JAL-Turn: Joint Acoustic–Linguistic Modeling for Real-Time and Robust Turn-Taking Detection in Full-Duplex Spoken Dialogue Systems

Anonymous ACL submission

Abstract

Despite recent advances, efficient and robust turn-taking detection remains a significant challenge in industrial-grade Voice AI agent deployments. Many existing systems rely solely on acoustic or semantic cues, leading to suboptimal accuracy and stability, while recent attempts to endow large language models with full-duplex capabilities require costly full-duplex data and incur substantial training and deployment overheads, limiting real-time performance. In this paper, we propose JAL-Turn, a lightweight and efficient speech-only turn-taking framework that adopts a joint acoustic–linguistic modeling paradigm, in which a cross-attention module adaptively integrates pre-trained acoustic representations with linguistic features to support low-latency prediction of hold vs. shift states. By sharing a frozen ASR encoder, JAL-Turn enables turn-taking prediction to run fully in parallel with speech recognition, introducing no additional end-to-end latency or computational overhead. In addition, we introduce a scalable data construction pipeline that automatically derives reliable turn-taking labels from large-scale real-world dialogue corpora. Extensive experiments on public multilingual benchmarks and an in-house Japanese customer-service dataset show that JAL-Turn consistently outperforms strong state-of-the-art baselines in detection accuracy while maintaining superior real-time performance.

1 Introduction

In recent years, the rapid proliferation of voice AI agents in applications such as intelligent customer service, personal assistants, and human–AI collaboration has substantially raised the demand for natural and high-quality spoken interaction. Unlike traditional dialogue systems, modern voice AI agents are often designed for extremely low-latency responses and continuous listening, which makes them prone to intervening before users have fully completed their utterances. In natural conversation, speakers frequently exhibit thinking pauses, hesitations, and self-repairs, which do not necessarily indicate a turn completion. Overly aggressive system responses under such conditions can lead to frequent interruptions, disrupted conversational flow, and degraded interaction quality, ultimately undermining user experience and trust. As a result, accurately determining whether a user has genuinely finished their speaking turn—while still maintaining rapid system responsiveness—has become a critical challenge for building effective and user-friendly voice AI agents.

Turn-taking is a fundamental property of human spoken interaction (Skantze, 2021). In everyday conversation, speaking turns are exchanged naturally with minimal delay. In contrast, due to inherent variability of speech signal and other factors (Pan et al., 2024; Inoue et al., 2024a, 2025), achieving similarly real-time and stable turn management remains challenging. Although existing turn-taking detection approaches (Gu et al., 2024; Skantze and Irfan, 2025) have made notable progress, they still yield erroneous decisions, resulting in excessive response latency, frequent interruptions, and unnatural interaction dynamics that substantially degrade user experience. Therefore, there is an urgent need for real-time and robust turn-taking techniques.

Traditional spoken dialogue systems typically rely on heuristic, silence-based turn-taking strategies (Skantze, 2021). A common approach is to wait for a fixed or adaptive duration of silence in the user's speech before deciding that the turn of user has ended, and then process the input and generate a response. However, silence alone is an unreliable cue for turn completion: users frequently produce within-utterance pauses that do not signal a handover (Majlesi et al., 2023; Inoue et al., 2024a). Previous studies (Stivers et al., 2009; Inoue et al., 2024b) revealed that turn transitions in many languages are rapid, often on the order of 100–500 ms,

suggesting listeners do not merely react to silences, but proactively anticipate upcoming turn completions using a combination of lexical, prosodic, and multimodal cues.

To this end, several works explored data-driven turn-taking models based on linguistic[1][2] or acoustic[3][4] features. However, due to stringent real-time constraints, these methods typically adopt relatively simple model architectures, which limits their ability to capture fine-grained discriminative cues and leaves substantial room for improvement in both detection accuracy and robustness. Motivated by recent success of large language models (LLMs) (Bai et al., 2023; Achiam et al., 2023) and speech language models (SLMs) (Fang et al., 2024; Pan et al., 2025), numerous works (Défossez et al., 2024; Yu et al., 2024; Zhang et al., 2025; Li et al., 2025) directly integrate full-duplex capabilities into LLM or SLM backbones. While such systems can improve detection quality, they exhibit several limitations that hinder practical deployment. First, they require large amounts of manually annotated dialogue data, which is expensive and time-consuming to obtain and difficult to scale across domains and languages. Second, their LLM or SLM backbones need to support multiple functionalities, such as automatic speech recognition (ASR), which often introduces additional latency (Li et al., 2025). Moreover, this architecture inherently prioritizes semantic information, thereby discarding fine-grained acoustic cues that are crucial for accurate turn-taking detection, and consequently suffers from notable performance degradation in complex real-world scenarios.

