ParsVoice: A Large-Scale Multi-Speaker Persian Speech Corpus for Text-to-Speech Synthesis

Anonymous ACL submission

Abstract

Persian remains substantially underrepresented in open speech-text resources, limiting progress in multi-speaker text-to-speech (TTS), speech-language modelling, and low-resource speech processing. We introduce **ParsVoice**, the largest publicly available Persian speech-text corpus tailored for training multi-speaker TTS systems, along with a scalable pipeline to construct high-quality speech-text data from long-form audiobook recordings. The pipeline combines a fine-tuned ParsBERT sentence-completion classifier, ASR-based boundary optimization, punctuation restoration, speaker identification, and a multi-dimensional quality assessment that covers both audio and Persian-specific text properties. The resulting release contains a 2,200-hour TTS-ready subset with 1.36 million aligned segments from 1,815 automatically identified speaker IDs, making it more than 25 times larger than the previously largest open Persian TTS dataset. To validate the corpus, we fine-tune XTTS, a zero-shot multilingual TTS model that operates directly on raw Persian text without phoneme representations, achieving a naturalness MOS of 3.6/5 and speaker similarity MOS of 4.0/5. ParsVoice, its metadata, and the construction pipeline are publicly released to support reproducible research on Persian speech synthesis and low-resource speech-language technologies.

1 Introduction

The rapid advancement of transformer architectures (Vaswani et al., 2017) and generative models has increased the demand for large-scale, high-quality speech data. Despite this growing need, resource availability remains highly uneven across languages, and Persian remains significantly underrepresented in speech corpora compared to high-resource languages such as English. This

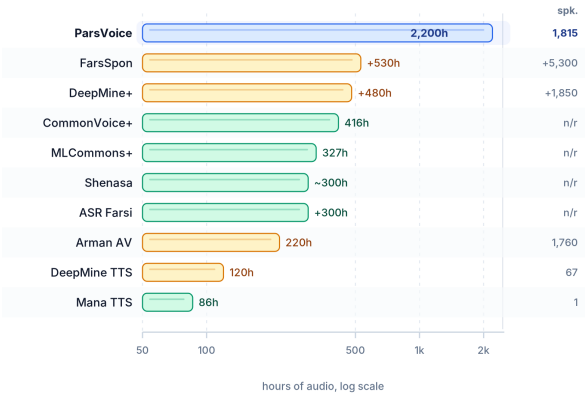

Figure 1: Comparison of Persian speech datasets by duration on a logarithmic scale.

lack of data limits progress not only in text-to-speech (TTS) synthesis, but also in related speech-language tasks such as automatic speech recognition (ASR), speech-text alignment, punctuation restoration, and audio-language modelling.

This scarcity extends beyond speech: insufficient Persian data affects a broad range of language technologies, including ASR, forced alignment, and optical character recognition (OCR). English benefits from thousands of hours of high-quality labelled speech (Panayotov et al., 2015), mature alignment tools with pretrained acoustic models (McAuliffe et al., 2017), and well-established OCR benchmarks—resources that are simply unavailable at a comparable scale for Persian. Moreover, Persian speech presents several language-specific challenges: short vowels are systematically omitted in written text, the *Ezafe*—a connective vowel between consecutive words—is never written yet must be pronounced, and a single letter can map to multiple phonemes depending on context (Adibian et al., 2025). These properties make raw-text Persian TTS challenging and have motivated many prior systems to rely on explicit phonetic or linguistic preprocessing. This resource gap has slowed development across mul-

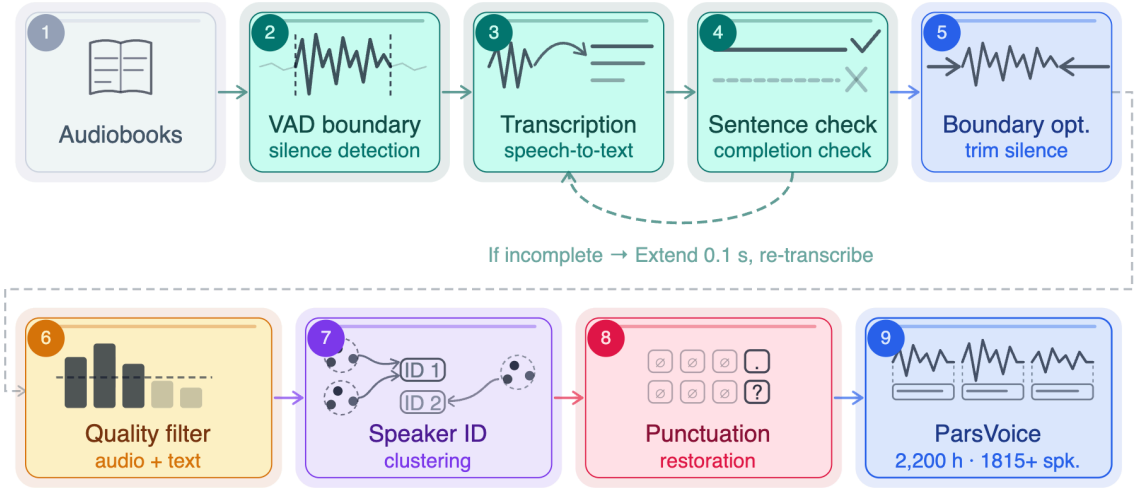

Figure 2: Overview of the ParsVoice corpus construction pipeline. Stages 1–5 handle segmentation and transcription; the dashed arc between stages 3 and 4 shows the iterative extension loop for incomplete sentences. Stages 6–9 handle filtering, speaker identification, punctuation restoration, and final output.

tiple areas of Persian language processing and widened the digital divide between Persian and better-resourced languages.

