Segmentation Strategy Matters: Benchmarking Whisper on Persian YouTube Content

Reihaneh Iranmanesh[1*], **Rojin Ziaei**[1*], **Joe Garman**[1]
[1]Georgetown University
{ri164, nz204, jhg28}@georgetown.edu

Abstract

Automatic Speech Recognition (ASR) transcription accuracy remains highly sensitive to audio segmentation strategies, yet most benchmarks assume oracle timestamps unavailable in deployment. We systematically evaluate how audio segmentation affects Whisper's performance on 10 hours of Persian YouTube content, comparing transcript-aligned (oracle) versus silence-based (realistic) approaches across contrasting acoustic conditions. Results reveal striking content-type dependency: podcast content benefits from timestamp segmentation (33% lower mean WER), while entertainment content favors silence-based segmentation (8% lower mean WER). This finding demonstrates that optimal segmentation must be content-aware, with silence detection better capturing natural boundaries in acoustically heterogeneous media while avoiding mid-utterance splits. We publicly release our evaluation framework, 10 hours of audio with gold transcripts, and segmentation results here: https://github.com/ri164-bolleit/persian-youtube-whisper-benchmark

1 Introduction

Automatic Speech Recognition (ASR) systems have achieved remarkable progress through neural architectures, large-scale training data, and robust pretraining objectives [Prabhavalkar et al., 2023, Malik et al., 2021, Kheddar et al., 2024, Radford et al., 2023]. Despite these advances, transcription accuracy in real-world settings remains highly sensitive to upstream design choices, particularly how continuous audio streams are segmented prior to inference [Kuhn et al., 2024, Arriaga et al., 2024]. Segmentation affects not only computational efficiency but also linguistic context, temporal alignment, and the stability of downstream evaluation metrics.

Many existing benchmarks rely on oracle transcript timestamps to define segment boundaries, an assumption that rarely holds outside controlled evaluation settings [Faria et al., 2022, Sklyar et al., 2022, von Neumann et al., 2023]. This creates a gap between benchmark performance and real-world deployment accuracy, where segmentation must be inferred directly from acoustic signals.

1.1 Research Questions

We address three key questions:
(RQ1) How does segmentation strategy–oracle timestamps versus acoustic silence detection–impact ASR accuracy?
(RQ2) Does the optimal strategy vary by content type (structured podcasts versus noisy entertainment)?
(RQ3) How does model scale affect robustness to segmentation choices?

1.2 Contributions

Using OpenAI Whisper at three scales (small: 0.2B, medium: 0.8B, large: 1.5B parameters), we evaluate two segmentation pipelines on a 10-hour Persian YouTube dataset spanning contrasting acoustic conditions. Our contributions include: (1) first systematic evaluation of segmentation strategies for Persian ASR, (2) evidence that optimal approaches are fundamentally content-type dependent, and (3) a publicly released benchmark addressing the critical gap in Persian real-world ASR evaluation.

2 Related Work

2.1 ASR Benchmarking and Evaluation

The development of robust ASR benchmarks has emerged as a critical research area, with increasing recognition of their limitations in capturing real-world speech variability [Aksënova et al., 2021, Szymański et al., 2020]. Traditional benchmarks such as LibriSpeech [Panayotov et al., 2015] pri-

[*] Equal contribution.

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 121–130
March 29, 2026 ©2026 Association for Computational Linguistics

marily focus on clean, well-structured read speech, rendering them inadequate for assessing ASR performance in spontaneous or conversational settings. Studies by [Del Rio et al., 2021] and [Cao et al., 2023] underscore the need for benchmarks that evaluate ASR systems under challenging conditions, including spontaneous speech, noisy environments, and multi-speaker interactions.

To address domain mismatch and improve cross-domain generalization, the End-to-End Speech Benchmark (ESB) [Gandhi et al., 2022] was introduced to evaluate ASR models across diverse domains without prior knowledge of data distributions. However, while ESB accounts for varied acoustic environments, it overlooks critical demographic factors such as speaker age, gender, and regional accents that significantly impact ASR performance [Aksënova et al., 2021].

2.2 Segmentation Strategies in ASR

While considerable attention has focused on model architecture and training data, the impact of audio segmentation strategies on ASR performance remains understudied. Most benchmarks assume oracle timestamps that align with transcript boundaries, an assumption that rarely holds in deployment [Faria et al., 2022, Sklyar et al., 2022, von Neumann et al., 2023]. Previous research [Kuhn et al., 2024, Arriaga et al., 2024] notes that standard benchmarks often fail to account for WER variability across segmentation approaches, creating a gap between benchmark performance and real-world accuracy.

