ManaTTS Persian: a recipe for creating TTS datasets for lower resource languages

Mahta Fetrat Qharabagh **Zahra Dehghanian** **Hamid R. Rabiee**
Sharif University of Technology / Tehran
m.fetrat@sharif.edu zahra.dehghanian97@sharif.edu rabiee@sharif.edu

Abstract

In this study, we introduce ManaTTS, the most extensive publicly accessible single-speaker Persian corpus, and a comprehensive framework for collecting transcribed speech datasets for the Persian language. ManaTTS, released under the open CC-0 license, comprises approximately 86 hours of audio with a sampling rate of 44.1 kHz. The dataset is supported by a fully transparent, MIT-licensed pipeline, a testament to innovation in the field. It includes unique tools for sentence tokenization, bounded audio segmentation, and a novel forced alignment method. This alignment technique is specifically designed for low-resource languages, addressing a crucial need in the field. With this dataset, we trained a Tacotron2-based TTS model, achieving a Mean Opinion Score (MOS) of 3.76, which is remarkably close to the MOS of 3.86 for the utterances generated by the same vocoder and natural spectrogram, and the MOS of 4.01 for the natural waveform, demonstrating the exceptional quality and effectiveness of the corpus.

1 Introduction

Text-to-speech conversion has long been an essential task. It is integrated with everyday life, including navigation systems, e-learning, content providing, and much more (Maps; Speechify; MurfAI). But one of the most vital applications of text-to-speech systems is providing accessibility for people with visual impairments, enabling written materials such as electronic device screens to be converted to speech that can be heard rather than read (NV).

The reason for emphasizing the latter application is the lack of open-access, high-quality systems. Some Persian text-to-speech models are embedded into applications like the Balad map (Balad) many commercial tools like Narakeet (Narakeet). However, no high-quality, freely available TTS models can be used by the more limited audience, including the visually impaired and speech domain researchers. To address these challenges, it is crucial to develop open-access text-to-speech tools, which primarily require a proper text-to-speech dataset.

An ideal text-to-speech dataset must meet several criteria (Naderi et al., 2022; Zen et al., 2019). First, it must exhibit minimal to no mismatches in transcripts. Second, it should have no background sound, including noise or background music. Third, it should have a high sampling rate (at least 24 kHz) to be useful for modern TTS models. It is also beneficial if the transcripts include exact punctuation to help detect stops and intonations. Additionally, the dataset should be large in terms of both total time duration and word coverage. Therefore, the data source must be diverse and not limited to a specific domain.

Our investigations show that many existing text-to-speech datasets for the Persian language are not publicly available. On the other hand, there are serious challenges with the available data, the most important being non-open licenses, along with issues such as small size, low quality, and limited domain, which will be discussed in the related works review. Hence, the first step toward open-source and open-access text-to-speech models for the Persian language, and the main focus of the current study, is to prepare such a clean, large-scale and open-source dataset.

In this work, we introduce a new dataset called "ManaTTS." The word Mana means "Enduring" and is derived from the name of a monthly magazine devoted to the blind community, called Nasl-e-Mana (NasleMana), which has been the source of our dataset. The magazine is publicly available, and the content providers were receptive to publishing the dataset with an open CC-0 license.[1] The ManaTTS corpus has the following characteristics:

- **Sampling rate:** All the audio files have a

[1] The dataset is available at https://huggingface.co/datasets/MahtaFetrat/Mana-TTS

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 9177–9206
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

sampling rate of 44.1 kHz.

- **Speakers:** The entire dataset is recorded by a single female speaker.
- **Duration:** It includes 86 hours and 24 minutes of processed and transcribed audio and is the largest single-speaker dataset in Persian.
- **License:** It is distributed under the open CC-0 1.0 license, enabling educational and commercial use.
- **Environment:** The data is mostly recorded in a silent environment and processed to remove potential background music.
- **Processing Method:** The entire processing pipeline of the dataset is available, making it fully reproducible. This pipeline introduces useful open-source tools for speech dataset creation, including a new sentence tokenization and forced alignment tool.
- **Extendable:** The dataset can be easily extended thanks to the monthly growing Nasl-e-Mana magazine and the fully open pipeline.
- **Coverage:** The dataset includes 24113 unique words and encompasses various topic domains.
- **Evaluation:** The dataset is used to train a TTS model and has demonstrated effectiveness and high-quality outputs.

We have also collected and processed Persian-Informal, a smaller dataset comprising informal Persian text and speech. This dataset is suitable for evaluating ASR models based on Character Error Rate (CER) and is used to prioritize the ASR models in the alignment tool for this work. For more details, refer to Appendix B.

The rest of the paper is organized as follows. The next section provides a comprehensive review of the available Persian speech datasets. Section 3 includes a detailed explanation of the data collection and processing methods. Section 4 describes the statistics of the dataset. Section 5 presents the experimental results. The final sections 6 and 7 summarize the achievements and limitations of this study.

2 Related Works

Our analysis encompasses various Persian datasets, including text-to-speech (TTS) datasets (Table 1) and other collections featuring speech-text pairs (Table 3). These include automatic speech recognition (ASR) datasets, audio-visual speech recognition (AVSR) datasets specific to Persian, a dataset for Persian phoneme recognition (PR), a Persian spoken digit recognition (DR) dataset, as well as multilingual datasets that incorporate Persian language components. While the primary focus of this study and the discussions in this section is on Persian speech corpora for text-to-speech systems, we have also included several well-known English TTS speech datasets for a more comprehensive comparison (Table 4). To see an extended discussion of the related works, please refer to Appendix A

3 Dataset Preparation

As mentioned earlier, the raw material of our dataset is crawled[2] from the website of the Nasl-e-Mana magazine (NasleMana) and is published under the CC-0 1.0 license with the consent of its owners. The majority of the audio files were recorded by a female speaker, and we manually removed any files not associated with her to ensure the dataset remained single-speaker. This data was processed through a pipeline to obtain speech and transcripts as output pairs. An overview of the entire pipeline is provided in Figure 1a.

The hardware utilized for the entire processing pipeline and model training consisted of a 12th Gen Intel Core i9-12900K CPU with 24 cores and an NVIDIA GeForce RTX 4090 GPU with 24,564 MiB of memory, supporting CUDA version 12.2.

3.1 Preprocessing

Before processing speech-text pairs, we preprocess audio and text files separately. Initially in MP3 format, the audio files are converted to WAV files. WAV format offers lossless compression, preserving audio quality throughout processing. Additionally, the audio undergoes processing with a source separation tool, namely Spleeter (Hennequin et al., 2020), to eliminate any potential background music and retain only the vocals.