TTS is particularly demanding in its data requirements. Unlike ASR models, which can tolerate noisy, real-world recordings, TTS models require data that satisfies several strict criteria simultaneously: audio must be free of background noise, music, and artifacts surrounding the speech signal, as background noise directly impairs alignment and degrades synthesised speech quality (Zen et al., 2019); transcripts must consist of complete, well-punctuated sentences, since punctuation governs prosody modelling and its absence measurably reduces naturalness (Zen et al., 2019; Qharabagh et al., 2024); speech and text must be precisely aligned at sentence boundaries rather than arbitrary silence points (Zen et al., 2019); and the corpus should cover a diverse range of speakers and content domains to enable generalisation across voices and topics (Zen et al., 2019; Seki et al., 2025). Meeting all of these requirements at scale is significantly more difficult and expensive than collecting ASR data, and the challenge is compounded for Persian by the linguistic properties discussed above. Existing Persian TTS resources are orders of magnitude smaller than their English counterparts (Panayotov et al., 2015; Ito and Johnson, 2017; Yamagishi et al., 2019), mostly single-speaker, and often proprietary—severely constraining multi-speaker TTS development.

To address this gap, we introduce **ParsVoice**, a large-scale publicly available Persian speech-text corpus designed as high-quality training data for Persian TTS and low-resource speech-language research. ParsVoice targets the harder setting of constructing a multi-speaker corpus from long-form audiobook recordings without relying on publicly available reference transcripts. We develop a scalable automated pipeline that combines sentence-aware segmentation, ASR-based transcription, a ParsBERT sentence-completion classifier, ASR-based boundary optimization, speaker identification, punctuation restoration, and Persian-specific audio/text quality assessment. The resulting TTS-ready release contains 2,200 hours of speech, 1.36 million aligned segments, and 1,815 automatically identified speaker IDs, making it more than 25 times larger than the previously largest open Persian TTS dataset (Qharabagh et al., 2024).

To validate ParsVoice, we fine-tuned XTTS (Casanova et al., 2024), a state-of-the-art zero-shot multilingual TTS model that operates directly on text without requiring phoneme representations—in contrast to traditional Persian TTS systems that rely on explicit phonetic transcription (Adibian et al., 2025). Our model achieved a naturalness Mean Opinion Score (MOS) of 3.6/5 and a Speaker Similarity MOS (SMOS) of 4.0/5, demonstrating that ParsVoice supports high-quality, phoneme-free multi-speaker Persian TTS. An overview of the complete pipeline is illustrated in Figure 2, with

each stage described in detail in Section 3.

2 Related Work

2.1 Speech Corpora for High-Resource Languages

Speech dataset development has been dominated by English resources. LibriSpeech (Panayotov et al., 2015) provides 960 hours of read speech from audiobooks across 2,484 speakers, LJSpeech (Ito and Johnson, 2017) offers 24 hours from a single speaker, and VCTK (Yamagishi et al., 2019) contains approximately 44 hours across 109 speakers. Multilingual efforts such as Common Voice (Ardila et al., 2020), Multilingual LibriSpeech (Pratap et al., 2020), and VoxPopuli (Wang et al., 2021) have expanded coverage to 20+ languages, but remain skewed toward European languages with variable quality across linguistic communities. Many widely spoken non-European languages—including Persian—have received comparatively little attention in these initiatives.

2.2 Persian Speech Datasets

Existing Persian speech resources suffer from critical limitations in scale, speaker diversity, and accessibility. Figure 1 compares ParsVoice with representative Persian speech datasets in terms of duration.

DeepMine+ (Zeinali et al., 2019) is the largest existing Persian speech resource, providing 480+ hours from 1,850+ speakers, but it is commercially restricted. Among TTS-specific datasets, DeepMine Multi-TTS (Adibian et al., 2025) provides 120 hours across 67 speakers under a restricted license with manually verified transcripts and phoneme-level annotations; ArmanTTS (Shamgholi et al., 2023) contains approximately 9 hours from a single speaker; and ManaTTS (Qharabagh et al., 2024) offers 86 hours from a single speaker under an open CC-0 license. Common Voice's Persian portion exists but lacks the audio quality required for TTS training.

2.3 Persian TTS Systems

Existing Persian TTS systems have predominantly relied on explicit phoneme-level intermediate representations, adding pipeline complexity and requiring language-specific lexicons or grapheme-to-phoneme models. ManaTTS (Qharabagh et al., 2024) introduced an end-to-end Persian TTS system built on a Tacotron-style architecture and trained on its accompanying 86-hour single-speaker corpus, with phoneme sequences derived from a rule-based front-end. ArmanTTS (Shamgholi et al., 2023) similarly targets single-speaker synthesis from a 9-hour corpus using a standard sequence-to-sequence TTS pipeline with phoneme inputs. The most closely related prior work is DeepMine Multi-TTS (Adibian et al., 2025), which trains both Tacotron2 and FastSpeech2 models on a 120-hour, 67-speaker corpus with manually verified transcripts and phoneme-level annotations; both models consume phoneme sequences rather than raw Persian text. Our system departs from this tradition by fine-tuning XTTS (Casanova et al., 2024), which operates *directly* on raw Persian text without any phoneme transcription step, lowering the barrier for future research and deployment. A quantitative comparison of MOS scores—with the important caveat that scores across studies are *not directly comparable* due to differing rater pools, test sentences, and evaluation protocols—is deferred to Section 5, where our system is evaluated under controlled conditions.

3 ParsVoice

We introduce a pipeline that transforms raw audiobook recordings into a structured, high-quality speech and text corpus through interconnected stages designed to maximize data quality. The pipeline addresses key challenges in Persian speech data creation: maintaining sentence integrity, ensuring audio-text alignment accuracy, and scaling to thousands of hours of content.

3.1 Data Collection and Source Selection

We selected IranSeda (`book.iranseda.ir`) as our primary data source based on several critical considerations. The platform hosts over 3,800 audiobooks across many categories, ensuring broad lexical and stylistic coverage essential for robust TTS training. Content is produced by professional narrators in controlled recording environments at a 44.1 kHz sampling rate, providing consistent audio quality crucial for neural TTS model training. Importantly, IranSeda audiobooks are publicly accessible, making the platform a practical large-scale source for Persian audiobook data.