This gap is particularly problematic as segmentation affects not only computational efficiency but also linguistic context preservation, temporal alignment quality, and the stability of evaluation metrics. In real-world deployments, segmentation must be inferred directly from acoustic signals using methods such as voice activity detection or silence-based splitting, yet most published results rely on oracle segmentation that provides an optimistic upper bound on achievable performance.

2.3 Persian ASR Resources and Challenges

For low-resource languages like Persian, comprehensive benchmarks remain scarce despite the language being spoken by over 100 million people. Existing Persian ASR datasets provide partial foundations but have significant limitations. The DeepMine dataset [Zeinali et al., 2019] offers a robust foundation with over 1,850 speakers and approximately 480 hours of audio, but its six-hour test set focuses primarily on formal speech, limiting applicability to informal or spontaneous contexts. Similarly, the Persian subset of Mozilla Common Voice [Ardila et al., 2019] suffers from demographic imbalances and data quality concerns due to its crowdsourcing approach. The FLEURS dataset [Conneau et al., 2023], while useful for few-shot learning, focuses on short, controlled utterances insufficient for evaluating conversational speech. [Sedghiyeh et al., 2025] introduce PSRB, a comprehensive benchmark for evaluating Persian ASR systems across diverse linguistic conditions including regional accents, speaker demographics, and acoustic environments. However, it does not explore segmentation strategies for handling long-form audio or optimal chunking approaches for Persian speech recognition.

Persian poses unique ASR challenges due to its linguistic diversity-spanning regional accents like Baluchi, Kurdish, and Dari-and features like variable word boundaries and the use of Zero Width Non-Joiner (ZWNJ) characters [Ghayoomi and Momtazi, 2009, Bijankhan et al., 2011]. These characteristics, combined with limited training data, exacerbate performance disparities in Persian ASR systems. Most existing Persian ASR research focuses on controlled conditions, with YouTube content representing a particularly challenging and underexplored domain combining spontaneous speech, variable acoustic quality, background music, and diverse speaking styles.

Our work addresses critical gaps in the literature by providing the first systematic evaluation of segmentation strategies for Persian ASR on real-world YouTube content, using two different segmentation approaches and various model scales.

3 Dataset

We curated a 10-hour audio–text dataset sourced from YouTube, designed to capture contrasting acoustic and conversational conditions. The dataset consists of two subsets, each totaling approximately five hours.

3.1 Roud Podcast Subset

Roud Podcast[1]: A conversational podcast featuring a single speaker in a structured format centered on book discussions. This content contains no background music, no speaker overlap, and min-

[1] https://www.youtube.com/@MojtabaShakoori

imal variability in acoustic conditions, providing a relatively clean and controlled speech environment.

Corpus Statistics: As shown in Table 1, this data includes six episodes totaling 5.1 hours containing 2,747 transcript-aligned segments. Segment lengths show low variance (mean: 16.9 words, SD: 5.5, range: 1–36), reflecting structured discourse with natural prosodic boundaries.

YouTube timestamps are generated by the content creator, naturally aligning with clause endings and prosodic boundaries. This production-time alignment creates favorable conditions for timestamp-based segmentation.

3.2 Kouman Entertainment Subset

Kouman Entertainment[2]: An entertainment-focused YouTube channel characterized by multiple hosts, frequent speaker changes, background music, and unpredictable audio content. This subset represents a substantially noisier and more heterogeneous speech setting.

Corpus Statistics: As shown in Table 2, this data includes ten episodes totaling 4.9 hours and containing 6,450 transcript segments. Segment lengths exhibit high variance (mean: 4.4 words, SD: 6.1, median: 1 word), reflecting fragmented timestamp-based splitting that produces single-word or phrase-level segments.

Episode	Segments	Words	Duration (min)
Episode 1	325	6,133	∼41
Episode 2	333	3,483	∼23
Episode 3	490	8,481	∼57
Episode 4	533	9,238	∼62
Episode 6	475	8,056	∼54
Episode 8	591	10,911	∼73
Total	**2,747**	**46,302**	**∼310**

Table 1: Roud Podcast corpus statistics. Type-Token Ratio: 0.142. Segment length: median=16, IQR=14–20 words. Timestamps created during production align with natural discourse boundaries.

Unlike Roud where timestamps align with discourse structure, Kouman timestamps are editorial markers added post-production for viewer navigation. These frequently split continuous speech during music-overlaid dialogue or multi-speaker exchanges, creating artificial mid-utterance boundaries. The mean 5.5:1 timestamp-to-silence ratio quantifies this fragmentation: timestamp segmentation produces more segments than acoustic boundaries warrant.

