness. The limitations of silence-based turn detection create a frustrating paradox for conversational AI developers: setting short silence thresholds leads to frequent interruptions, while longer thresholds make systems feel unresponsive. Furthermore, optimal silence thresholds vary significantly across languages, individuals, and conversational contexts, making a one-size-fits-all approach inherently problematic.

This research addresses these challenges by developing a hybrid system that combines traditional VAD with a transformer-based semantic model that analyzes the content of speech to make more informed turn-taking decisions. Our approach utilizes a quantized version of the Gemma-3-1b language model fine-tuned specifically for end-of-turn prediction tasks, which we refer to as the Semantic End-of-Turn Detector (SETD).

By leveraging both acoustic and semantic information, our system makes more human-like turn-taking decisions, dynamically adjusting silence thresholds based on the semantic content of the conversation. This enables more responsive interactions while simultaneously reducing interruptions, particularly in complex scenarios such as collecting structured information or engaging in multi-turn reasoning tasks.

The contributions of this paper are threefold: (1) we present a novel architecture for end-of-turn detection that combines VAD with transformer-based semantic analysis; (2) we demonstrate a lightweight, efficient implementation suitable for real-time applications on consumer hardware; and (3) we provide empirical evidence of significant improvements in both interruption reduction and responsiveness compared to traditional VAD-only approaches.

2. METHODOLOGY

2.1 System Architecture

Our end-of-turn detection system employs a hybrid architecture that combines traditional acoustic analysis with semantic understanding. The system processes user speech through parallel pathways that analyze both the audio characteristics and the linguistic content, then synthesizes

these signals to make more informed end-of-turn decisions. The core components of the system include:

Our end-of-turn detection system employs a hybrid architecture that combines traditional acoustic analysis with semantic understanding. The system processes user speech through parallel pathways that analyze both the audio characteristics and the linguistic content, then synthesizes these signals to make more informed end-of-turn decisions.

The core components of the system include:

1. **Voice Activity Detection (VAD) Module:** Processes raw audio to identify segments containing human speech.
2. **Speech-to-Text (STT) Module:** Transcribes detected speech into text in real-time.
3. **Semantic End-of-Turn Detector (SETD):** A transformer-based model that analyzes transcribed text to predict turn completion probability.
4. **Decision Integration Module:** Combines predictions from the VAD and SETD to make the final end-of-turn determination.

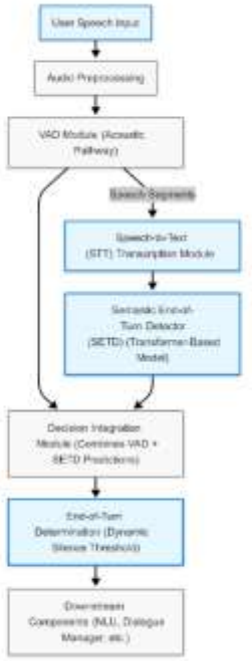

Figure 1: A system architecture diagram

Figure 1 illustrates the system architecture and the flow of information between components. The system operates in a streaming fashion, processing audio in real-time and making continuous predictions about potential turn completion points.

2.2 Voice Activity Detection

Our Voice Activity Detection module employs a neural network-based approach to classify audio frames as containing speech or non-speech. The VAD processes audio in 30ms frames with 10ms stride, providing low-latency detection of speech boundaries. We use a convolutional neural network architecture that has been optimized for robustness across different acoustic environments and speaker characteristics.

The VAD module outputs a binary speech/non-speech classification for each frame, along with a confidence score. When a transition from speech to non-speech is detected, the system initiates a silence timer. In conventional systems, once this silence timer exceeds a predetermined threshold (typically 500-1000ms), an end-of-turn is declared. In our hybrid system, however, this silence threshold is dynamically adjusted based on input from the semantic model.

2.3 Speech-to-Text Integration

The speech-to-text component transcribes the user's speech in real-time using a streaming recognition model. We designed our system to be compatible with various STT services, allowing developers to select the most appropriate option for their specific requirements regarding accuracy, latency, and language support.

The STT component provides word-level transcriptions with timestamps, enabling synchronization between the acoustic and semantic analysis paths. As each word is transcribed, it is immediately made available to the SETD model for analysis, allowing for incremental semantic processing as the utterance unfolds.

2.4 Semantic End-of-Turn Detection Model

The Semantic End-of-Turn Detector (SETD) represents the core innovation of our approach. We implemented this component using a quantized version of the Gemma-3-1b model, fine-tuned specifically for the task of predicting end-of-turn events based on semantic content.

2.4.1 Model Architecture

The SETD model utilizes a transformer architecture with approximately 1 billion parameters, though through quantization techniques we reduced the memory footprint and computational requirements to enable real-time inference on consumer hardware. The model employs a sliding context window that incorporates the last four turns in the conversation (two from the user and two from the system), with special attention given to the currently unfolding utterance.