Although IranSeda provides audiobook recordings together with metadata, the corresponding

written transcripts are not publicly available. We considered obtaining transcripts from scanned or digitised book editions using OCR, but this approach was unreliable at scale. Audiobook narrations often differ from printed editions due to abridgement, narrator edits, translation differences, or edition-specific wording, and many titles have no accessible digitised text version. Even minor text–audio mismatches would introduce systematic alignment errors and propagate noise throughout the corpus. We therefore adopted an ASR-based transcription strategy that is edition-agnostic, does not require access to source texts, and can be applied automatically at scale.

3.2 Intelligent Audio Segmentation

Raw audiobook files typically span several hours and must be segmented into short, linguistically coherent chunks for TTS training. This is challenging because silence-based Voice Activity Detection (VAD) boundaries do not necessarily coincide with sentence boundaries: a segment may start at a natural pause but still end before the sentence is complete. We therefore use a three-phase segmentation pipeline that combines acoustic boundary detection, ASR transcription, and text-based completeness validation.

Phase 1: Acoustic Boundary Detection. We apply WebRTC (Google) VAD with aggressiveness level 0 to identify silence-based candidate boundaries in each audio file; see Appendix A for a comparison of alternative configurations.

Phase 2: Transcription. Each candidate segment is transcribed using the Google Web Speech backend through the SpeechRecognition library (Zhang, 2017), which achieved the lowest Persian ASR error rates among the tested alternatives (Appendix B). Since the API returns only the recognised text string and does not provide word-level timestamps, timestamp-based alignment is not possible. This motivates the boundary-extension procedure described below. Residual transcription errors are handled downstream by the quality assessment frameworks described in Section 3.4.

Phase 3: Completeness Validation and Boundary Extension. Each transcript is analysed by a ParsBERT-based sentence-completion classifier fine-tuned to distinguish complete from incomplete Persian sentences. Segments predicted as incomplete undergo iterative boundary extension in 0.1-second increments, up to a maximum of 5 seconds. After each extension, the segment is re-transcribed and re-evaluated by the classifier. This loop continues until the transcript satisfies the completeness criterion or the maximum extension limit is reached. Segments that remain incomplete are retained with an incompleteness flag rather than discarded, allowing downstream users to apply their own quality thresholds. Full details of the classifier training procedure, including the construction of complete and truncated training examples from PersianPunc (Kalahroodi et al., 2026), are provided in Appendix C.

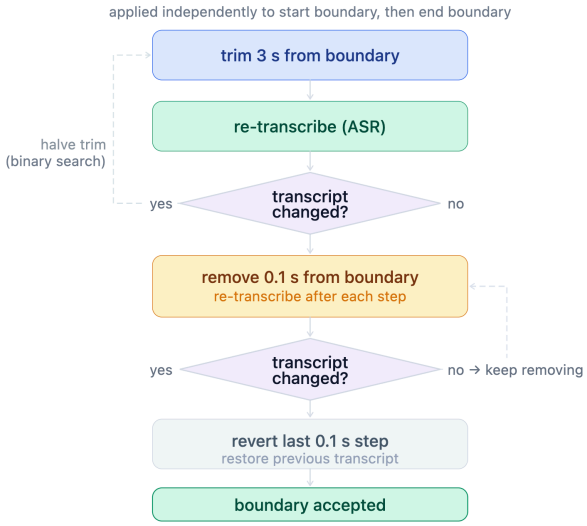

Figure 3: Boundary optimization algorithm, applied independently to the start and end boundaries of each segment. Binary search first converges on a coarse optimal trim; linear fine-tuning then removes 0.1 s increments until the transcript changes, at which point the last step is reverted and The boundary is accepted.

3.3 Boundary Optimization Algorithm

Even with accurate transcription, audio segments may contain unwanted silence, background noise, or acoustic artifacts at boundaries that degrade TTS model performance (Zen et al., 2019; Jung et al., 2024; Seki et al., 2025). Our boundary optimization algorithm employs binary search to determine optimal trimming points for both segment start and end boundaries. Figure 3 illustrates the full algorithm.

The boundary optimization is applied independently to each boundary: the start boundary is processed first, then the end boundary, using the same procedure. A hard ceiling of 3 seconds is set for the initial trimming step to prevent over-trimming in edge cases.

The procedure follows a two-stage search strat-

egy:

1. Initial Adjustment: 3 seconds are removed from the boundary, and the trimmed audio is re-transcribed.

2. Stability Verification: If the new transcription differs from the original in any way—that is, if even a single character changes—the trimming is deemed excessive and must be reduced. Any textual difference is treated as a signal that speech content has been removed.

3. Binary Search Optimization: The algorithm iteratively halves the trimming interval to converge on the maximum trim length that still produces an identical transcription. Binary search terminates when the transcription is character-for-character identical to the original.

4. Fine-Grained Linear Search: After binary search converges, a linear search with 0.1-second increments is applied to achieve precise boundary alignment, continuing until the transcript changes, at which point the last step is reverted, and the boundary is accepted.

In practice, the binary search converges well within the 3-second ceiling for the vast majority of segments. Leading silence is more prevalent than trailing noise in the source recordings; full trimming statistics are reported in Appendix E (Table 9).

3.4 Text-Audio Quality Assessment

High-quality speech is critical for TTS training data, as low-quality samples can degrade model performance. We implement a comprehensive assessment across both audio and text dimensions. All thresholds were determined through human review of a random subset of segments: candidate threshold values were applied, and the resulting accepted and rejected samples were manually inspected until the authors judged the boundary to produce consistently clean output. The distribution of quality scores across the full dataset is provided in Appendix E.

3.4.1 Persian Text Quality Metrics

The Persian Text Quality Framework evaluates transcriptions across five dimensions: character quality, length quality, repetition score, linguistic complexity, and phonetic coverage. Each metric is normalised to $[0, 1]$ and combined using fixed weights to produce a total score on $[0, 1]$. Segments are categorised into three descriptive tiers: high quality ($\geq$ 0.78), mid-quality (0.62–0.78), and low quality ($<$ 0.62). These tier boundaries are used for reporting and analysis only; the threshold applied to exclude segments from the TTS-ready subset is 0.5 (see Section 3.6). Full metric descriptions, weights, and tier thresholds are provided in Appendix D.