Episode	Segments	Words	Avg W/Seg	TS:Sil.
amazon	799	2,057	2.6	6.5:1
eshgh	655	3,624	5.5	6.9:1
hoosh	338	1,411	4.2	1.8:1
madrese	751	2,331	3.1	6.7:1
mia	693	1,391	2.0	7.2:1
nooshabe	901	4,827	5.4	6.6:1
norooz	718	3,191	4.4	7.7:1
safar	325	2,857	8.8	2.9:1
soal	827	2,514	3.0	6.9:1
youtuber	443	4,279	9.7	2.0:1
Total/Avg	**6,450**	**28,482**	**4.4**	**5.5:1**

Table 2: Kouman Entertainment corpus statistics. Ten episodes, approximately 4.9 hours. Type-Token Ratio: 0.276. TS:Silence Ratio indicates timestamp segmentation produces 5.5× more segments than silence-based segmentation (mean), revealing severe over-fragmentation. Segment length: min=0, max=64, median=1 word.

3.3 Data Preparation

For both subsets, the original YouTube videos were converted into WAV audio files (16 kHz, mono). Corresponding time-stamped transcripts were retrieved directly from YouTube and stored as CSV files aligned with each audio file. All transcripts were manually created by channel managers and content creators as YouTube lacks auto-transcription for Persian. All spoken content is in standard Persian.

4 Methodology

We evaluate automatic speech recognition performance using two distinct experimental pipelines that differ in how audio is segmented and aligned with reference human-generated transcripts. Our pipeline employs three OpenAI Whisper models [Radford et al., 2023] (small, medium, and large) and is evaluated using Word Error Rate (WER) and Character Error Rate (CER).

4.1 Transcript-Aligned Timestamp Segmentation

In the first setup, audio segmentation is directly derived from YouTube's time-stamped reference transcripts. Each transcript CSV consists of alternating timestamp and text lines, where timestamps indicate the start time of each spoken segment. Segment end times are inferred from the subsequent

[2] https://www.youtube.com/c/Kouman

timestamp, with the final segment extending to the end of the audio file.

Using these timestamps, the corresponding WAV audio files are segmented exactly to match the transcript boundaries. Each segment is then transcribed independently using Whisper. Since the audio segments and reference transcripts are perfectly aligned in time by construction, evaluation is performed via a one-to-one comparison between each reference segment and its corresponding Whisper transcription. This pipeline provides an optimistic estimate of transcription performance under ideal alignment conditions.

4.2 Silence-Based Segmentation

The second setup removes reliance on transcript-based segmentation and instead segments audio using acoustic silence detection. Audio is split into non-silent regions using energy-based thresholds, followed by post-processing to enforce segment duration constraints between 5 and 30 seconds. Short segments are merged, long segments are split, and brief silences are optionally retained at segment boundaries to preserve natural speech context.

Because these segments do not necessarily align with the reference transcript timestamps, evaluation requires a fuzzy time-alignment procedure. For each predicted audio segment, a reference text is constructed by concatenating all transcript entries whose timestamps overlap with the segment's time window. This pipeline more closely reflects real-world transcription scenarios.

5 Results

Here, we report results for the Roud podcast, a single-speaker, low-noise conversational dataset, and Kouman, a noisy, multi-speaker entertainment channel, using two segmentation strategies.

5.1 Segment Length and Granularity

The two segmentation strategies produce substantially different segment distributions for both datasets. Transcript-based segmentation yields a larger number of shorter segments, tightly aligned to sentence- or phrase-level transcript boundaries. In contrast, silence-based segmentation produces fewer but longer segments by merging contiguous speech between silence intervals and enforcing minimum and maximum duration constraints (5–30 seconds).

Timestamp-based segmentation produces 2.5–8× more segments per episode (319–901 segments) compared to silence-based segmentation (93–216 segments). This fragmentation creates shorter, more numerous boundaries that may interrupt mid-utterance in content with music, sound effects, and overlapping speech.

Across all evaluated episodes for both Roud and Kouman, silence-based segmentation consistently results in a lower total number of segments per episode, indicating longer average segment durations. This difference in granularity has direct implications for both transcription quality and evaluation stability, as longer segments preserve more linguistic context while increasing acoustic variability.

5.2 Roud Podcast: Clean Speech Results

Table 3 presents comprehensive results for the Roud podcast across six episodes and three model sizes. Across all episodes, performance improves monotonically with model size. Whisper large consistently achieves the lowest WER and CER, followed by medium and small.

We assess the statistical significance of performance differences using both parametric (independent two-sample t-tests) and non-parametric (Mann–Whitney U) tests on per-segment WER and CER distributions. Across evaluated episodes and models, differences between silence-based and timestamp-based segmentation are statistically significant at $\alpha = 0.05$ for both WER and CER.

Effect size analysis using Cohen's d indicates small-to-moderate effects, with stronger effects observed for WER than CER. In all cases, silence-based segmentation exhibits statistically significant degradation relative to transcript-based timestamp segmentation.