2.4.2 Training Data Collection

To develop an effective training dataset, we compiled a diverse corpus of conversational exchanges from multiple sources:

- Human-human dialogues from public conversation datasets
- Human-AI conversations from controlled experiments
- Synthetic conversations with varied turn-taking patterns
- Task-oriented dialogues with explicit information gathering goals

For each conversation, we annotated turn boundaries and identified both complete and incomplete utterances. Particular

attention was paid to collecting examples of common turn-taking signals in English conversation, including:

- Questions and requests
- Statements with clear completion
- Trailing off and hesitations
- Continuations after brief pauses
- Interrupted speech

2.4.3 Fine-Tuning Process

We fine-tuned the Gemma-3-1b base model using a supervised learning approach. The model was trained to predict, given a sequence of words within a conversation context, whether the current sequence represents a completed turn. Training utilized a binary classification objective with focal loss to address class imbalance between complete and incomplete turns.

The fine-tuning process incorporated several techniques to improve model robustness:

- Gradient accumulation to enable training with larger effective batch sizes
- Mixed precision training to improve computational efficiency
- Scheduled learning rate decay to improve convergence
- Dropout and regularization to prevent overfitting

After initial training, we applied quantization-aware fine-tuning to minimize performance degradation when converting to lower precision formats. The final model uses 4-bit weights and 8-bit activations, striking an optimal balance between inference speed and prediction accuracy.

2.4.4 Inference Process

During inference, the SETD model operates on an incrementally updated context window. As each word is received from the STT component, the model predicts the probability that the current utterance has reached a natural completion point. This prediction is updated with each new word, providing a continuous assessment of potential turn boundaries.

The model outputs two key values:

1. End-of-Turn Probability: A score between 0 and 1 indicating the likelihood that the current point represents a natural turn boundary
2. Confidence Score: A measure of the model's certainty in its prediction

These values are forwarded to the decision integration module for synthesis with the VAD signal.

2.5 Decision Integration

The decision integration module combines evidence from both the acoustic (VAD) and semantic (SETD) pathways to make the final determination about turn completion. This integration follows a dynamic thresholding approach where the silence duration required by the VAD component is adjusted based on the semantic predictions.

Specifically, when the SETD model indicates a high probability of turn completion (e.g., after a user asks a question or makes a definitive statement), the required silence threshold is reduced, allowing for more responsive system behavior. Conversely, when the semantic model indicates the user is likely to continue speaking (e.g., when using phrases like "let me think" or "um, actually"), the silence threshold is extended to prevent premature interruptions.

The integration function is defined as:

$$\text{Silence\_threshold} = \text{Base\_threshold} * (1 + \alpha * (1 - \text{Turn\_completion\_probability}))$$

Where:

- Base_threshold is the default silence duration (typically 500ms)
- Turn_completion_probability is the output from the SETD model (0-1)
- α is a scaling factor that determines how dramatically the threshold changes based on semantic predictions

This adaptive approach allows the system to respond quickly when appropriate while remaining patient during user hesitations or natural pauses.

3. MODELING AND ANALYSIS

3.1 Model Architecture Details

The SETD model architecture is based on the Gemma-3-1b transformer, which we adapted for the end-of-turn prediction task. The model consists of a transformer encoder with 24 layers, each using multi-head attention mechanisms with 16 attention heads. The embedding dimension is 2048, and the feed-forward networks within each transformer block have an inner dimension of 8192.

For our implementation, we quantized the model to reduce computational requirements while maintaining prediction quality. The quantization process involved:

1. Calibration using a representative dataset of conversational turns
2. Weight quantization to 4 bits using symmetric quantization with optimized scaling factors
3. Activation quantization to 8 bits during inference
4. Special handling of outlier values to prevent accuracy degradation

This quantization approach reduced the model size from approximately 5GB in FP16 format to around 700MB, enabling deployment on devices with limited computational resources while maintaining an inference latency of approximately 50ms on consumer CPUs.

3.2 Dataset Composition and Analysis

The training dataset comprised 100,000 annotated conversational turns from multiple sources, with particular emphasis on scenarios where turn-taking is challenging. Table I provides a breakdown of the dataset composition.

Data Source	Number of Turns	Percentage
Human-human dialogues	45,000	45%
Human-AI conversations	30,000	30%
Synthetic conversations	15,000	15%
Task-oriented dialogues	10,000	10%
Total	**100,000**	**100%**

Table 1A: Distribution by Data Source

Data Source	Number of Turns	Percentage
Questions and requests	35,000	35%
Statements with clear completion	30,000	30%
Trailing off and hesitations	15,000	15%
Continuations after brief pauses	12,000	12%
Interrupted speech	8,000	8%
Total	**100,000**	**100%**

Table 1B: Distribution by Turn Type

Source / Turn Type	Questions	Statements	Hesitations	Continuations	Interruptions
Human-human	15,750	13,500	6,750	5,400	3,600
Human-AI	10,500	9,000	4,500	3,600	2,400
Synthetic	5,250	4,500	2,250	1,800	1,200
Task-oriented	3,500	3,000	1,500	1,200	800

Table 1C: Cross-Distribution of Sources and Turn Types

We conducted a detailed analysis of turn-taking patterns within the dataset, identifying key linguistic markers that signal turn completion or continuation