3.4.2 Audio Quality Metrics

The audio quality framework assesses recordings for overall clarity, absence of distortions, and suitability for speech processing. Metrics include estimated signal-to-noise ratio, dynamic range, clipping ratio, silence ratio, background music presence (detected via inaSpeechSegmenter (Doukhan et al., 2018)), and duration. Each segment receives a composite score on a 0–100 scale; recordings are classified into three tiers: high quality ($\geq$ 90), acceptable (75–89), and low quality ($<$ 75, excluded from TTS use). Full scoring rules, metric descriptions, and score distributions are provided in Appendices D and E.

3.5 Speaker Identification

Since approximately 40% of audiobook entries lacked narrator names and some books featured multiple narrators, we apply a two-stage speaker identification pipeline based on ECAPA-TDNN embeddings (Desplanques et al., 2020) to assign consistent speaker labels across the corpus. The first stage clusters segments within each audiobook into local speaker identities; the second merges these local identities into global speaker labels across the entire corpus. The number of speakers per book is estimated by consensus among Silhouette, gap statistic, and DBSCAN-based criteria, and clustering uses an ensemble of Agglomerative, Spectral, and Gaussian Mixture Model methods, with the result selected by highest silhouette score. Each segment receives a confidence score combining its silhouette contribution and distance to the assigned centroid; low-confidence segments and spuriously small clusters are excluded. Global identities are formed by agglomerative clustering of confidence-weighted centroid embeddings across all books, with a cosine similarity threshold of 0.85 and a secondary validation step to prevent spurious cross-book merges. Full implementation details, including outlier removal, dimensionality reduction, confidence thresholds, and merging criteria, are provided in Appendix F.

Table 1: ParsVoice Dataset Statistics. Token and vocabulary counts are computed using the hazm Persian tokeniser (Roshan, 2014) after text normalisation. We refer to the boundary-optimised corpus before TTS-specific filtering as the processed ASR-oriented subset.

Metric	ASR Subset	Final TTS Subset
Total Hours	4,096.2	2,199.7
Segments	2,999,296	1,364,671
Total Tokens	29,566,314	17,347,614
Unique Word Forms	309,406	229,860

3.6 Final Data Cleaning and Preparation

We apply audio and text quality filtering to the boundary-optimised corpus: segments with audio quality scores below 75 (out of 100) and segments with text quality scores below 0.5 are removed, reducing the corpus to 2,089,749 segments (3,007.8 hours). The final TTS-ready subset is then constructed by additionally excluding segments that contain background music, segments assigned to low-confidence speaker clusters, and segments flagged as incomplete by the sentence-completion classifier. This yields 1,364,671 segments totalling 2,199.7 hours. Since ASR transcriptions lack punctuation, we apply the ParsBERT-based Persian punctuation restoration model of (Kalahroodi et al., 2026) to all segments prior to release.

3.7 Dataset Release

ParsVoice is released publicly with restored transcripts, speaker IDs, audio-quality scores, and text-quality scores. This allows users to reconstruct different subsets according to their application requirements, including stricter TTS-only subsets or broader ASR-oriented subsets.

4 ParsVoice Corpus Analysis

The current release contains 2,231 processed audiobooks, of which 1,706 are retained in the final TTS subset after quality filtering.

Table 2: ParsVoice pipeline statistics at each processing stage.

Stage	Segments	Hours
VAD output	5,161,459	5,824.7
After removing empty transcripts	3,323,679	5,127.3
After boundary optimization	2,999,296	4,096.2
After text-audio quality filtering	2,089,749	3,007.8
Final TTS subset	1,364,671	2,199.7

Quality Assessment. Of the 3,323,679 non-empty transcript segments, 2,999,296 were retained after boundary optimization; the remaining 324,383 segments were discarded due to transcription instability or failed re-transcription during the boundary search, accounting for approximately 524 hours of audio. For the 2,999,296 retained segments (4,603.2 hours of pre-trim audio), the boundary optimization algorithm removed a mean of 426.9 ms from segment starts and 181.6 ms from segment ends, totalling 507.0 hours (11.0% of the pre-trim duration of retained segments) of unwanted silence and artifacts. Overall, 81.2% of segments required start-boundary trimming, while 50.4% required end-boundary trimming. After trimming, the average segment duration is 4.92 seconds, with a median duration of 3.60 seconds.

Linguistic Analysis. Using the hazm Persian tokeniser (Roshan, 2014) after text normalisation, the ASR subset contains 29,566,314 tokens and 309,406 unique word forms. The final TTS subset contains 17,347,614 tokens and 229,860 unique word forms, with an average of 12.8 words and 60.7 characters per segment.

Speaker Statistics. The final TTS subset contains 1,815 automatically identified global speaker IDs, covering both metadata-linked named narrators and audiobooks with unknown narrator names. Gender analysis was performed using narrator names present in the available metadata; approximately 40% of audiobooks lacked narrator name information and were therefore excluded from the gender count. Among audiobooks with available metadata, approximately 33% of narrators are female, and 67% are male. As a metadata-based clustering consistency check, we computed narrator purity: the duration-weighted share of each named narrator's audio assigned to its dominant global speaker ID. Among 325 narrators with sufficient retained audio, the median purity is 1.00, the mean is 0.869, and 214 narrators (65.8%) have at least 90% of their duration assigned to a single speaker ID.

5 Evaluation: TTS Model Training

To assess whether ParsVoice is suitable for downstream TTS training, we fine-tuned XTTSv2 (Casanova et al., 2024), a multilingual zero-shot TTS model, on the proposed corpus. The goal of this experiment is not to introduce a new TTS architecture, but to validate that ParsVoice can support high-quality Persian speech synthesis and speaker adaptation.