Paired Episode-Level Analysis:

Figures 1a and 1b show paired comparisons for the Whisper *large* model, while Figures 1c and 1d report the same analysis for the *medium* model.

For the large model, the mean paired improvement from silence-based to timestamp-based segmentation is 0.161 in WER (33% relative improvement) and 0.227 in CER (67% relative improvement), with paired t-tests yielding $p < 0.001$ and Wilcoxon signed-rank tests yielding $p = 0.031$. Similar trends are observed for the medium model, with mean paired improvements of 0.084 in WER and 0.167 in CER. The performance gap between segmentation strategies widens as model size increases, suggesting that larger models are more capable of exploiting high-quality segmentation

Episode	Model	Timestamp Seg.			Silence Seg.		
		WER	CER	Segs	WER	CER	Segs
1	large	**0.349**	0.114	325	0.505	0.339	139
1	medium	0.499	0.201	325	0.582	0.364	139
1	small	0.653	0.259	325	0.700	0.410	139
2	large	**0.337**	0.110	333	0.409	0.225	94
2	medium	0.480	0.187	333	0.494	0.258	94
2	small	0.656	0.282	333	0.623	0.310	94
3	large	**0.358**	0.122	490	0.527	0.373	266
3	medium	0.555	0.232	490	0.629	0.412	266
3	small	0.703	0.305	490	0.741	0.464	266
4	large	**0.331**	0.124	533	0.504	0.371	278
4	medium	0.500	0.227	533	0.585	0.400	278
4	small	0.653	0.285	533	0.697	0.440	278
6	large	**0.268**	0.098	475	0.484	0.372	263
6	medium	0.417	0.178	475	0.553	0.394	263
6	small	0.586	0.238	475	0.667	0.435	263
8	large	**0.307**	0.098	591	0.490	0.349	289
8	medium	0.476	0.184	591	0.590	0.383	289
8	small	0.640	0.254	591	0.700	0.427	289
Average (large)		**0.325**	0.111	458	0.487	0.338	222
Average (medium)		0.488	0.202	458	0.572	0.369	222
Average (small)		0.649	0.271	458	0.688	0.431	222

Table 3: Detailed results for Roud podcast across all episodes and models. Timestamp segmentation consistently outperforms silence-based segmentation across all three models. Whisper large achieves the lowest WER (and CER) across all episodes using timestamp segmentation.

while being penalized more by noisy segmentation.

5.3 Kouman Entertainment: Noisy Speech Results

Table 4 presents comprehensive results for the Kouman entertainment channel across ten episodes. Strikingly, the performance pattern reverses from the podcast results: silence-based segmentation now *outperforms* timestamp-based segmentation in 7 out of 10 episodes for both model sizes. (Results for Whisper Small are excluded as the model achieved WER values above 0.90.)