In summary, this paper makes the following contributions:

- We propose **JAL-Turn**, a lightweight speech-only turn-taking model that jointly leverages pre-trained acoustic and linguistic encoders and supports parallel inference with ASR through encoder sharing, enabling low-latency deployment.
- We introduce a scalable data construction pipeline that automatically derives reliable turn-taking labels from large-scale real-world dialogue corpora without manual annotation, enabling effective training across domains and languages.
- We present extensive experiments on a public multilingual benchmark and an in-house Japanese customer-service corpus, along with ablation and attribution analyses, demonstrating that JAL-Turn consistently outperforms strong audio-only and LLM-based baselines while satisfying real-time constraints.

[1] https://github.com/ten-framework/ten-turn-detection
[2] https://github.com/pipecat-ai/smart-turn/tree/main
[3] https://github.com/pipecat-ai/smart-turn/tree/filipi/smart-turn
[4] https://github.com/inokoj/VAP-Realtime

2 Data Pipeline

To enhance the robustness of our turn-taking model, we develop an efficient data pipeline that automatically derives reliable training labels from large-scale real-world conversational corpora without requiring any manual annotation.

Given stereo conversational audio, we first extract frame-level voice activity detection (VAD) at 50 Hz. For each channel $c \in \{0, 1\}$, we obtain a binary sequence $\mathbf{v}_c = [v_c^1, \ldots, v_c^T] \in \{0, 1\}^T$, where $v_c^t = 1$ indicates speech activity at frame t.

2.1 Future-Window Labeling

Motivated by previous studies on future voice activity projection (Ekstedt and Skantze, 2022), we employ a future-window strategy that leverages upcoming conversational dynamics. For each frame t, we compute a weighted VAD score over a 2-second future window:

$$s_c^t = \sum_{i=0}^{\tau f_s} w(i)\, v_c^{t+i}, \tag{1}$$

where $\tau = 2$ seconds, $f_s = 50$ Hz, and $w(i)$ is a non-negative weighting function. Three temporal weighting schemes—linear, square-root, and exponential—independently assign Hold/Shift labels (Hold if $s_c^t \geq s_{1-c}^t$), and a label is kept only when all schemes agree. This future-window design effectively suppresses backchannels: instantaneous VAD comparisons are easily confounded by brief listener responses, whereas future VAD patterns provide more reliable cues for genuine turn transitions.

2.2 Context Construction

After labeling, training samples are extracted from each VAD falling edge following a speech segment, which corresponds to natural turn-taking decision

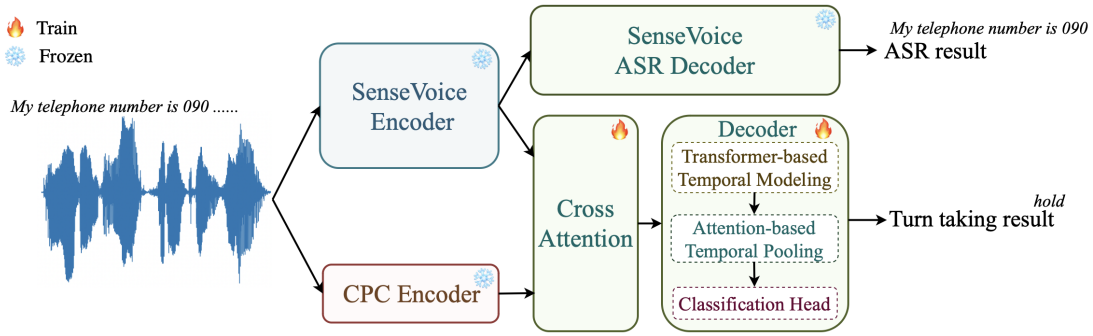

Figure 1: Overall training architecture of the proposed JAL-Turn framework.

points. For each such point, we construct a context window by extending backward to the previous long silence (duration ≥ 2 seconds) and then normalize it to a fixed 10-second length through left-padding or truncation as needed, ensuring that the window always includes a complete preceding speech segment to provide sufficient semantic context for turn prediction.

2.3 Dataset Generation

Applying this pipeline to 1,128 hours of in-house stereo conversational data yields approximately 2,299 hours of trainable segments, with an estimated labeling accuracy of about 85% based on manual inspection.