Table 3: Subjective evaluation results. MOS = naturalness, SMOS = speaker similarity, Acc. = intelligibility MOS, and Spk.Sim. is ECAPA-TDNN cosine similarity reported as a percentage. † Results are taken directly from the original papers and were not re-evaluated by our raters. For XTTS + ParsVoice, 95% confidence intervals are shown.

System	MOS	SMOS	Acc.	Spk.Sim. (%)
Real Persian Speech†	4.72	4.86	—	—
XTTS + ParsVoice	**3.59±0.09**	**4.03±0.08**	**4.03±0.08**	**80.0**
DeepMine Tacotron2† (Adibian et al., 2025)	3.94±0.04	—	—	—
DeepMine FastSpeech2† (Adibian et al., 2025)	4.12±0.06	3.98±0.11	—	—

5.1 Model Adaptation

We adapted the GPT component of XTTSv2 to Persian while keeping the remaining modules frozen, including the DVAE. A BPE tokeniser was trained on Persian text, and 2,500 additional Persian tokens extracted from the Copera corpus were added to the XTTSv2 GPT vocabulary. The model was trained on a single NVIDIA A100 GPU using AdamW with a learning rate of 5×10^{-6} and weight decay of 10^{-2} for 170000 steps. We used a per-device batch size of 21 with gradient accumulation over 8 steps, resulting in an effective batch size of 168.

5.2 Evaluation Setup

We evaluated the fine-tuned model using 90 synthesised samples generated from random sentences and unseen Persian reference speakers drawn from the Persian Common Voice dataset. These speakers are drawn from Persian Common Voice rather than ParsVoice and are therefore external to the training corpus. The age and gender distribution of the reference speakers is reported in Appendix G.3. Subjective evaluation was conducted using ratings from 22 retained native Persian raters, selected from 41 recruited raters. Raters scored naturalness using MOS, speaker similarity using SMOS, and intelligibility using an additional 1–5 intelligibility MOS. Further details about the rater pool and annotation procedure are provided in Appendix G.1. For objective evaluation, we measured intelligibility using WER and CER from three ASR systems: Google Web Speech, the transcription system used during corpus construction; a Persian-finetuned Whisper model (Nezamisafa, 2025), which is independent of our pipeline; and Qwen ASR 1.7B (Shi et al., 2026). To contextualise these scores, we also evaluated the same ASR systems on the Google FLEURS Persian test set (Conneau et al., 2023), which serves as a real-speech reference. Speaker similarity was additionally measured using ECAPA-TDNN cosine similarity between reference and synthesised speaker embeddings.

5.3 Subjective Evaluation Results

Table 3 reports the subjective evaluation results. The fine-tuned XTTS model achieves a naturalness MOS of 3.60, a speaker similarity SMOS of 4.00, and an intelligibility MOS of 4.00. These results indicate that ParsVoice supports intelligible and speaker-consistent Persian synthesis in a zero-shot adaptation setting. The real-speech MOS and SMOS values are taken from (Adibian et al., 2025) and are included only as approximate upper-bound references. Likewise, the DeepMine Tacotron2 and FastSpeech2 results are copied from the original paper and were not evaluated by our rater pool. Since these systems were evaluated with different raters, test sentences, and experimental conditions, direct numerical comparison should be interpreted with caution.

5.4 Objective Results

Table 4 shows WER and CER across the three ASR systems. Because Google Web Speech was also used during corpus construction, we treat it only as a diagnostic evaluator. The Persian-finetuned Whisper model, which is independent of the pipeline, provides the primary non-circular intelligibility check. Google ASR obtains 10.40% WER and 5.92% CER on synthesised speech, compared with 6.53% WER and 2.55% CER on real Persian speech from FLEURS. This corresponds to a WER gap of 3.87 percentage points. Importantly, the Persian-finetuned Whisper model, which is independent of our transcription pipeline, shows a similar trend: 17.20% WER on synthesised speech versus 13.07% on FLEURS, corresponding to a 4.13 percentage-point gap. This suggests that the low WER is not merely an arte-

Table 4: Multi-ASR intelligibility evaluation. WER and CER are reported on 90 synthesised samples and on the Google FLEURS Persian test set as a real-speech reference. Lower WER/CER is better; the gap column shows TTS−Real degradation in WER.

	TTS		Real		Gap
ASR System	**WER**	**CER**	**WER**	**CER**	**WER**
Google Web Speech	10.40	5.92	6.53	2.55	+3.87
Whisper fine-tuned (Nezamisafa, 2025)	17.20	7.79	13.07	5.05	+4.13
Qwen ASR 1.7B	34.08	15.32	22.92	8.56	+11.16

All values are percentages. The Persian-finetuned Whisper model is independent of our transcription pipeline and serves as the primary unbiased intelligibility reference.

fact of using Google Web Speech during corpus construction. Instead, the synthesised speech remains broadly intelligible under an independent ASR evaluator. Qwen ASR 1.7B reports higher absolute error rates on both real and synthesised speech, but preserves the same overall pattern. Finally, speaker similarity was assessed using ECAPA-TDNN (Desplanques et al., 2020) embeddings extracted via the SpeechBrain toolkit (Parcollet et al., 2022; Ravanelli et al., 2024), computing cosine similarity between the reference and synthesised speaker representations. The resulting score of 80.0% indicates effective zero-shot speaker adaptation across the 42 unseen reference speakers.

6 Conclusion

In this work, we address the scarcity of high-quality Persian speech datasets by introducing ParsVoice, the largest publicly available Persian speech-text corpus tailored for TTS training to date. ParsVoice consists of 2,200 hours of clean, segmented speech suitable for TTS training, covering 1,364,671 segments from 1,815 automatically identified speaker IDs. In addition, the broader processed corpus contains 4,096.2 hours of audio, making it useful for a wider range of Persian speech research tasks.

Alongside the dataset, we provide a scalable and automated pipeline for dataset creation. This pipeline incorporates several key parts, including a BERT-based model to ensure sentence completeness, a binary search algorithm for precise audio-text boundary optimization, and comprehensive audio-text quality assessment frameworks specifically designed for Persian. Together, the dataset and the pipeline provide a valuable resource for advancing Persian speech research and TTS development.