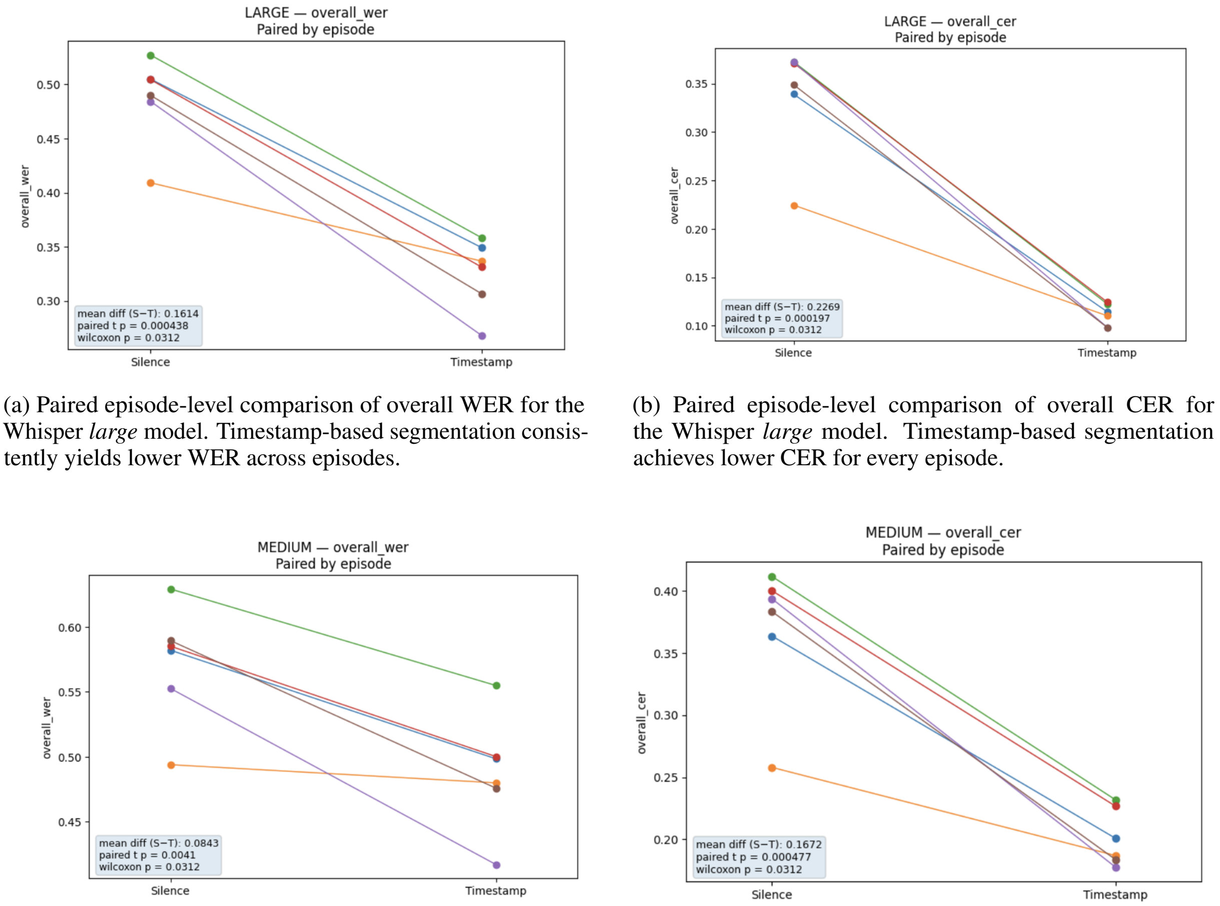
For the large model, WER differences are statistically significant (paired t-test $p = 0.020$, Wilcoxon $p = 0.027$), with silence-based segmentation achieving a mean improvement of 0.0604 (7.9% relative improvement) over timestamp-based segmentation. CER differences are not statistically significant (paired t-test $p = 0.476$, Wilcoxon $p = 1.000$), indicating that character-level accuracy is relatively stable across segmentation methods for larger models.

For the medium model, WER differences are highly significant (paired t-test $p = 0.007$, Wilcoxon $p = 0.027$), with silence-based segmentation showing a mean improvement of 0.0488 (5.7% relative improvement). CER differences approach significance (paired t-test $p = 0.114$, Wilcoxon $p = 0.037$), with a mean improvement of 0.0439.

Paired Episode-Level Analysis:

Figures 2a and 2b show paired comparisons for the Whisper *large* model, while Figures 3a and 3b report the same analysis for the *medium* model. Each line corresponds to a single episode, connecting silence-based results to timestamp-based results.

Across most (seven out of ten) episodes and both model sizes, silence-based segmentation yields lower or comparable WER and CER compared to timestamp-based segmentation. This trend contrasts with expectations that transcript-aligned segmentation would uniformly outperform silence-based approaches. Paired statistical tests confirm that for WER, silence-based segmentation significantly outperforms timestamp-based segmentation. For the *large* model, the mean paired difference (Silence minus Timestamp) is -0.0604 in WER and -0.0191 in CER, with WER differences reaching statistical significance at $\alpha = 0.05$. Similar trends are observed for the *medium* model, with mean paired differences of -0.0488 in WER and



(a) Paired episode-level comparison of overall WER for the Whisper *large* model. Timestamp-based segmentation consistently yields lower WER across episodes.

(b) Paired episode-level comparison of overall CER for the Whisper *large* model. Timestamp-based segmentation achieves lower CER for every episode.

(c) Paired episode-level comparison of overall WER for the Whisper *medium* model. Improvements from timestamp-based segmentation remain consistent across episodes, but are smaller than for the *large* model, indicating increased sensitivity to segmentation at lower model capacity.

(d) Paired episode-level comparison of overall CER for the Whisper *medium* model. Timestamp-based segmentation consistently reduces CER across episodes, with statistically significant paired differences.

Figure 1: Episode-level paired comparison of WER (left) and CER (right) for the Whisper *large* (top) and *medium* (bottom) models. Each line corresponds to a single episode, connecting silence-based segmentation to YouTube timestamp–based segmentation. Reported statistics include the mean paired difference (Silence – Timestamp), paired t-test p-value, and Wilcoxon signed-rank p-value.

-0.0439 in CER, both showing statistically significant improvements for silence-based segmentation in WER.

These results suggest that for the Kouman dataset, silence-based segmentation produces segments that are better suited to the acoustic and linguistic characteristics of the content, possibly by avoiding mid-utterance boundaries introduced by timestamp-based segmentation.