To further increase data diversity, we also incorporate a 95-hour in-house dataset in which each recording contains a single complete utterance spoken by one speaker (e.g. "My phone number is 090-8987-2023"). Applying the same segmentation procedure produces 749 hours of training segments, where all frames except the final one are labeled as Hold. Although this dataset achieves nearly 100% labeling accuracy, it lacks the conversational variability present in real dialogues.

We therefore train our model on a mixture of both datasets, combining large-scale conversational dynamics with high-quality utterance-level supervision.

3 JAL-Turn

As illustrated in Fig. 1, our proposed **JAL-Turn** framework primarily comprises five components: 1) a dual-path encoder for acoustic and semantic feature extraction, 2) a cross-attention-based fusion module for integrating heterogeneous representations, 3) a self-attention-based Transformer module for temporal modeling, 4) an attention pooling module for utterance-level temporal pooling, and 5) a binary classification head for hold/shift prediction.

3.1 Dual-Encoder Architecture

In contrast to conventional single-encoder approaches that rely on acoustic or linguistic cues, we adopt a dual-path encoder architecture to explicitly capture complementary aspects of the speech signal. Concretely, the primary encoder derives from the pretrained SenseVoice-Small (An et al., 2024) model, a multilingual speech foundation model based on a convolution-augmented transformer backbone (Gulati et al., 2020; Yang et al., 2023), which is particularly effective at extracting high-level semantic and linguistic elements. The secondary encoder is a pretrained contrastive predictive coding (CPC) model (Riviere et al., 2020), which emphasizes low-level acoustic regularities via self-supervised contrastive learning.

Given an input waveform $\mathbf{x} \in \mathbb{R}^{1\times L}$, where L denotes the waveform length, the two encoders produce frame-level feature sequences:

$$\mathbf{h}_l = \text{Encoder}_{\text{Sense}}(\mathbf{x}) \in \mathbb{R}^{T_1\times d_1}, \quad (2)$$

$$\mathbf{h}_a = \text{Encoder}_{\text{CPC}}(\mathbf{x}) \in \mathbb{R}^{T_2\times d_2}, \quad (3)$$

where $d_1 = 512$ and $d_2 = 256$ denote the respective output dimensions. Both encoders are kept frozen during training in order to preserve their pretrained knowledge. We then apply linear projection layers to map the features into a shared latent space. Overall, this design allows JAL-Turn to jointly exploit high-level linguistic cues from SenseVoice and fine-grained acoustic patterns from CPC, yielding richer and more informative representations for turn-taking detection.

During training, both the CPC and SenseVoice

encoders are kept frozen. Importantly, the SenseVoice encoder is shared between ASR and JAL-Turn, enabling turn-taking predictions to be computed synchronously with ASR during inference. This design avoids inserting any additional processing stages before or after ASR decoding, allowing turn-taking decisions and transcriptions to be obtained in parallel from a single forward pass of the shared encoder.

3.2 Cross-Attention-based Fusion

To effectively integrate the heterogeneous features produced by the dual-path encoders, we introduce a cross-attention-based fusion module, enabling JAL-Turn to dynamically attend to the most informative regions across different feature spaces.

Concretely, the fusion module consists of $L = 2$ stacked cross-attention layers. In each layer, the SenseVoice features h'_l act as queries, while CPC features h'_a serve as keys and values:

$$\mathrm{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{4}$$

$$\mathbf{Q} = \mathrm{LayerNorm}(\mathbf{h}'_l) \tag{5}$$

$$\mathbf{K}/\mathbf{V} = \mathrm{LayerNorm}(\mathbf{h}'_a) \tag{6}$$

where $d_k = d/H$ denotes the dimensionality of each head, and d and H denote the model dimension and the number of attention heads, which are set to 256 and 4, respectively. Each layer additionally incorporates residual connections and a position-wise feed-forward network (FFN).

3.3 Transformer-based Module

On top of the fused representation, we employ a causal self-attention-based Transformer module. Notably, we adopt Attention with Linear Biases (ALiBi) (Press et al., 2022) as the positional bias mechanism, which replaces conventional absolute positional embeddings by adding a distance-dependent bias directly to the attention logits. The intuition behind is that ALiBi has been shown to improve length extrapolation and naturally encodes a recency bias, which aligns well with the local temporal patterns that govern turn-taking behavior.

3.4 Attention-based Temporal Pooling Module

To obtain an utterance-level representation while allowing JAL-Turn to automatically focus on the most informative temporal segments, we apply a lightweight attention-based temporal pooling mechanism over the sequence of hidden states $\mathbf{h}tt = 1^T$ produced by the Transformer-like module. Specifically, we compute:

$$\alpha_t = \mathrm{softmax}\left(\mathbf{w}^\top \mathbf{h}_t + b\right) \tag{7}$$

$$\mathbf{h}_{\text{pool}} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{h}_t \tag{8}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are trainable parameters, and α_t denotes the normalized attention weight for time step t.