We validated ParsVoice by fine-tuning XTTS, achieving a naturalness MOS of 3.6/5 and a speaker similarity SMOS of 4.0/5 (Section 5). Multi-ASR intelligibility evaluation using both the transcription-pipeline ASR and a fully independent Persian-finetuned Whisper model confirms that the synthesised speech is genuinely intelligible. These results confirm the corpus's quality and its suitability for developing robust multi-speaker TTS systems, addressing a critical resource gap in Persian language technology.

7 Limitations

Despite the scale and quality of ParsVoice, several limitations should be acknowledged. First, the corpus is derived entirely from audiobooks, which means the speaking style is predominantly formal and narrative. Spontaneous conversational speech, which differs substantially in prosody, pace, and vocabulary, is not represented. TTS models trained on ParsVoice may therefore produce speech that sounds natural in read-aloud contexts but less so in conversational applications. Second, our transcription pipeline relies on a commercial ASR system without access to the original book texts, which introduces a small but non-zero transcription error rate. While our multi-stage quality filtering mitigates this, some residual errors may remain in the dataset. Third, the evaluation of ParsVoice was conducted by fine-tuning XTTS, a single model architecture. Due to limited access to GPU compute, we were unable to train multiple model architectures or conduct extensive hyperparameter searches, which limits the generalisability of the validation results. Similarly, we were unable to evaluate against restricted-access baseline models such as DeepMine Multi-TTS, meaning our MOS comparison relies on scores reported in the original papers rather than a controlled head-to-head evaluation with shared raters. Fourth, the metadata-linked portion of the corpus is gender-imbalanced, with approximately 67% male and 33% female narrators among audiobooks with available narrator names. Because around 40% of audiobooks lack narrator metadata, the full-corpus gender distribution remains only partially observed. This imbalance may affect

the quality of TTS voices synthesised for female speakers relative to male speakers. Finally, a substantial portion of the corpus lacks complete narrator metadata, meaning speaker identity for these recordings is assigned through automated clustering rather than verified ground truth. Although the ECAPA-TDNN-based pipeline provides useful global speaker IDs, these labels should be interpreted as automatically inferred identities rather than manually verified speaker annotations.

References

Majid Adibian, Hossein Zeinali, and Soroush Barmaki. 2025. Deepmine-multi-tts: a persian speech corpus for multi-speaker text-to-speech. *Language Resources and Evaluation*, 59:2245–2264.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982.

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.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834.

David Doukhan, Eliott Lechapt, Marc Evrard, and Jean Carrive. 2018. Inas mirex 2018 music and speech detection system. In *Music Information Retrieval Evaluation eXchange (MIREX 2018)*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):38313847.

Google. Webrtc voice activity detector. https://github.com/wiseman/py-webrtcvad. Accessed: 2025-09-14.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Jee-weon Jung, Wangyou Zhang, Soumi Maiti, Yihan Wu, Xin Wang, Ji-Hoon Kim, Yuta Matsunaga, Seyun Um, Jinchuan Tian, Hye-jin Shim, and 1 others. 2024. Text-to-speech synthesis in the wild. *arXiv preprint arXiv:2409.08711*.

Mohammad Javad Ranjbar Kalahroodi, Heshaam Faili, and Azadeh Shakery. 2026. PersianPunc: A large-scale dataset and BERT-based approach for Persian punctuation restoration. In *The Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 105–113, Rabat, Morocco. Association for Computational Linguistics.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*, pages 498–502.

Nezamisafa. 2025. whisper-persian-v4. https://huggingface.co/nezamisafa/whisper-persian-v4. Fine-tuned version of OpenAI Whisper-large-v3 for Persian automatic speech recognition. Accessed: 2026-05-25.

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.

Titouan Parcollet, Mirco Ravanelli, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2022. SpeechBrain: A General-Purpose Speech Toolkit. Preprint.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, interspeech_2020, page 27572761. ISCA.

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R. Rabiee. 2024. Manatts persian: a recipe for creating tts datasets for lower resource languages. *Preprint*, arXiv:2409.07259.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain {de Langen}, Cem Subakan, Peter Plantinga,

Yingzhi Wang, Pooneh Mousavi, {Luca Della} Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, {Sung Lin} Yeh, Pierre Champion, Aku Rouhe, and 14 others. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25. Publisher Copyright: © 2024 Mirco Ravanelli, Titouan Parcollet, et al.

Roshan. 2014. hazm: Python library for digesting persian text. Available at https://github.com/roshan-research/hazm.

Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari. 2025. Ttsops: A closed-loop corpus optimization framework for training multi-speaker tts models from dark data. *Preprint*, arXiv:2506.15614.

Mohammd Hasan Shamgholi, Vahid Saeedi, Javad Peymanfard, Leila Alhabib, and Hossein Zeinali. 2023. Armantts single-speaker persian dataset. *Preprint*, arXiv:2304.03585.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

SileroTeam. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

Hossein Zeinali, Luká Burget, and Jan "Honza" ernocký. 2019. A multi purpose and large scale speech corpus in persian and english for speaker and speech recognition: the deepmine database. *Preprint*, arXiv:1912.03627.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech 2019*, pages 1526–1530.

Anthony Zhang. 2017. Speech recognition (version 3.11). Software. Available from https://github.com/Uberi/speech_recognition. Accessed: 2025.

A VAD Method Comparison

To justify the choice of the VAD algorithm used in Phase 1 of the segmentation pipeline (Section 3.2), we evaluated six VAD configurations on a random sample of ten audiobooks. The evaluation metric is *completion rate*: the percentage of produced segments that the ParsBERT sentence-completion classifier identifies as complete sentences. WebRTC Level 0 achieved the highest completion rate while remaining lightweight and scalable; Silero and Whisper-based VAD are additionally heavier to run. We therefore selected WebRTC Level 0 as the segmentation backend.