5.4 Error Distribution Analysis

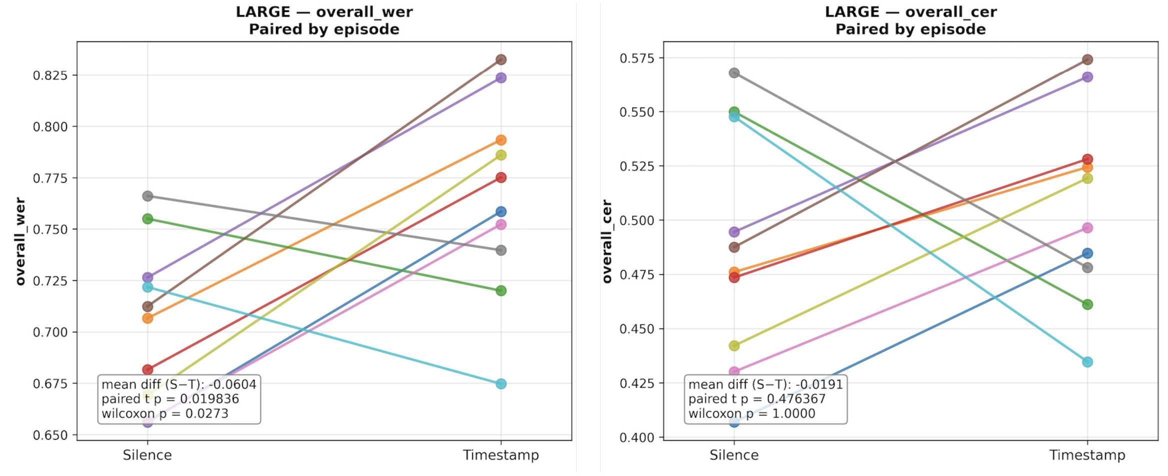
While silence-based segmentation generally outperforms timestamp-based segmentation for entertainment content, three episodes (Hoosh, Safar, YouTuber) exhibit degradation in average WER. The "YouTuber" episode shows the most degradation: for the large model, WER increased by 0.0471 (6.98%) and CER by 0.1129 (25.97%).

However, error distribution analysis reveals a critical engineering tradeoff. For the YouTuber episode, silence-based segmentation produces dramatically tighter distributions with lower variance compared to timestamp-based segmentation.

Variance-Bias Tradeoff: This pattern represents a fundamental tradeoff-silence-based segmentation trades marginally higher average error (6–26% degradation in these three episodes) for significantly improved consistency and predictability. The absence of extreme outliers suggests that even when silence detection produces suboptimal boundaries for content with rapid speech or heavy background audio, the resulting errors remain bounded.

Episode	Model	Timestamp Seg.			Silence Seg.		
		WER	CER	Segs	WER	CER	Segs
amazon	large	0.759	0.485	794	**0.656**	0.407	123
amazon	medium	0.848	0.563	794	0.769	0.466	123
eshgh	large	0.793	0.524	654	**0.707**	0.476	95
eshgh	medium	0.872	0.607	654	0.817	0.538	95
hoosh	large	**0.720**	0.461	337	0.755	0.550	183
hoosh	medium	0.819	0.527	337	0.833	0.594	183
madrese	large	0.775	0.528	748	**0.682**	0.474	111
madrese	medium	0.862	0.617	748	0.781	0.523	111
mia	large	0.824	0.566	692	**0.726**	0.495	96
mia	medium	0.895	0.637	692	0.817	0.537	96
nooshabe	large	0.833	0.574	901	**0.712**	0.488	137
nooshabe	medium	0.883	0.637	901	0.809	0.532	137
norooz	large	0.752	0.496	714	**0.657**	0.430	93
norooz	medium	0.854	0.588	714	0.772	0.489	93
safar	large	**0.740**	0.478	319	0.766	0.568	109
safar	medium	0.837	0.560	319	0.851	0.625	109
soal	large	0.786	0.519	821	**0.669**	0.442	119
soal	medium	0.861	0.594	821	0.777	0.509	119
youtuber	large	**0.675**	0.435	435	0.722	0.548	216
youtuber	medium	0.798	0.509	435	0.816	0.588	216
Average (large)		0.766	0.507	642	**0.705**	0.488	128
Average (medium)		0.853	0.584	642	0.804	0.540	128

Table 4: Detailed results for Kouman entertainment channel. Bold indicates lowest WER performance. Silence-based segmentation outperforms timestamp in 7/10 episodes for both models.



(a) Paired episode-level comparison of overall WER for the Whisper *large* model.