3.5 Classification Head and Training Objective

Given the pooled representation $\mathbf{h}_{\text{pool}} \in \mathbb{R}^d$, we apply a lightweight linear classification head to produce the shift logit $\hat{y} \in \mathbb{R}$, parameterized by trainable weights $\mathbf{w}_{\text{cls}} \in \mathbb{R}^d$ and bias $b_{\text{cls}} \in \mathbb{R}$. The predicted probability is obtained via a sigmoid activation, and the final decision (hold vs. shift) is made using a fixed threshold $\tau = 0.5$.

The model is trained end-to-end with the standard binary cross-entropy objective over mini-batches, where $y \in \{0, 1\}$ denotes the ground-truth turn label (1 for shift and 0 for hold).

4 Experiments

4.1 Experimental Setups

4.1.1 Datasets

To comprehensively assess the proposed method, we conduct experiments on the public Mandarin dataset Easy-Turn (approximately 1145 hours), multilingual STurn-v3[5,6] dataset (containing approximately 700 hours of speech in 23 languages), and a large-scale in-house corpus of real-world Japanese dialogues introduced in the previous section.

For Easy-Turn and STurn-v3, we use the original test set for evaluation, while splitting the training set into training and validation sets with a 9:1 ratio. For the in-house Japanese corpus, we partition all in-house data into training and validation sets using the same 9:1 split as well. Regarding the test set, we additionally collected 500 samples from real-world business data for evaluation which are

[5] https://huggingface.co/datasets/pipecat-ai/smart-turn-data-v3-train

[6] https://huggingface.co/datasets/pipecat-ai/smart-turn-data-v3-test

labeled by human. These data covered various attributes such as gender, age, and business scenario to ensure a comprehensive evaluation of the proposed method.

4.1.2 Implementation Details

In all experiments, JAL-Turn is trained end-to-end using a single H100 GPU for 10 epochs within the PyTorch framework. We use the AdamW optimizer with an initial learning rate of 1×10^{-4}, a weight decay of 0.001, and a batch size of 64. The learning rate is scheduled using a cosine annealing strategy, decaying to a minimum value of 1×10^{-6}.

Regarding evaluation metrics, we use accuracy and F1-score for turn-taking detection. In addition, we use latency to quantify the responsiveness of the system in full-duplex scenarios.

4.2 Main Results

To comprehensively assess the proposed approach, we evaluate JAL-Turn against three categories of baselines on two public multilingual benchmarks and an in-house Japanese corpus. Specifically, we consider: (1) audio-only methods, including STurn-v2 and STurn-v3; (2) LLM-based pipelines, including GPT-5.1[7], Qwen3-0.6B[8], and Gemini-2.5-Flash[9], where SenseVoice is first used to produce ASR transcripts and the resulting text is then fed into the LLMs for turn-state prediction; and (3) an SLM-based system, represented by EasyTurn (Li et al., 2025), which performs turn-taking detection using a speech language model backbone. GPT-5.1 and Gemini-2.5-Flash are evaluated via their official APIs, whereas Qwen3-0.6B is fine-tuned on the training data and served with vLLM[10] (Kwon et al., 2023) for efficient inference. All LLM-based experiments use the same prompt, provided in Appendix A.

4.2.1 Comparison with SLM-based Systems

We first compare JAL-Turn with representative strong baselines on the Mandarin Easy-Turn corpus (Table 1). Here, Acc_{cp}, Acc_{incp}, Acc_{bc}, and Acc_{wait} denote the turn-taking detection accuracy for the *complete*, *incomplete*, *backchannel*, and *wait* states, respectively (higher is better). JAL-Turn achieves the best performance on the cp state (96.67%) while operating at an extremely low end-to-end latency of 12 ms. Compared with Paraformer (Gao et al., 2022)+TEN Turn Detection and STurn-v2, JAL-Turn yields markedly higher accuracy of the complete and incomplete states, and reduces latency from 204 ms / 27 ms to 12 ms.

Against EasyTurn, JAL-Turn attains comparable performance on `incp` and `wait` (within 4.0 and 6.0 points, respectively) and slightly improves cp accuracy (96.67% vs. 96.33%). However, JAL-Turn underperforms on the bc state (80% vs. 91%), which we conjecture stems from the intrinsically context-dependent nature of backchannels: they are often short, semantically light responses whose role is better determined with explicit lexical/semantic cues. Crucially, despite this gap on bc, JAL-Turn offers a substantially more favorable quality–latency trade-off overall, delivering competitive state-wise accuracy under strict real-time constraints.