[2] The crawling script and the entire processing source code are available at https://github.com/MahtaFetrat/ManaTTS-Persian-Speech-Dataset

Table 1: List of Persian text-to-speech corpora. **Size** refers to the total duration in hours. **N.T.** (Natural Text) and **N.A.** (Natural Audio) indicate whether the text or audio is natural or synthesized, with a × used to denote synthesized content.

Dataset	Size	Speakers	N.T.	N.A.	Availability	License
Mana TTS	**$\sim$ 86**	**1**	**√**	**√**	**Avail.**	**CC-0 1.0**
Arman TTS (2023)	$\sim$ 9	1	√	√	Not Avail.	Unknown
AmerAndish (2022)	21	1	√	√	Not Avail.	Unknown
tts dataset (2024a)	$\sim$ 16	1	√	√	Avail.	Unknown
TTS audio (2024)	$\sim$26	1	√	√	Avail.	Proprietary
Persian TTS (2024)	+30	1	√	√	Not Avail.	Unknown
tts-famale (2024b)	$\sim$ 30	1	√	×	Avail.	CC-0 1.0
tts-male (2024c)	$\sim$ 38	1	√	×	Avail.	CC-0 1.0
Persian Speech (2017)	$\sim$ 2.5	1	√	√	Avail.	CC BY-NC-SA 4.0
ParsiGoo (2024)	$\sim$ 5	6	√	×	Avail.	CC BY-SA 4.0
DeepMine Multi-TTS (2023)	120	67	×	√	Avail. on req.	Unknown

There is a distinct pipeline for processing the text files, as summarized in Figure 1b. The text data undergoes normalization using Hazm (RoshanAI, 2024) normalizer. This step is crucial as it standardizes the words, ensuring consistency. This simplification reduces the unnecessary details that the TTS model must handle.

The subsequent three steps aim to remove links and references that are typically not meant to be read aloud. These encompass all inline references in the form `[NUM]`, end-of-text references such as author names and book details, and end-of-text links, including lines containing URLs.

In the next phase, we addressed how numbers are written differently from how they're spoken. We used the parsi-io (ParsIO, 2023) tool to detect numbers in the text and convert them into spoken equivalents. Afterward, we trimmed non-essential symbols to streamline the text and decrease the input given to the model. Lastly, we eliminated any extra whitespace, including empty lines, to read the text for alignment with audio components in later phases, as explained in more detail below.

3.2 Alignment

Alignment involves matching audio to its transcript. We have divided this task into two phases. Firstly, we ensure that each audio file contains the same initial and final content as the corresponding text file, which we refer to as start-end alignment. Secondly, we segment the extensive audio and text files into smaller pieces, typically a few seconds and a few words, in a process known as forced alignment.

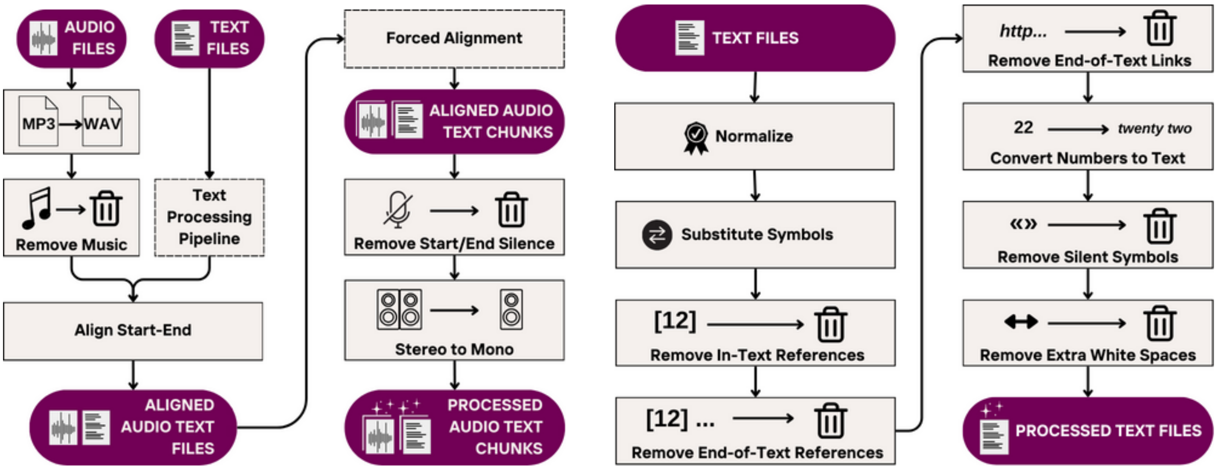
Manual alignment is an arduous task, especially on large datasets. Therefore, we opted to automate this process. Our approach for both alignment phases is rooted in automatic speech recognition (ASR) models. We developed a module that generates reliable hypothesis transcripts for each audio chunk. Subsequently, we match the hypothesis with the corresponding segment in the ground truth text.

3.2.1 Transcription Module

The transcription task can be as straightforward as utilizing a single reliable ASR model to obtain the transcript of a given audio chunk. However, this wasn't our scenario because no openly accessible Persian ASR model is sufficiently reliable to handle this task alone. For instance, one common issue with ASR models was the occasional generation of truncated transcripts for some input cases. Consequently, we chose to integrate multiple ASR models into a transcription module and implement a form of majority voting among them. This approach allows errors from one or two models to be concealed, significantly reducing the likelihood of such defects appearing in the output transcripts.

We utilized five of the top open-access Persian ASR models. We deliberately selected the ASR models and all tools in our pipeline from the open-source domain, enabling us to publish our work under non-restrictive licenses. A list of the tools used, including the ASR models, can be found in Table 11.

Some ASR models were accompanied by a re-



(a) Processing pipeline for audio and text files.

(b) Processing pipeline for text files.

Figure 1: Dataset processing pipelines.

ported Word Error Rate (WER). However, they were assessed on different test sets, making them incomparable. To rank and compare the ASR models based on their error rates, we gathered and processed the PersianInformal dataset, evaluating the models accordingly. Further details regarding this dataset, including its collection method and evaluation results, are provided in the Appendix B.

The input to the transcription module is a small audio chunk, typically less than 20 seconds in duration. The output comprises a list of eligible transcripts sorted by the reliability of their corresponding ASR models. Figure 2 depicts an overview of this module.

The transcripts are generated as follows: initially, a given audio chunk is input into all ASR models. Subsequently, any transcripts shorter than 80% of the longest is discarded. This step helps address the issue of incomplete transcripts, which was mentioned before. The remaining transcripts that meet the length criteria are then sorted based on the performance of their respective ASR models and returned in a list. Some insightful statistics of this module can be found in Appendix C.