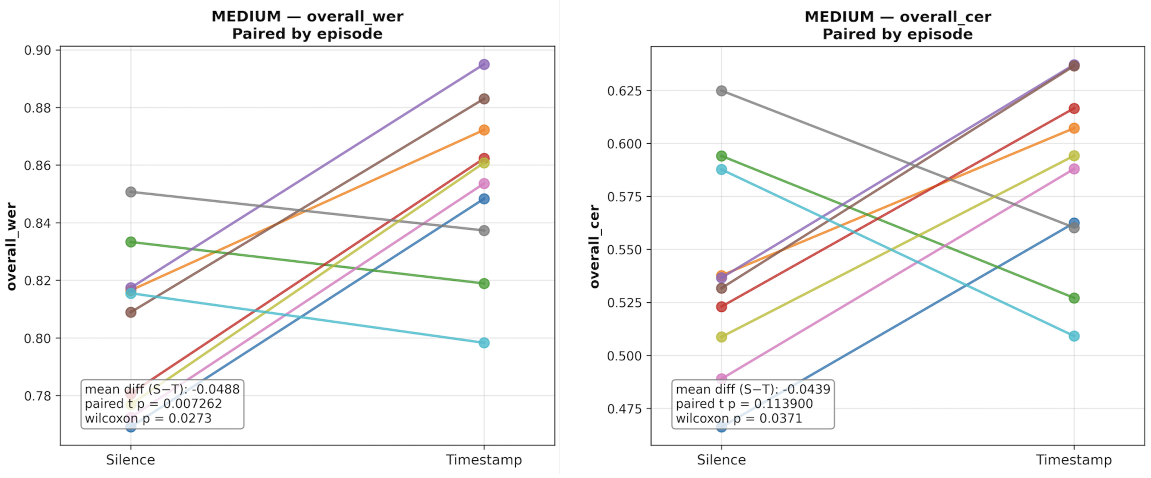
(b) Paired episode-level comparison of overall CER for the Whisper *large* model.

Figure 2: Kouman's Episode-level paired comparison of WER (left) and CER (right) for the Whisper *large* model. Each line corresponds to a single episode, connecting silence-based segmentation to YouTube timestamp–based segmentation. Reported statistics include the mean paired difference (Silence – Timestamp), paired t-test p-value, and Wilcoxon signed-rank p-value.

6 Discussion

6.1 Content-Type Dependency

Our results reveal that optimal segmentation strategy is fundamentally content-type dependent. For podcast content (Roud), timestamp-based segmentation significantly outperforms silence-based approaches across all models, with mean improvements of 33% in WER and 67% in CER.



(a) Paired episode-level comparison of overall WER for the Whisper *medium* model.

(b) Paired episode-level comparison of overall CER for the Whisper *medium* model.

Figure 3: Kouman's Episode-level paired comparison of WER (left) and CER (right) for the Whisper *medium* model. Each line corresponds to a single episode, connecting silence-based segmentation to YouTube timestamp–based segmentation. Reported statistics include the mean paired difference (Silence – Timestamp), paired t-test p-value, and Wilcoxon signed-rank p-value.

Conversely, for entertainment content (Kouman), silence-based segmentation achieves lower WER in 70% of episodes, with mean improvements of 7.9% for large models and 5.7% for medium models.

These results reflect fundamental differences in how timestamps relate to acoustic structure. Podcast timestamps, created during production, coincide naturally with speaker pauses and phrase boundaries–they capture the same acoustic events that silence detection would identify. Entertainment timestamps, added post-production, often split continuous audio arbitrarily, particularly during music, sound effects, or overlapping dialogue.

Silence detection inherently respects acoustic structure, avoiding mid-word or mid-phrase splits even when timestamps don't align with natural boundaries. This explains why longer, acoustically-grounded segments outperform shorter, arbitrarily-aligned ones in heterogeneous audio–they preserve utterance integrity and reduce boundary-induced errors.

6.2 Model Scaling Effects

Performance improves consistently as model size increases, regardless of content type or segmentation strategy. Importantly, larger models do not eliminate the content-type effect: podcasts continue to benefit from timestamp segmentation while entertainment content favors silence-based segmentation, even at the largest model size (1.5B parameters). This persistence confirms that the segmentation preferences reflect systematic differences rather than limitations that could be overcome with more model capacity.

For Roud, the absolute segmentation gap increases from small to large models, and the relative gap also widens indicating that while larger models achieve better absolute accuracy, they remain sensitive to segmentation quality. For Kouman, gaps remain proportionally consistent, suggesting that the acoustic boundary alignment effect scales with model capacity.

Larger models demonstrate superior generalization across both segmentation approaches, with lower variance in error rates across episodes. This robustness is particularly evident in Kouman's challenging acoustic conditions, where the large model shows more stable performance across diverse episodes.