4.2.2 Comparison with Audio-only Methods

As shown in Tables 2 and 3, JAL-Turn consistently achieves the best detection accuracy and F1-score among audio-only methods on both the public multilingual benchmark and the in-house Japanese corpus. On the public STurn benchmark, JAL-Turn attains 93.27% accuracy and 0.934 F1, providing small but consistent relative gains of approximately 0.2% in accuracy and 0.3% in F1 over STurn-v3, while outperforming STurn-v2 by about 43% relative accuracy and 38% relative F1. On the in-house Japanese corpus, the improvements are much more substantial: JAL-Turn reaches 92.03% accuracy and 0.925 F1, corresponding to relative gains of roughly 25.6% accuracy and 24.8% F1 over STurn-v3 (73.29%, 0.741), and about 41.3% accuracy and 36.2% F1 over STurn-v2 (65.12%, 0.679).

In terms of efficiency, STurn-v3 achieves the lowest latency on both benchmarks (12 ms and 23 ms), but JAL-Turn still operates comfortably within the real-time regime, with end-to-end latencies of 22 ms on the public dataset and 43 ms on the in-house corpus. Compared with the older STurn-v2 system, JAL-Turn not only delivers dramatically higher accuracy and F1, but also reduces latency by about 85% on the public benchmark (149 ms $\rightarrow$ 22 ms) and by roughly 69% on the in-house corpus (138 ms $\rightarrow$ 43 ms). These results demonstrate that the proposed joint acoustic–linguistic modeling paradigm effectively integrates fine-grained acoustic and linguistic cues, yielding clearly supe-

[7] https://openai.com/zh-Hans-CN/index/gpt-5-1
[8] https://huggingface.co/Qwen/Qwen3-0.6B
[9] https://poe.com/Gemini-2.5-Flash
[10] https://github.com/vllm-project/vllm

Table 1: Performance comparison of turn-taking detection methods on Mandarin Easy-Turn corpus.

Model	Acc_{cp}	Acc_{incp}	Acc_{bc}	Acc_{wait}	Latency (ms)
Paraformer+TEN Turn Detection	86.67	89.3	-	91	204
STurn-v2	78.67	62	-	-	27
Easy-Turn	96.33	**97.67**	**91**	**98**	263
JAL-Turn	**96.67**	93.67	80	92	**12**

Table 2: Performance comparison of audio-only turn-taking detection methods on multilingual STurn-v3.

Model	Acc	F1	Latency (ms)
STurn-v2	65.12	0.679	149
STurn-v3	93.10	0.931	**12**
JAL-Turn	**93.27**	**0.934**	36

Table 3: Performance comparison of audio-only turn-taking detection methods on in-house Japanese corpus.

Model	Acc	F1	Latency (ms)
STurn-v2	55.46	0.427	140
STurn-v3	71.94	0.736	**13**
JAL-Turn	**92.03**	**0.925**	38

rior detection quality while maintaining low end-to-end latency suitable for deployment in real-time dialogue systems.

4.2.3 Comparison with LLM-based Methods

As shown in Table 4, JAL-Turn also compares favorably with LLM-based turn-taking detectors on the in-house benchmark.

Table 4: Performance comparison of JAL-Turn against LLM-based turn-taking detection methods on the in-house benchmark.

Model	Acc	F1	Latency (ms)
Gemini-2.5-Flash	76.91	0.817	595
Qwen3-0.6B	78.70	0.782	124
GPT-5.1	85.52	0.874	1205
JAL-Turn	**92.03**	**0.925**	**38**

Compared with Gemini-2.5-Flash and Qwen3-0.6B, JAL-Turn improves accuracy by 15.1 and 13.3 absolute points (92.03% vs. 76.91% / 78.70%), respectively, and increases F1 by 0.108 and 0.143, while reducing latency from 595 ms and 124 ms to only 38 ms. Relative to GPT-5.1, JAL-Turn further raises accuracy from 85.52% to 92.03% and F1 from 0.874 to 0.925, and lowers latency by more than a factor of five (205 ms $\rightarrow$ 38 ms). These results indicate that JAL-Turn not only matches or surpasses the detection quality of substantially heavier LLM-based pipelines, but also offers an order-of-magnitude lower response latency and avoids the additional overhead of full ASR decoding and large-scale language model inference, making it a lightweight yet competitive alternative for real-time full-duplex dialogue systems.