3.2.2 Start/End Alignment

In the raw dataset, each audio file is paired with a corresponding text file. However, certain factors can cause inconsistencies between the starting/ending points of the audio and its associated text, necessitating start-end alignment processes that may involve removing a few seconds or words from each. The primary factors include:

- The title and author name are read by the speaker, even though not included in the text.
- Additional resources at the end of the text that are typically unread.

The general workflow of the start-end alignment is as follows: the audio is initially segmented based on silent moments, then a search iterates over these segments as potential starting (or ending) points for the audio. For each segment, the most reliable hypothesis transcript is obtained from the transcription module, and the text is searched to find the best matching interval. The pair of trimmed audio and text with the lowest Character Error Rate (CER) between the hypothesis transcript and reference text is selected to determine the start and end of the files.

3.2.3 Forced Alignment

The start-end alignment phase generates pairs of audio and text files that are perfectly matched at the beginning and end. However, this format isn't suitable for feeding into a TTS model. The audio and text files must be divided into smaller chunks, typically a few to 15 seconds each. This process is commonly known as forced alignment.

In our search for a forced alignment tool for Persian, we considered Aeneas (ReadBeyond), which is known for its large community and high performance. However, as noted in their project limitations, "Audio should match the text: large portions of spurious text or audio might produce a wrong sync map." This limitation made Aeneas unsuitable for our needs, as the audio and text could have mismatches due to factors such as:

- The speaker was provided with a slightly different version of the text to read.

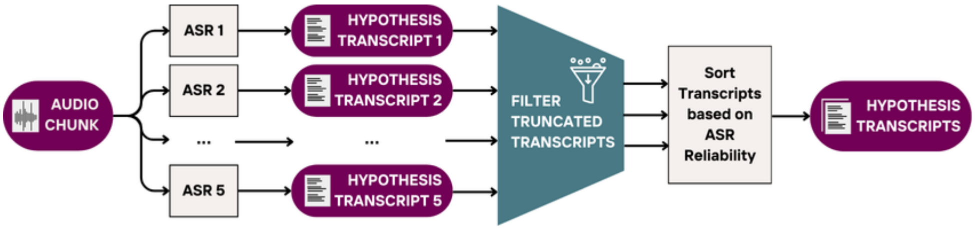


Figure 2: Transcription Module

- The speaker might censor some parts of the text.
- The speaker might make a mistake and repeat herself to correct it.

As a result, we decided to develop our method for forced alignment. The primary workflow of our forced alignment algorithm is illustrated in Figure 11 in Appendix E.

Initially, the audio is divided into smaller parts using silent intervals. We ensure these audio parts are between 2 and 12 seconds by combining smaller segments or changing the silence detection setting to create smaller parts. The last step is to find a matching section from the reference text. We use the hypothesis transcripts provided by the transcription module until we find a subsection of the text that meets the desired similarity criteria.

The algorithm employs two search methods to find matching text: Interval Search and Gapped Search. Interval Search seeks all sub-strings of the text in the form of $text[s:i]$ within a defined range. As the name suggests, the Gapped Search would let a missing gap in the text and look for sub-strings of the form $text[s:j] + text[k:i]$.

The search process halts immediately upon finding a match with $CER \leq 0.05$. Moreover, due to the lower computational cost of Interval Search, matches with $0.05 < CER \leq 0.2$ at the end of this search are also accepted, avoiding the initiation of the Gapped Search. Suppose neither of the search methods can find a matching substring with $CER \leq 0.2$. In that case, the process iteratively tests the next hypothesis transcripts until all options are exhausted, resulting in the complete rejection of the chunk.

3.3 Post-Processing

In this phase, the audio chunks undergo processing to eliminate any silent segments lasting more than 1 second. It's worth noting that silence removal occurs after forced alignment because silent moments are utilized in segmenting the audio into smaller chunks and should not be removed beforehand. For this task, we utilize the Pydub (James Robert, 2022) silence module, which is also used for segmenting the audio into chunks. The audio chunks are converted from stereo to mono as the final step.

4 Statistics

Raw Files: Nasl-e-Mana maintains an archive of over 41 magazines spanning over three years. The archive comprises a total of over 568 audio-text files. The audio files underwent manual inspection to ensure the dataset consisted of single-speaker recordings. Following this review, four files were excluded from the raw material, resulting in 564 files for automated processing. The duration of the audio files ranged from approximately 0.5 to 34 minutes, with an average duration of about 10 minutes. Similarly, the lengths of the text files varied, with word counts ranging from 44 to 3951 and an average length of 1234 words.

Processed Chunks Count: Executing the pipeline on the raw material yielded a minimum, maximum, and average of approximately 4, 398, and 118 chunks per file, respectively, totaling 66172 chunks overall. Roughly 97.98% of these chunks were automatically accepted as having good quality, while 1338 (about 2.02%) were rejected due to an unacceptable CER between the hypothesis transcript and the matching text. Consequently, the final dataset comprises 64834 accepted audio-text chunks.

Accepted Chunks Duration: As previously mentioned, our pipeline's chunking method guarantees that audio chunks have durations ranging from a minimum of 2 seconds to a maximum of 12 seconds. The histogram depicting the duration distribution of the audio chunks is illustrated in Figure 3.

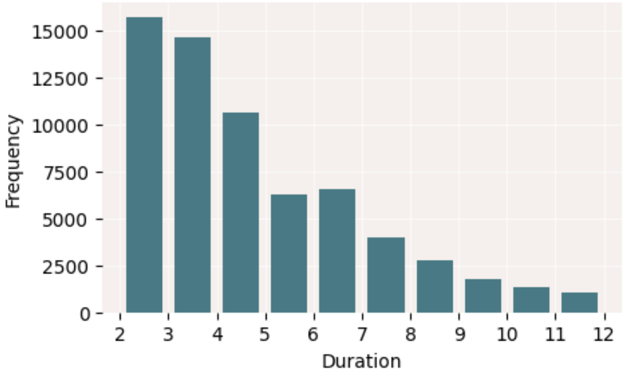


Figure 3: Distribution of the duration of audio chunks.

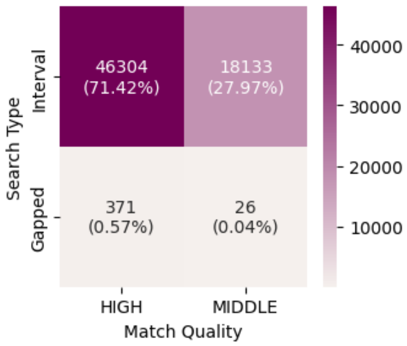


Figure 4: Distribution of search type and match quality of accepted chunks.