While silence-based segmentation consistently underperformed for podcast content, its superiority for entertainment content appears counter-intuitive. The key difference lies in timestamp semantics: podcast timestamps are production-aligned, coinciding naturally with speaker turns and phrase boundaries, while entertainment timestamps are editorial markers added post-production for navigation (scenes, acts, highlights). These editorial timestamps frequently split continuous dialogue or complex audio scenes at arbitrary points that violate acoustic structure. Silence detection outperforms in entertainment precisely because it respects natural speech boundaries, avoiding the mid-utterance splits that editorial timestamps introduce.

7 Conclusion

This work presents the first systematic evaluation of audio segmentation strategies for Persian YouTube content, revealing that optimal approaches are fundamentally content-type dependent. Our key findings: (1) podcast content favors timestamp segmentation with 33% lower WER, (2) entertainment content paradoxically benefits from silence-based segmentation with 8% lower WER, (3) this effect persists across model scales, confirming it reflects acoustic characteristics rather than capacity limitations, and (4) silence segmentation offers better worst-case guarantees through reduced error variance.

Our publicly released dataset addresses a critical gap in Persian ASR research, providing a systematic benchmark for real-world informal content. This enables researchers to build better tools and democratizes professional-quality transcription for Persian creators.

Future work should explore: (1) content-type classifiers coupled with adaptive segmentation systems, (2) hybrid timestamp-acoustic methods that validate boundaries acoustically, (3) alternative metrics beyond WER/CER, (4) parameter optimization per content type, and (5) extension to other languages and domains.

Limitations

While this study provides the first systematic evaluation of segmentation strategies for Persian ASR, several limitations warrant consideration. First, our analysis is restricted to two content types from Persian YouTube, conversational podcasts and entertainment videos, which may not generalize to other domains such as formal lectures, broadcast news, parliamentary proceedings, or telephone conversations where acoustic characteristics and speaking styles differ substantially. Second, we evaluate only OpenAI's Whisper models at three scales; comparisons with other state-of-the-art ASR systems (e.g., wav2vec 2.0, Conformer-based models, or Persian-specific architectures) would strengthen conclusions about whether content-type dependency is a general phenomenon or specific to Whisper's architecture and training. Third, our silence-based segmentation employs fixed energy thresholds and duration constraints (5–30 seconds) that were not systematically optimized; content-specific parameter tuning might yield different relative performance. Fourth, our evaluation relies solely on WER and CER metrics, which may not fully capture perceptual quality, semantic preservation, and complexity of Persian vocabulary. Finally, our dataset represents standard Persian from digital media creators; our findings may not extend to dialectal variations, code-switched content, or Persian spoken in different geographic regions or sociolinguistic contexts. Future work should address these limitations through multi-domain evaluation, cross-system comparison, and user-centered metrics.

References

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Carlos Arriaga, Alejandro Pozo, Javier Conde, and Alvaro Alonso. 2024. Evaluation of real-time transcriptions using end-to-end asr models. *arXiv preprint arXiv:2409.05674*.

Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a persian written corpus: Peykare. *Language Resources and Evaluation*, 45:143–164.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 250–262.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*.

Arlo Faria, Adam Janin, Sidhi Adkoli, and Korbinian Riedhammer. 2022. Toward Zero Oracle Word Error Rate on the Switchboard Benchmark. In *Interspeech 2022*, pages 3973–3977.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. 2022. Esb: A benchmark for multi-domain end-to-end speech recognition. *arXiv preprint arXiv:2210.13352*.

Masood Ghayoomi and Saeedeh Momtazi. 2009. Challenges in developing persian corpora from online resources. In *2009 International Conference on Asian Language Processing*, pages 108–113. IEEE.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information fusion*, 109.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2024. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4):1–23.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nima Sedghiyeh, Sara Sadeghi, Reza Khodadadi, Farzin Kashani, Omid Aghdaei, Somayeh Rahimi, and Mohammad Sadegh Safari. 2025. Psrb: A comprehensive benchmark for evaluating persian asr systems.

Ilya Sklyar, Anna Piunova, and Christian Osendorfer. 2022. Separator-Transducer-Segmenter: Streaming Recognition and Segmentation of Multi-party Speech. In *Interspeech 2022*, pages 4451–4455.

Piotr Szymański, Piotr Żelasko, Mikołaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. 2020. Wer we are and wer we think we are. *arXiv preprint arXiv:2010.03432*.

Thilo von Neumann, Keisuke Kinoshita, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach. 2023. Segment-less continuous speech separation of meetings: Training and evaluation criteria. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:576–589.

Hossein Zeinali, Lukáš Burget, and Jan Henrik Černockỳ. 2019. A multi purpose and large scale speech corpus in persian and english for speaker and speech recognition: the deepmine database. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 397–402. IEEE.