4.3 Ablation Studies

To assess the contribution of each component in JAL-Turn, we conduct ablation experiments on the in-house Japanese corpus, as summarized in Table 5.

Table 5: Ablation studies of the proposed JAL-Turn. w/o CrossATT denotes using concatenation, w/o ATTPooling represents using the last layer feature.

Model	Acc	F1	Latency (ms)
JAL-Turn	**92.03**	**0.925**	38
w/o Sense	72.01	0.698	12
w/o CPC	84.18	0.839	26
w/o CrossATT	88.59	0.873	41
w/o ATTPooling	90.23	0.895	48

Removing either encoder leads to a clear degradation in detection performance. Dropping the SenseVoice encoder (w/o Sense) causes accuracy to fall from 92.03% to 72.01% and F1 from 0.925 to 0.698, indicating that linguistically enriched representations are the primary driver of performance. Removing the CPC encoder (w/o CPC) is less catastrophic but still non-trivial, reducing accuracy to 84.18% and F1 to 0.839. This shows that CPC contributes complementary fine-grained acoustic cues that further enhance robustness.

Eliminating the cross-attention module (w/o CrossATT) while retaining both encoders also results in a noticeable drop, to 88.59% accuracy and 0.873 F1. This confirms that explicitly modeling interactions between acoustic and linguistic streams is more effective than simply co-presenting their features.

Finally, removing the attention-based temporal pooling (w/o ATTPooling) leads to a moder-

ate degradation in performance (90.23% accuracy and 0.895 F1) and simultaneously increases latency from 38 ms to 48 ms. Thus, the proposed lightweight attention pooling not only yields more informative utterance-level representations, but also provides a better accuracy–latency trade-off than simpler temporal aggregation schemes.

4.4 Analysis

To investigate how the proposed model exploits acoustic and linguistic cues for turn-taking prediction, we analyze both encoder-level and temporal-level contributions on the STurn-v3 and our in-house Japanese test sets. We adopt a unified gradient-based attribution framework to quantify representation-level contributions, complemented by controlled encoder ablations to assess the functional role of each information source.

4.4.1 Encoder-Level Contribution

For each sample, we backpropagate the classification score to the encoder outputs and compute encoder-wise scores via the element-wise gradient–activation product. The contribution ratio for encoder i is defined as

$$\rho_i = \frac{\|\nabla_{x_i} \odot x_i\|_2}{\sum_j \|\nabla_{x_j} \odot x_j\|_2},$$

where x_i denotes the output features of encoder i. We group samples into four duration bins: 0–3 s, 3–6 s, 6–9 s, and >9 s.

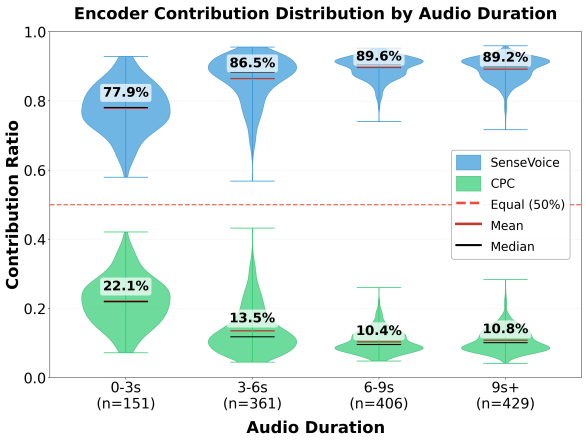

Figure 2: Encoder-wise contribution ratios across utterance durations for SenseVoice (left) and CPC (right).

As shown in Fig. 2, SenseVoice clearly dominates across all bins, with ρ_{Sense} increasing from roughly 0.78 (0–3 s) to 0.90 (6–9 s), indicating that linguistically enriched representations are the primary driver of turn-taking decisions and become even more influential for longer contexts. CPC exhibits complementary but smaller contributions, with ρ_{CPC} decreasing from about 0.22 (0–3 s) to 0.10 (6–9 s) as duration grows, suggesting that prosodic and low-level acoustic cues are most useful for short utterances when semantic context is limited. Overall, this reveals a length-dependent division of labor: SenseVoice provides the dominant semantic signal, while CPC supplies auxiliary acoustic evidence, particularly in short-context scenarios.

4.4.2 Temporal-Level Contribution

To further analyze how the model allocates attention over time, we perform *position-regularized temporal attribution* using the gradient–input product. Temporal contribution scores are projected onto a normalized 0–100% axis to facilitate direct comparison across variable-length utterances.