Accepted chunks search type: As outlined in previous sections, two search methods matched a hypothesis transcript with the ground truth text: Interval Search and Gapped Search. It should be pointed out that the Interval Search method is preferred because of its lower computational cost. Consequently, if a match meeting the acceptance criteria is found through Interval Search, further searching with the Gapped Search is unnecessary. Gapped Search is primarily utilized when the ground truth text does not perfectly align with the hypothesis. Analysis of chunk information reveals that approximately 99.39% of matching text chunks are identified through Interval Search, while the remaining 0.61% (397 chunks) are the result of Gapped Search.

Accepted chunks match quality: As mentioned, there are two threshold levels for CER of audio chunks. The first, labeled HIGH, signifies a match between the hypothesis and ground truth text with a CER less than 0.05. The second, labeled MIDDLE, denotes an acceptable CER of 0.05 to 0.2. Attaining the HIGH threshold during the search prompts an immediate acceptance of the chunk, whereas achieving the MIDDLE threshold would only accept the chunk at the end of each search type.

Approximately 71.46% of the chunks (46330 in total) have the HIGH and about 28.54% of the chunks (18504 in total) have the MIDDLE match quality. Figure 4 illustrates the joint distribution of the search type and match quality of the audio-text chunks, as they are correlated.

It's also intriguing to visualize the distribution of CER values for all the chunks that passed through the dataset creation pipeline. As illustrated in Figure 5, there is only a small number of rejected chunks compared to the matched chunks in the other groups. Manual investigations indicate that these rejected chunks are primarily associated with an underlying discrepancy between the raw audio and text files. Other reasons for rejection include utterances that differ in their written and spoken forms, which will be further discussed in the section 8.

Word count: The accepted chunks exhibit a range of 1 to 38 words, with an average of approximately 11 words per chunk. The histogram illustrating the distribution of word counts can be found in Figure 6. Overall, the dataset contains a total of approximately 24,113 unique words.

5 Experiments

In this section, we present the experiments conducted to evaluate the quality and accuracy of the ManaTTS dataset and the TTS model trained on it. The experiments are divided into two main parts: (1) assessing the quality of the trained TTS model and (2) measuring the transcription accuracy of the dataset.

5.1 Trained TTS Model Quality

5.1.1 Training Setup

To assess the quality and efficacy of the ManaTTS dataset, we trained a Tacotron2-based TTS model (Jia et al., 2018) from the ground up using this corpus. This model comprises three main components: First, a speaker encoder trained on extensive untranscribed datasets, enabling extraction of speaker characteristics from mere seconds of speech. Second, a sequence-to-sequence Tacotron2-based model is tasked with converting text into mel-spectrograms, and the final component is a vocoder responsible for transforming the

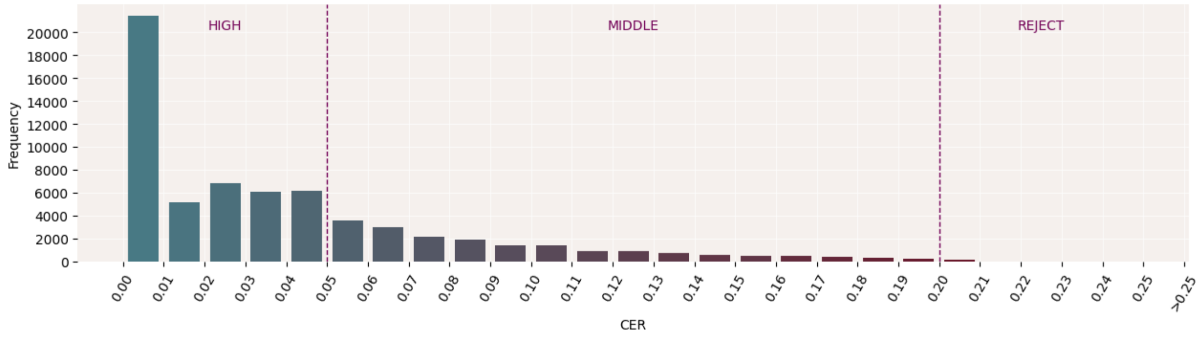


Figure 5: Distribution of CER values across all chunks. The vertical lines denote the threshold values for the HIGH, MIDDLE, and REJECT match qualities as discussed in the section 3.2.3.

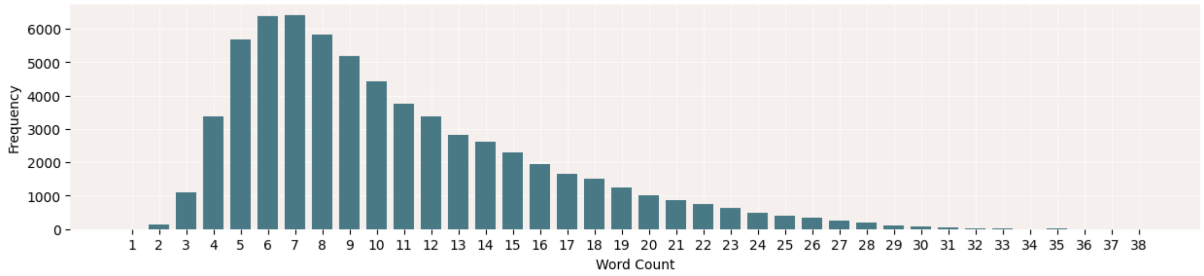


Figure 6: Word count distribution of accepted text chunks.

mel-spectrogram into the speech waveform.

We adopt a Persian language setup for the text-to-mel-spectrogram module (Adibian). The input data undergo resampling to 24 kHz and preprocessing using an FFT size of 2048 and 80 mel-frequency filter banks. We then conduct six training sessions. The learning rate begins at 1e-3 and gradually reduces to 1e-5 in the final session, while the batch size is fixed at 16. With these parameters, the model undergoes training for 320,000 steps and is subsequently used for synthesizing samples for evaluation. [3]

For the speaker encoder and vocoder components, we used the pre-trained encoder and HiFi-GAN (Kong et al., 2020) vocoder from the previously mentioned work (Adibian). HiFi-GAN is trained adversarially, where the generator synthesizes waveforms from spectrograms, and the discriminator distinguishes between synthetic and real waveforms. Due to its non-auto-regressive nature, HiFi-GAN operates faster than earlier vocoders while achieving superior speech quality.