VAD Method	Completion Rate (%)
WebRTC Level 0	**75.31**
WebRTC Level 1	73.43
WebRTC Level 2	59.62
WebRTC Level 3	31.60
Silero (SileroTeam, 2024)	42.02
Whisper-based VAD	52.10

Table 5: Completion rates of segments produced by different VAD methods on ten randomly selected audiobooks.

B ASR System Comparison

We compared three ASR systems on 1,000 randomly selected samples from Persian Common Voice (Ardila et al., 2020) to select a transcription backend.

ASR System	WER	CER
Google Speech Rec.	**20.10**	**7.03**
Whisper-large-v3 Persian-finetuned (Nezamisafa, 2025)	32.50	11.14
Whisper-large-v3 (Radford et al., 2023)	43.77	15.62

Table 6: Persian ASR comparison on 1,000 Persian Common Voice samples. Lower is better.

Google Speech Recognition achieves the lowest WER (20.10%) and CER (7.03%) and was therefore selected as the transcription backend.

C Sentence Completion Classifier

The classifier is fine-tuned from ParsBERT (Farahani et al., 2021) on synthetic data from Persian-Punc (Kalahroodi et al., 2026). Positive examples are complete sentences; negatives are generated by removing up to five words from the sentence end or truncating the final word by up to five characters, reflecting the tendency of VAD boundaries to cut sentences at their endings. Incomplete examples are oversampled at a 3:2 ratio. Training runs for 3 epochs with Adam (lr $= 2 \times 10^{-5}$, batch size 128, weight decay 0.01). On a 10,000-sample held-out set the classifier achieves 97.4% F1 across both classes.

D Audio and Text Quality Scoring Details

D.1 Audio Quality Scoring

Each recording receives a composite score on a 0–100 scale. Scores of ≥ 90 are *high quality*; 75–89 are *acceptable*; below 75 are *low quality* and excluded from TTS use. Scoring rules are listed in Table 7.

Table 7: Audio quality scoring rules. Adjustments are summed and clamped to $[0, 100]$.

Metric	Condition	Adj.
SNR	> 20 dB	$+35$
	10–20 dB	$+25$
	5–10 dB	$+15$
Dynamic range	> 20 dB	$+15$
	15–20 dB	$+12$
	10–15 dB	$+10$
Duration	3–15 s	$+10$
	15–30 s	$+5$
No background music	clean (binary)	$+15$
Clipping	$> 5\%$	-20
	1–5%	-10
	0.1–1%	-5
Silence	$> 50\%$	-15
	30–50%	-10
	20–30%	-5

D.2 Text Quality Scoring

Each transcription is scored on a 0–1 scale across five weighted dimensions. (Table 8). Tier thresholds for descriptive reporting: ≥ 0.78 = *high quality*; 0.62–0.78 = *mid quality*; < 0.62 = *low quality*. Note that these tiers are used only for analysis and reporting; the actual filtering threshold that excludes segments from the TTS-ready corpus is 0.5 (Section 3.6).

E Corpus Statistics

E.1 Pipeline Stage Statistics

Segment counts and hours at each processing stage are reported in Table 2 in Section 4. After transcription and before boundary optimization, segments have a mean duration of 5.55 s (me-

Table 8: Text quality dimensions, weights, and descriptions.

Dimension	Weight	Description
Character quality	0.25	Fraction of valid Persian characters or digits; penalises Arabic-script substitutes.
Length quality	0.20	Word and character counts in an acceptable range (3–30 words, 20–200 characters).
Repetition score	0.20	Lexical diversity; penalises excessive word repetition.
Linguistic complexity	0.175	Combines common-word ratio, average word length.
Phonetic coverage	0.175	Diversity of Persian characters; favours broad phoneme coverage.

Table 9: Boundary optimization trimming statistics across 2,999,296 retained segments.

Metric	Start	End
Segments trimmed	2,434,154 (81.2%)	1,510,777 (50.4%)
Mean trim, all segs (ms)	426.9	181.6
Mean trim, trimmed (ms)	526.0	360.5
Total audio removed by trimming: 507.0 h (11.0% of pre-trim duration of retained segments, 4,603.2 h)		

dian 4.20 s). Only 1.1% of segments are shorter than 1 s and 1.9% exceed 20 s, confirming that VAD-based segmentation produces broadly TTS-suitable lengths even before further refinement. Of the 3,323,679 segments with non-empty transcripts, 86.3% were classified as complete by the ParsBERT classifier; the remaining 13.7% underwent iterative boundary extension with a mean of 17.7 extension steps.

E.2 Boundary optimization

The boundary optimization algorithm processed 3,323,679 non-empty transcript segments. Of these, 324,383 segments were discarded due to transcription instability or failed re-transcription during the boundary search, accounting for approximately 524 hours of audio. The remaining 2,999,296 segments were retained and trimmed. Table 9 summarises trimming statistics for the retained segments.

E.3 Audio Quality

The mean composite audio quality score is 89.5 (median 92.0). Of all scored segments, 83.1% reach the high-quality threshold ($\geq$ 90), 13.4% fall in the acceptable range (75–89), and 3.5% are excluded. Mean SNR is 41.8 dB (96.5% of segments exceed 20 dB). No clipping was detected. Background music was present in 4.3% of segments; these are excluded from the TTS-ready subset.

E.4 Text Quality

In the final TTS subset, the mean text-quality score is 0.737 (median 0.748). Sub-metric means are: character quality 0.851, length quality 0.900, repetition score 0.597, linguistic complexity 0.586, and phonetic coverage 0.699.

F Speaker Identification Details

Local clustering. Segment-level ECAPA-TDNN embeddings are preprocessed by removing statistical outliers (z-score threshold 3.0), L2-normalised, and PCA-reduced when the embedding dimension exceeds 512. The number of speakers k^* per book is estimated by consensus among the Silhouette score, the gap statistic, and DBSCAN-based estimates. Clustering uses an ensemble of Agglomerative (cosine distance, average linkage), Spectral, and Gaussian Mixture Model methods, with the result selected by the highest silhouette score. Each segment receives a confidence score $c_i = 0.6\,\tilde{s}_i + 0.4\,d_i$, where $\tilde{s}_i$ is the normalised silhouette score and d_i is the distance-based confidence. Segments with $c_i < 0.4$ and clusters smaller than $1/k^*$ of the high-confidence pool are excluded; speaker groups with fewer than 10% of the book's high-confidence segments are discarded entirely, preventing fragmented utterances from entering the final subset.