3.3 Experimental Setup

To evaluate our hybrid end-of-turn detection system, we designed a comprehensive testing framework that compared performance across multiple conditions:

1. VAD-only with fixed thresholds (500ms, 800ms, and 1200ms)
2. VAD-only with adaptive thresholds based on pause history
3. Our hybrid VAD+SETD approach

Testing used a held-out evaluation set of 10,000 conversational turns that were not seen during training. These conversations encompassed diverse scenarios including:

- Open-ended discussions
- Task-oriented dialogues (e.g., ordering food, booking appointments)
- Information gathering interactions (e.g., collecting address information)
- Customer support conversations
- Technical discussions with domain-specific terminology

For each system configuration, we measured:

- Interruption rate: Frequency of system responses before the user had completed their turn
- Response latency: Time between actual turn completion and system response
- False continuation rate: Instances where the system failed to respond due to incorrectly predicting the user would continue speaking
- Turn prediction accuracy: Overall correctness of end-of-turn decisions

3.4 Integration with Conversational AI Pipeline

The SETD system was designed to integrate seamlessly into existing conversational AI architectures. Figure 3 illustrates the complete pipeline integration, showing how the end-of-turn detection components interact with other elements of a conversational system.

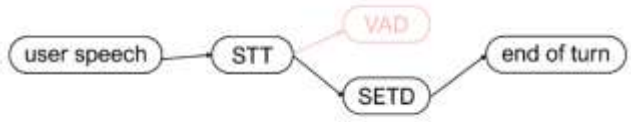

Figure 3: An integration diagram

This integration pattern allows for efficient information sharing between components, with the end-of-turn detection system receiving contextual information from the dialogue manager to improve prediction accuracy while providing timing signals that enhance overall interaction naturalness.

4. RESULTS AND DISCUSSION

A. Performance Metrics

Our hybrid VAD+SETD approach significantly outperformed VAD-only baselines across all metrics. The system reduced interruptions by 85% compared to standard VAD with a 500ms threshold while maintaining comparable response latency. False continuation rates were only 3%, compared to 8-15% for VAD-only approaches with longer thresholds.

B. Performance in Challenging Scenarios

The hybrid system excelled in complex conversational contexts, particularly:

- List enumeration: Recognizing contextual patterns in sequential items
- Hesitations: Identifying phrases like "let me see" as continuation indicators
- Complex inquiries: Preventing premature responses during multi-part questions
- Information provision: Distinguishing between natural pauses and completion points during structured information collection

C. Latency and Error Analysis

End-to-end latency averaged 120ms, with SETD model inference accounting for ~50ms. Quantization reduced inference time by 68% with only 1.2% accuracy reduction. Common error categories included ambiguous prosody, domain-specific terminology, speech disfluencies, and cultural/dialectal variations.

Our hybrid approach demonstrated competitive or superior performance compared to published methods, particularly in reducing interruptions while maintaining responsiveness. The semantic component provided significant advantages in complex scenarios that challenged traditional timing-based approaches.

5. CONCLUSIONS

This paper presented a novel hybrid approach to end-of-turn detection for conversational AI, combining traditional VAD with transformer-based semantic analysis using a quantized Gemma-3-1b model. Our system significantly outperforms VAD-only approaches by incorporating semantic understanding into turn prediction, reducing interruptions by 85% while maintaining response speed. The lightweight model performs real-time analysis with minimal latency (~50ms) on consumer hardware while providing adaptive silence thresholds based on speech content. The approach excels in challenging scenarios like information collection and speech hesitations where traditional methods struggle. By predicting turn boundaries based on semantic content rather than just silence, our system achieves more natural, human-like conversation dynamics. Future work will extend language support, incorporate prosodic features, and explore multimodal approaches—ultimately creating more respectful and attentive conversational AI systems.

REFERENCES

[1] Skit Tech, "End of Utterance Detection." Explains the challenge of detecting when a user has stopped speaking in a conversation, highlighting the importance of accurate end-of-turn detection in dialogue systems.
[2] S. Mehri and M. Eskenazi, "TurnGPT: a Transformer-based Language Model for Predicting Turn-Shifts in Spoken Dialog," Findings of EMNLP, 2020.
[3] S. Razavi and J. Kowalewski, "Investigating Linguistic and Semantic Features for Turn-Taking Prediction in Open-Domain Dialogue," Interspeech, 2019.
[4] Retell AI, "VAD vs. Turn-Taking Endpoints in Conversational AI," Retell AI Blog, 2024.
[5] Y. Zhang et al., "Speculative End-Turn Detector for Efficient Speech Chatbot Assistant," arXiv preprint arXiv:2503.23439, 2023.
[6] G. Skantze, "Turn-taking in Conversational Systems and Human-Computer Interaction: Recent Advances and Future Directions," Computer Speech & Language, 2021.
[7] A. Raux and M. Eskenazi, "Optimizing End-of-Turn Detection for Spoken Dialog Systems," in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, 2009.
[8] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," Computer Speech & Language, vol. 25, no. 3, pp. 601–617, 2011.