Across all three languages, contributions from the SenseVoice encoder exhibit a consistent monotonic increase toward the end of the utterance, revealing a strong bias toward utterance-final regions for turn-taking prediction (Fig. 3). However, the degree of temporal concentration varies substantially across languages. In Japanese, the majority of the SenseVoice contribution is concentrated within the final **5%** of the audio context. In contrast, Chinese shows a broader concentration around the final **10%**, while English displays a more gradual accumulation, with peak contributions occurring around the final **25%** of the utterance.

This temporal localization reflects language-specific turn-completion mechanisms. In Japanese, turn-taking cues are tightly associated with utterance-final phonetic structures and sentence-final morphemes, such as polite-form endings (e.g., "*-masu*" and "*-desu*"), which are fully realized only in *Shift* instances. As a result, the model exhibits highly concentrated attention near the end of the utterance, particularly for *Shift* predictions. In contrast, English relies more heavily on anticipatory prosodic patterns that unfold over a longer temporal span, while Chinese again occupies an intermediate position.

Notably, the CPC encoder does not exhibit a comparable temporal bias. Its contribution remains relatively uniform across the entire utterance for all three languages, indicating that low-level acoustic representations primarily capture global prosodic characteristics rather than temporally localized turn-completion cues. Together, these findings

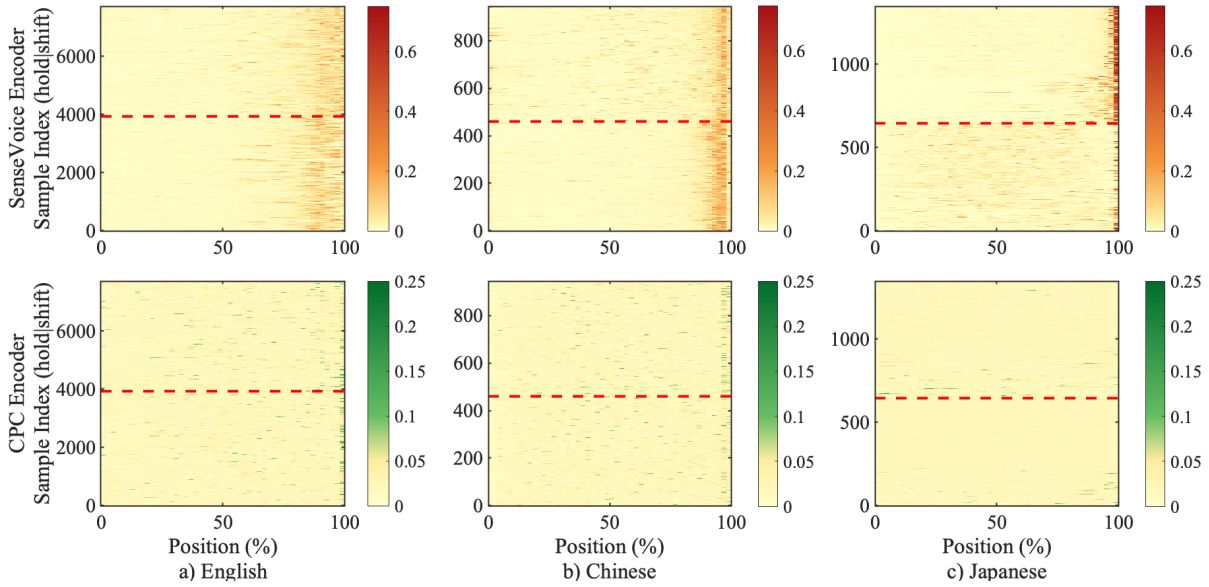

Figure 3: Temporal-level attribution heatmap of three language (normalized to 0–100%). The upper and lower rows correspond to SenseVoice and CPC encoders, respectively. Red dashed lines indicate the boundary between Hold and Shift samples.

suggest that SenseVoice is responsible for modeling temporally localized, language-dependent turn-taking signals, whereas CPC provides complementary, globally distributed acoustic information. Across languages, these findings further imply that SenseVoice, as an ASR encoder, primarily encodes rich phonetic and sub-lexical information, rather than high-level semantic representations in the traditional sense. Although its overall contribution is relatively smaller, CPC complements SenseVoice by encoding broader prosodic variations that are not fully captured by ASR-style encoders, thereby enriching the model's representation space along additional acoustic dimensions.