5.1.2 Evaluation

To evaluate the TTS model trained on ManaTTS, we selected five utterances from the latest issue of Nasl-e-Mana magazine that were not included in the training data. We then generated the speech waveform of the selected utterances using the following sources:

1. **Baseline Model 1:** A VITS-based TTS model trained for the Persian language with an open-access model (Kamtera, 2024b).
2. **Baseline Model 2:** Another open-access model based on Glow-TTS trained for the Persian language (Kamtera, 2024a).
3. **GT Spec:** Waveforms generated from the ground truth spectrogram of the natural speech using the same HiFi-GAN vocoder as our work.
4. **Ours:** The model used in our study, which generates spectrograms for a given utterance, and the waveform is generated using a HiFi-GAN vocoder.

The inclusion of the **GT Spec** source is critical because the vocoder used in our model was not trained on our dataset. To ensure a fair evaluation of the model's spectrogram generation capability,

[3] The full settings and scripts used for training are available at https://github.com/MahtaFetrat/Persian-MultiSpeaker-Tacotron2

we extracted the mel-spectrograms from the natural utterances and synthesized their waveforms using the pre-trained vocoder. This allowed us to compare the model's outputs with waveforms generated from ground-truth spectrograms.

We complemented these four sources with the natural audio chunks of the five selected samples and conducted a subjective Mean Opinion Score (MOS) test involving 76 native Persian speakers.[4]

The subjects were prompted as follows: "Rate the voices you hear based on how natural they sound and how likely they are to have been uttered by a human. If you think the voice is completely natural and has no problems, rate it 5. Otherwise, decrease the rating down to 1 based on how robotic it sounds and the problems or noises you notice."

The order of the models for each of the five utterances was shuffled to prevent bias towards consistently rating a specific model the same or being influenced by an increasing/decreasing naturalness trend. They were also not informed which utterances were related to our work or that there was a natural utterance for each sample.

The final MOS scores are presented in Table 2. For a more detailed breakdown of the results, please refer to Appendix D.

5.2 Transcription Accuracy

In the processing pipeline, text-audio pairs are accepted into the dataset only if the character error rate (CER) between the audio transcript and the selected text is at most 0.2. However, this threshold inherently includes errors introduced by the inaccuracy of automatic speech recognition (ASR) systems. To evaluate the actual difference between the audio content and the suggested texts, we conducted an experiment in which we randomly selected 100 audio-text chunks and manually transcribed their audio content. We then calculated the CER between these manual transcripts and the corresponding text pairs from the dataset. The resulting CER was only 0.01, which is remarkably small, indicating that the text files are nearly perfect matches for the audio content.

6 Discussion

The absence of high-quality, open-source/open-access text-to-speech models and datasets for the Persian language has been highlighted in section A. Below are some of the critical challenges associated with the available corpora.

- The dataset is either inaccessible, lacks a specified license, or is under a restrictive license.
- The dataset contains utterances from a limited domain, such as religious contexts exclusively.
- The speech is synthesized using a text-to-speech model.
- The text is synthesized using a speech-to-text model and has not been adequately verified.
- The dataset is limited in size.

The ManaTTS dataset introduced in this work is the first Persian language text-to-speech corpus that addresses all the above challenges. This dataset is publicly available under an open CC-0 1.0 license. It includes utterances from a monthly magazine over three years and covers various Persian language utterances. The speech and ground truth text in this dataset are collected by human agents, not synthesized. Most notably, it is the largest single-speaker text-to-speech dataset available to date.

Another notable contribution of this work is that it represents the first fully open-source text-to-speech data collection project for the Persian language. Due to the open code base, all steps, including data crawling, processing, and model training, are reproducible. This approach helps develop additional Persian speech datasets and offers two fundamental benefits.

First, in addition to the standard tools used in typical speech data processing, this project introduces a new sentence tokenization method and a new start-end alignment and forced alignment tool capable of aligning speech and text pairs that are not exact matches but are slightly different. Our work demonstrates that this tool can be effectively used with publicly available Persian ASR models of moderate accuracy, thus contributing yet another open-license tool to the community.

Second, the ever-growing monthly magazine of Nasl-e-Mana and the fully available data collection and processing pipeline make this dataset easily extendable. Future work can obtain an even larger dataset by executing a few scripts.

The experiments reveal that the TTS model, trained on the ManaTTS dataset, achieved a MOS of 3.76, slightly lower than natural spectrograms

[4] For other objective methods, please refer to Appendix D.

Table 2: Subjective assessment of outcomes of the TTS models. **GT Spec** refers to the utterances with ground truth spectrograms but HiFi-GAN-synthesized waveforms, and **GT Waveform** refers to the natural speech samples. The values are presented as mean opinion score (MOS) ± standard deviation (std).

Source	VITS	Glow	**Ours**	GT Spec	GT Waveform
MOS	1.68 ± 0.80	1.34 ± 0.70	$\mathbf{3.76 \pm 1.04}$	3.86 ± 1.04	4.01 ± 1.14

at 3.86 and natural speech at 4.01. It has even outperformed the ground truth spectrogram and waveform samples in some of the test utterances (refer to Appendix D). This implies that our dataset has acceptable quality and can be used effectively to train Persian text-to-speech models.

7 Conclusion

In this work, we proposed a processing pipeline to create a TTS dataset from raw speech and text files. Applying the pipeline to the archive of the Nasl-e-Mana magazine, we published ManaTTS, the largest single-speaker Persian TTS dataset, under the open CC-0 1.0 license. Additionally, we released a smaller transcribed speech dataset, Persian-Informal, which serves as a valuable test dataset for evaluating ASR models and is utilized in our novel forced alignment method. Evaluating ManaTTS involved training a Tacotron2-based TTS model. The samples synthesized by this model exhibited remarkable naturalness, comparing favorably to both the utterances generated from gold speech spectrograms and natural speech waveforms.

8 Limitations

Despite our work's contributions, there are some limitations to it. Firstly, although the transcripts are acquired using multiple ASR models and are used to find a match from the ground truth text only if they satisfy certain CER thresholds, this process is not entirely deterministic and is prone to minor errors that might not significantly affect the CER value. Thus, employing superior ASR systems with stricter CER thresholds might further reduce such errors.

Secondly, owing to English's pervasive nature, text in other languages may incorporate English words and phrases. However, our pipeline lacks an explicit mechanism to match these phrases between speech and text. Therefore, if the TTS model is expected to detect and vocalize English subtexts, the pipeline should be modified to include more of these examples in the dataset.

Thirdly, while the pipeline incorporates a mechanism to match the spoken form of numbers, some specifically formatted numeric words remain challenging. For instance, a phone number might be pronounced in various ways, such as date or time. Therefore, a tool capable of converting between these differences in the spoken and written forms of specific numeric data and symbols for the Persian language would be highly beneficial.