Global merging. Local cluster centroids are compared across all books using pairwise cosine similarity. Agglomerative clustering with average linkage and a similarity threshold of 0.85 forms global identities. Cross-book merges are accepted only when the matched speakers share at least 100 segments in total; otherwise they are kept as separate single-book identities.

G Evaluation Details

G.1 Annotation Interface and Procedure

Ratings were collected via *Goyar*, a custom Telegram-based annotation bot developed for this study. Raters accessed the interface through their personal Telegram accounts at their own convenience. Before the session began, each rater received the following standardised instructions: complete the evaluation in a quiet environment,

and use in-ear headphones (earphones) throughout, to ensure consistent and distraction-free listening conditions.

Each session comprised 90 evaluation items presented sequentially. For each item, the bot displayed (i) the item index out of 90, (ii) the Persian reference text, (iii) a reference audio clip recorded by the target speaker, and (iv) the corresponding synthesised audio clip. Raters were explicitly instructed to listen to the *reference audio* first in order to familiarise themselves with the target speaker's voice characteristics before listening to the *synthesised audio*. Ratings were submitted directly via the bot's inline response buttons.

Each synthesised sample was scored on three criteria using the following scales:

1. Naturalness (MOS). How natural and human-like does the synthesised voice sound? Raters were asked to consider whether the voice sounded robotic, whether prosody and intonation were unusual, and whether pauses and pronunciation were natural.

- 1 = Completely unnatural, machine-like
- 2 = Somewhat unnatural with clear defects
- 3 = Acceptable but not fully natural
- 4 = Almost natural with minor imperfections
- 5 = Completely natural and human-like

2. Speaker Similarity (SMOS). How closely does the synthesised voice resemble the reference speaker? Raters were asked to judge whether the quality, pitch, and vocal characteristics of the synthesised voice matched those of the reference recording.

- 1 = Completely different, like another person
- 2 = Somewhat different, low similarity
- 3 = Moderate similarity
- 4 = Very similar with minor differences
- 5 = Identical to the reference speaker

3. Intelligibility (Intel. MOS). How well does the synthesised audio match the written text? Raters were asked to note whether any words were missing, added, or mispronounced. This scale used 0.5-point increments to permit finer-grained assessment of intelligibility:

- 1 = Completely mismatched with the text
- 2 = Some correspondence but many errors
- 3 = Acceptable correspondence
- 4 = Good correspondence with minor errors
- 5 = Perfectly and exactly matched

Figure 4 shows a screenshot of the Goyar interface as displayed to raters.

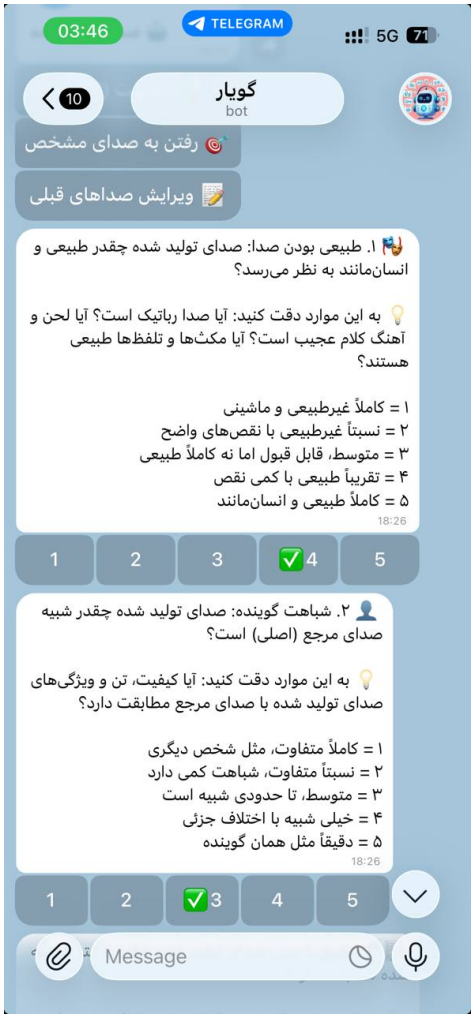

Figure 4: Screenshot of the Goyar Telegram-based annotation interface. The bot presents the reference text, delivers the reference and synthesised audio clips in sequence, and collects naturalness, speaker similarity, and intelligibility ratings via inline response buttons.

G.2 Raters

Subjective ratings were collected via a Telegram-based interface. Of 41 recruited native Persian raters, 22 who completed at least 10 ratings were retained. Table 10 summarises rater demographics, and Table 11 reports mean scores grouped by rater gender.

Table 10: Retained rater demographics (22 raters).

Property	Value
Female / Male	11 / 11
Mean age (years)	28.6
Age range	19–60

Table 11: Mean scores by rater gender.

Rater gender	MOS	SMOS	Intel. MOS
Female	3.46	3.82	3.84
Male	3.76	4.30	4.26

G.3 Reference Speakers

The 90 synthesised samples correspond to 42 unique reference speakers stratified from Persian Common Voice (Ardila et al., 2020), unseen during training. Speakers were stratified by gender and age group to ensure demographic coverage. Table 12 summarises their demographics. Female reference speakers received marginally higher scores across all three dimensions (MOS 3.69 vs. 3.53, SMOS 4.13 vs. 3.99, intelligibility MOS 4.14 vs. 3.94).

Table 12: Reference speaker demographics (42 speakers).

Property	Count	%
Gender		
Female	17	40.5
Male	20	47.6
Not reported	5	11.9
Age group		
Teens	6	14.3
Twenties	10	23.8
Thirties	10	23.8
Forties	7	16.7
Fifties	1	2.4
Sixties	1	2.4
Not reported	7	16.7