5 Conclusion

In this study, we present JAL-Turn, a lightweight and robust turn-taking detection framework for full-duplex spoken dialogue systems. By jointly and adaptively modeling acoustic and linguistic elements of the given utterance, together with lightweight transformer and temporal attention pooling modules, JAL-Turn supports low-latency and accurate prediction of turn taking decisions. A scalable data construction pipeline further enables automatic extraction of reliable turn-taking labels from large-scale real-world corpora. Experiments on multilingual public benchmarks and an in-house Japanese customer-service dataset show that JAL-Turn consistently outperforms strong baselines in both accuracy and robustness while maintaining real-time performance.

6 Limitations

Our work has several limitations. First, our scalable data pipeline derives training supervision from stereo VAD with a future-window rule, yielding an estimated labeling accuracy of ∼85% based on manual inspection. This supervision is therefore noisy and only indirectly reflects pragmatic turn-taking intent: ambiguous overlaps, long hesitations, or VAD errors may lead to mislabeled hold/shift decisions. Moreover, the future-window labeling explicitly suppresses backchannels, which can limit the model's ability to recognize backchannel-specific interaction patterns and partially explains the performance gap on the backchannel state observed on Easy-Turn. Second, JAL-Turn focuses on binary hold/shift prediction with a fixed-length (10 s) context construction. While effective for low-latency deployment, this formulation does not fully capture richer turn-management behaviors (e.g., multi-state dialog acts or explicit backchannel prediction), and extending the method to finer-grained conversational states remains future work. Third, part of our evaluation is conducted on an in-house Japanese customer-service corpus with a manually labeled test subset, which may limit full

reproducibility and independent verification.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Erik Ekstedt and Gabriel Skantze. 2022. Voice activity projection: Self-supervised learning of turn-taking events. *arXiv preprint arXiv:2205.09812*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*.

Nianlong Gu, Kanghwi Lee, Maris Basha, Sumit Kumar Ram, Guanghao You, and Richard HR Hahnloser. 2024. Positive transfer of the whisper speech transformer to human and animal voice activity detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7505–7509. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. Real-time and continuous turn-taking prediction using voice activity projection. *arXiv preprint arXiv:2401.04868*.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024b. 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)*, pages 11873–11883.

Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2025. Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7171–7181.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Guojian Li, Chengyou Wang, Hongfei Xue, Shuiyuan Wang, Dehui Gao, Zihan Zhang, Yuke Lin, Wenjie Li, Longshuai Xiao, Zhonghua Fu, and 1 others. 2025. Easy turn: Integrating acoustic and linguistic modalities for robust turn-taking in full-duplex spoken dialogue systems. *arXiv preprint arXiv:2509.23938*.

Ali Reza Majlesi, Ronald Cumbal, Olov Engwall, Sarah Gillet, Silvia Kunitz, Gustav Lymer, Catrin Norrby, and Sylvaine Tuncer. 2023. Managing turn-taking in human-robot interactions: the case of projections and overlaps, and the anticipation of turn design by human participants. *Social Interaction. Video-based Studies of Human Sociality*, 6(1).

Yu Pan, Yuguang Yang, Yanni Hu, Jianhao Ye, Xiang Zhang, Hongbin Zhou, Lei Ma, and Jianjun Zhao. 2025. S2st-omni: An efficient and scalable multilingual speech-to-speech translation framework via seamlessly speech-text alignment and streaming speech decoder. *arXiv preprint arXiv:2506.11160*.

Yu Pan, Yuguang Yang, Yuheng Huang, Tiancheng Jin, Jingjing Yin, Yanni Hu, Heng Lu, Lei Ma, and Jianjun Zhao. 2024. Gmp-tl: Gender-augmented multi-scale pseudo-label enhanced transfer learning for speech emotion recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 496–501. IEEE.

Ofir Press, Noah A Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418. IEEE.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.

Gabriel Skantze 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)*, pages 859–868. IEEE.

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, and 1 others. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Yuguang Yang, Yu Pan, Jingjing Yin, Jiangyu Han, Lei Ma, and Heng Lu. 2023. Hybridformer: Improving squeezeformer with hybrid attention and nsr mechanism. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. 2024. Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and 1 others. 2025. Omniflatten: An end-to-end gpt model for seamless voice conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14570–14580.

A Prompt for LLM-based Turn-Taking Detection

You are analyzing a conversation turn. Your task is to determine whether the current speaker should continue speaking or should stop and let the other speaker talk.

Rules: If the current input text indicates that the speaker has NOT finished their thought or sentence (the speech is incomplete or continuing), output: hold. If the current input text indicates that the speaker HAS finished their thought or sentence (the speech is complete and ready for the other speaker to respond), output: shift.

Analyze the following text and output your decision.