Finally, despite the public availability of the raw data, it's important to recognize potential misuse concerns arising from our dataset or processing pipeline, such as voice impersonation. Anonymization emerges as a solution to mitigate these risks, ensuring responsible dataset usage in alignment with privacy and ethical considerations.

Acknowledgments

We thank Nasl-e-Mana magazine for its invaluable work and for being so generous with the published dataset license. We also extend our gratitude to the Iran Blind Non-governmental Organization for its support and guidance regarding the need for open access initiatives in this domain.

Funding Disclosure: No funding was received for conducting this study.

References

Blizzard Challenge 2013 Dataset. https://www.synsig.org/index.php/Blizzard_Challenge_2013#Data_download. Accessed: April 22, 2024.

Common Voice Dataset. https://commonvoice.mozilla.org/en/datasets. Accessed: April 22, 2024.

Farsspon persian speech recognition system. https://asr-gooyesh.com/fa/shop/%D8%AF%D8%A7%D8%AF%DA%AF%D8%A7%D9%86-farsspon/. Accessed: April 22, 2024.

Virgool. https://virgool.io/. Accessed: June 1, 2024.

2001. Farsdat (farsi speech database) corpus. ELRA Catalog. Accessed: April 22, 2024.

2016. Large farsdat corpus. ELRA Catalog. Accessed: April 22, 2024.

2017. Persian speech corpus website. https://fa.persianspeechcorpus.com/. Accessed: April 22, 2024.

2024. JiWER. https://github.com/jitsi/jiwer. Accessed: May 3, 2024.

Majid Adibian. Persian-MultiSpeaker-Tacotron2. https://github.com/Adibian/Persian-MultiSpeaker-Tacotron2. Accessed on: May 4, 2024.

Majid Adibian, Hossein Zeinali, and Soroush Barmaki. 2023. Deepmine-multi-tts: A persian speech corpus for multi-speaker text-to-speech. Available at SSRN 4673655.

AlisterTA. 2024. Persian Text-to-Speech Repository. https://github.com/AlisterTA/Persian-text-to-speech?tab=readme-ov-file. Accessed: April 22, 2024.

Alpha Cephei Inc. 2024. Vosk - offline speech recognition api. https://alphacephei.com/vosk/. Accessed: May 23, 2024.

Amir Pourmand. 2024. Automatic Speech Recognition Farsi YouTube. https://www.kaggle.com/datasets/amirpourmand/automatic-speech-recognition-farsi-youtube. Accessed: April 22, 2024.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-fi multi-speaker english tts dataset. arXiv preprint arXiv:2104.01497.

Balad. Balad. https://balad.ir/. Accessed: June 1, 2024.

Mohammad Hadi Bashari. 2021. Perpos: Cross platform persian pos tagger. https://github.com/mhbashari/perpos. Accessed: May 23, 2024.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5971–5975. IEEE.

Caito. M-AILABS Speech Dataset. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/. Accessed: April 22, 2024.

Elham Pourabbas and Kaveh Taghipour and Fatemeh Salehi. 2022. Persian speech emotion recognition dataset. https://zenodo.org/records/7486182. Accessed: April 22, 2024.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. arXiv preprint arXiv:2111.09344.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. Communications of the ACM, 64(12):86–92.

Rafael Mosquera Gómez, Julián Eusse, Juan Ciro, Daniel Galvez, Ryan Hileman, Kurt Bollacker, and David Kanter. 2023. Speech wikimedia: A 77 language multilingual speech dataset. arXiv preprint arXiv:2308.15710.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, 5(50):2154. Deezer Research.

Hugging Face. LJ Speech Dataset. https://huggingface.co/datasets/lj_speech. Accessed: April 22, 2024.

James Robert. 2022. Pydub. https://github.com/jiaaro/pydub. Accessed: May 3, 2024.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Advances in neural information processing systems, 31.

Kamtera. 2024. ParsiGoo. https://huggingface.co/datasets/Kamtera/ParsiGoo. Accessed: April 22, 2024.

Kamtera. 2024a. Persian tts female glow_tts. https://huggingface.co/Kamtera/persian-tts-female-glow_tts. Accessed: May 31, 2024.

Kamtera. 2024b. Persian tts female vits. https://huggingface.co/Kamtera/persian-tts-female-vits. Accessed: May 31, 2024.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in neural information processing systems, 33:17022–17033.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of IEEE pacific rim conference on communications computers and signal processing, volume 1, pages 125–128. IEEE.

Magnoliasis. 2024a. Persian TTS Dataset. https://www.kaggle.com/datasets/magnoliasis/persian-tts-dataset. Accessed: April 22, 2024.

Magnoliasis. 2024b. Persian TTS Dataset (Female). https://www.kaggle.com/datasets/magnoliasis/persian-tts-dataset-famale. Accessed: April 22, 2024.

Magnoliasis. 2024c. Persian TTS Dataset (Male). https://www.kaggle.com/datasets/magnoliasis/persian-tts-dataset-male. Accessed: April 22, 2024.

Saber Malekzadeh, Mohammad Hossein Gholizadeh, and Seyed Naser Razavi. 2018. Persian vowel recognition with mfcc and ann on pcvc speech dataset. arXiv preprint arXiv:1812.06953.

Maps. Google maps. https://www.google.com/maps. Accessed: June 1, 2024.

Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, et al. 2021. Multilingual spoken words corpus. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

M Mehdadfi. 2021. Wav2vec2-large-xlsr-persian-v3. https://huggingface.co/m3hrdadfi/wav2vec2-large-xlsr-persian-v3. Accessed: May 23, 2024.

Moradi. 2024. Persian Text-to-Speech Audio Dataset. https://www.kaggle.com/datasets/moradi/persian-texttospeech-audio. Accessed: April 22, 2024.

MurfAI. Murf ai. https://murf.ai/. Accessed: June 1, 2024.

Navid Naderi, Babak Nasersharif, and Amirhossein Nikoofard. 2022. Persian speech synthesis using enhanced tacotron based on multi-resolution convolution layers and a convex optimization method. Multimedia Tools and Applications, 81(3):3629–3645.

Narakeet. Persian text to speech. https://www.narakeet.com/languages/persian-text-to-speech/. Accessed: June 1, 2024.

NasleMana. Nasl-e-mana magazine. https://naslemana.com/. Accessed: June 1, 2024.

NV. Nv access. https://www.nvaccess.org/. Accessed: June 1, 2024.

OpenAI. 2023. Gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-10-15.

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.

Seyyed Mohammad Masoud Paparnchi. 2021. wav2vec2-xlsr-multilingual-53-fa. https://github.com/Hamtech-ai/wav2vec2-fa. Accessed: May 23, 2024.

ParsiIO. 2023. Parsi.io. https://github.com/language-ml/parsi.io. Accessed: May 3, 2024.

PersianDataset. 2024. PersianSpeech: Persian Speech Dataset. https://github.com/persiandataset/PersianSpeech. Accessed: April 22, 2024.

Javad Peymanfard, Samin Heydarian, Ali Lashini, Hossein Zeinali, Mohammad Reza Mohammadi, and Nasser Mozayani. 2024. A multi-purpose audio-visual corpus for multi-modal persian speech recognition: The arman-av dataset. Expert Systems with Applications, 238:121648.

Javad Peymanfard, Ali Lashini, Samin Heydarian, Hossein Zeinali, and Nasser Mozayani. 2022. Word-level persian lipreading dataset. In 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), pages 225–230. IEEE.

Ralireza. 2024. PSDR: Persian Speech Dataset for Recognition. https://github.com/Ralireza/PSDR. Accessed: April 22, 2024.

Mirco Ravanelli, Titouan Parcollet, Aku Rouhe, Peter Plantinga, Elena Rastorgueva, Loren Lugosch, Nauman Dawalatabad, Chou Ju-Chieh, Abdel Heba, Francois Grondin, William Aris, Chien-Feng Liao, Samuele Cornell, Sung-Lin Yeh, Hwidong Na, Yan Gao, Szu-Wei Fu, Cem Subakan, Renato De Mori, and Yoshua Bengio. 2021. Speechbrain. https://github.com/speechbrain/speechbrain. Accessed: May 23, 2024.

ReadBeyond. Aeneas. https://github.com/readbeyond/aeneas. Accessed: May 3, 2024.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), volume 2, pages 749–752. IEEE.

RoshanAI. 2024. Hazm. https://www.roshan-ai.ir/hazm/docs/index.html. Accessed: May 3, 2024.

Sadra Sabouri, Elnaz Rahmati, Soroush Gooran, and Hossein Sameti. 2022. naab: A ready-to-use plug-and-play corpus for farsi. arXiv preprint arXiv:2208.13486.

Seyed Saleh Hosseini. 2021. Persian Speech Recognition. https://github.com/seyedsaleh/persian-speech-recognition. Accessed: April 22, 2024.

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

Aryan Shekarlaban and Pooya Mohammadi Kazaj. 2023. Hezar: The all-in-one ai library for persian. https://github.com/hezarai/hezar. Accessed: May 3, 2024.

shenasa-ai. 2024. Speech2Text Repository. https://github.com/shenasa-ai/speech2text. Accessed: April 22, 2024.

SpeechBrain. 2023. Whisper large-v2 fine-tuned on commonvoice persian. https://huggingface.co/speechbrain/asr-whisper-large-v2-commonvoice-fa. Accessed: June 4, 2024.

Speechify. Speechify for education. https://speechify.com/edu/. Accessed: June 1, 2024.

Veaux, Christophe and Yamagishi, Junichi and MacDonald, Kirsten. 2017. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. https://datashare.ed.ac.uk/handle/10283/2651. Accessed: April 22, 2024.

Rohola Zandie, Mohammad H Mahoor, Julia Madsen, and Eshrat S Emamian. 2021. Ryanspeech: A corpus for conversational text-to-speech synthesis. arXiv preprint arXiv:2106.08468.

Hossein Zeinali, Lukáš Burget, and Jan Honza 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.

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. arXiv preprint arXiv:1904.02882.

Anthony Zhang. 2017. Speech recognition (version 3.8). https://github.com/Uberi/speech_recognition. Software.

A Extended Discussion of Related Works

Recent advancements in speech recognition and synthesis techniques have led to numerous projects developing systems for the Persian language. Each project has its own dataset, each with unique advantages and limitations. This section comprehensively reviews Persian speech datasets and their corresponding transcripts, detailing their respective merits and drawbacks.

Our analysis encompasses a diverse range of Persian datasets, including text-to-speech (TTS) datasets (Table 1), automatic speech recognition (ASR) datasets, audio-visual speech recognition (AVSR) datasets tailored for Persian, a dataset designed explicitly for Persian phoneme recognition (PR), a Persian spoken digit recognition (DR) dataset, and multilingual datasets incorporating Persian language components (Table 3). While our primary focus and discussions in this section center around Persian speech corpora for text-to-speech systems, we have also included notable English TTS speech datasets for a comprehensive comparison (Table 4).

It is worth noting that the availability of single-speaker TTS datasets in Persian is notably limited compared to their English counterparts. Moreover, the average size of English TTS datasets significantly surpasses that of Persian datasets. This highlights a crucial gap and emphasizes the pressing need for comprehensive, single-speaker Persian datasets to drive progress in research and application development within this domain. In the subsequent part of this section, we will delve into the specifics of each Persian TTS dataset listed in Table 1.

ArmanTTS (Shamgholi et al., 2023) is a prominent single-speaker TTS dataset for the Persian language, comprising approximately 9 hours of audio recorded in a professional studio setting at a sampling rate of 22.05 kHz. The audio files are typically about 2.5 seconds long (with a maximum of 12.5 seconds), corresponding to approximately 5 words (and up to a maximum of 30 words), with an average signal-to-noise ratio of 25 dB. Unfortunately, this dataset has not been publicly available yet and the authors have not provided options for access upon request or specified any licensing terms for its use.

AmerAndish (Naderi et al., 2022) introduced by Naderi et al., is derived from audio books read by a single female speaker. They used a set of automatic tools to read text of PDF files, remove audio noise, and remove audio clips with a different speaker. However, the task of splitting the audio to chunks and matching the chunks with some reference text was performed manually by human agents and later double checked with an ASR system to remove potential mismatches. The resulting dataset includes chunks of 1-12 seconds and summing up to 21 hours of audio and matching text. Unfortunately, the authors have not provided any means of accessing the dataset or issued a license for its use.

persian tts dataset (Magnoliasis, 2024a) represents another single-speaker Persian resource, featuring approximately 15.6 hours of audio. While the dataset owners have not provided a detailed description, it is evident that the audio is derived from a Persian translation of the Holy Quran. Regrettably, the owners have also not provided any licensing information for the dataset, and it remains unclear whether the audiobook and the corresponding text are free from copyright restrictions. Furthermore, the dataset's exclusive focus on the Holy Quran means it lacks topical and lexical diversity, which is a substantial limitation for developing TTS systems that require a broad range of vocabulary and expressions to perform effectively in varied contexts.

Persian text-to-speech audio (Moradi, 2024) is a single-speaker corpus with 26 hours of content. This dataset, derived from a Persian translation of the Holy Quran, lacks a detailed description. Similarly to prior datasets, copyright details and licensing status are not provided by the dataset owners, leaving all rights reserved to the original authors.

Persian-text-to-speech (AlisterTA, 2024) details a Persian TTS model project. Researchers compiled a dataset from over 30 hours of audio sourced from commercially purchased audiobooks, narrated by a female speaker. They segmented the audio into chunks ranging from 3 to 14 seconds using silence detection, then manually aligned these chunks with their corresponding texts. Notably, the purchase of the audiobooks implies copyright restrictions, rendering the dataset non-public and unlicensed.

persian tts dataset (female) (Magnoliasis, 2024b) is a single-speaker Persian dataset under a CC-0 license, comprising 30 hours of audio synthesized

Table 3: List of other Persian datasets including speech and text. The datasets indicated by a plus sign are multilingual, but only the information for the Persian part is shown. **Size** shows the duration in hours, **Spks.** stands for the number of speakers, and the columns **N.T.** and **N.A.** abbreviate Natural Text and Natural Audio as in Table 1. **Comm.** stands for a commercial license.

Dataset	**Usage**	**Size**	**Spks.**	**N.T.**	**N.A.**	**Availability**	**License**
DeepMine+ (2019)	ASR	+480	+1850	√	√	Paid	Proprietary
CMU Wilderness+ (2019)	ASR	5	1	√	√	Avail.	CC-0 1.0
MLCommons+ (2021)	ASR	327	-	√	√	Avail.	CC BY 4.0
Speech Wikimedia+ (2023)	ASR	∼ 0.13	-	√	√	Avail.	CC BY-SA
PersianSpeech (2024)	ASR	∼ 3	-	√	√	Avail.	MIT
PersianSpeech (2024)	ASR	∼ 86	-	-	-	Avail. on req.	MIT
Persian STT (2022)	ASR	-	-	×	√	Not Avail.	CC BY 4.0
Small Farsdat (2001)	ASR	5	300	√	√	Paid	Non comm. Comm.
Large Farsdat (2016)	ASR	∼ 73	100	√	√	Paid	Non comm. Comm.
ASR Farsi (2024)	ASR	+300	-	√	√	Avail.	CC-0 1.0
CommonVoice+ ()	ASR	416	-	√	√	Avail.	CC-0 1.0
FarsSpon ()	ASR	+530	+5300	√	√	Paid	Proprietary
Shenasa (2024)	ASR	∼300	-	√	√	Avail.	GPL-3.0
Persian-SR (2021)	ASR	-	-	√	√	Avail. on req.	MIT
Arman AV (2024)	AVSR	220	1760	√	√	Avail. on req.	Proprietary
PLRW (2022)	AVSR	30	1800	×	√	Not Avail.	CC BY 4.0
PCVC (2018)	PR	∼ 1	12	√	√	Avail.	GPL-3.0
PSDR (2024)	DR	-	-	√	√	Avail.	GPL-3.0

Table 4: List of well-known English datasets including speech and text. **Size** indicates the duration in hours, and **Non comm.** stands for a non-commercial license.

Dataset	**Usage**	**Size**	**Speakers**	**Availability**	**License**
Hi-Fi TTS (2021)	TTS	292	10	Avail.	CC BY 4.0
Libri TTS (2019)	TTS	585	2456	Avail.	CC BY 4.0
BC2013 ()	TTS	300	1	Avail. on req.	Non comm.
VCTK (2017)	TTS	44	109	Avail.	ODC-BY v1.0
LibriSpeech (2015)	ASR	982	2484	Avail.	CC BY 4.0
People's Speech (2021)	ASR	30k	-	Avail.	CC BY-SA
RyanSpeech (2021)	ASR/TTS	10	1	Avail. on req.	CC BY 4.0
LJSpeech ()	ASR/TTS	24	1	Avail.	CC-0 1.0
MAILABS ()	ASR/TTS	75	2	Avail.	BSD

from the Persian text corpus Naab (Sabouri et al., 2022). The dataset's primary limitation is its fully synthesized audio content, which restricts the performance of TTS models trained on it, as they cannot reach the naturalness of human speech due to the inherent constraints of the used synthesizer.

persian-tts-dataset-male (Magnoliasis, 2024c) unveils a CC-0 licensed, single-speaker Persian dataset, containing approximately 38 hours of audio. The dataset documentation lacks specifics regarding the source and data collection methodology. However, Initial manual analysis of several audio samples indicates that the content was synthesized using another TTS model.

Persian Speech Corpus (per, 2017) presents a Persian TTS dataset, licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Featuring recordings from a single male speaker, this corpus encompasses professionally produced studio-quality audio, but totals only 2.5 hours, which may be considered brief for extensive TTS research and development.

ParsiGoo (Kamtera, 2024) introduces a multi-speaker TTS dataset tailored for the Persian language, secured under the CC BY-SA 4.0 license. This collection comprises about 5 hours of audio, recorded at a sampling rate of 22.05 kHz. The dataset features four distinct speaking styles across six speakers, enhancing its diversity. However, detailed provenance of the audio sources remains unspecified. Manual examination reveals that audio from five speakers is synthesized, while recordings from the sixth speaker are authentically vocal. Regrettably, there is no information provided on the copyright status of these audio files.

DeepMine Multi-TTS (Adibian et al., 2023) is the first large-scale multi-speaker Persian TTS dataset. It encompasses 120 hours of audio recordings sampled at a rate of 22.05 kHz, featuring contributions from 67 speakers. The dataset primarily consists of audio files obtained from a platform hosting public and freely accessible audio-books. The audio tracks have been processed manually to remove parts that included background music. The transcripts of this dataset were generated using a specific ASR system and then checked manually. The resulting chunks vary in length from 0 to 14 seconds but are mostly between 1 to 10 seconds long. Although this dataset has not been published publicly and lacks a specified license, the authors note that the data will be available on request for only research purposes.

B Evaluation of ASR Models

As detailed in Section 3.2, alignment tools necessitate the ASR models to be arranged based on their reliability. This section elucidates the process undertaken to conduct this assessment.

B.1 PersianInformal Dataset

To ensure a proper assessment of the ASR models, we required a dataset that had not been seen by these models during their training phase. However, many existing ASR corpora, such as CommonVoice, had been utilized in training these ASRs. Consequently, we opted to create a small, high-quality dataset sourced from a collection of text files for the evaluation process. We deliberately selected informal Persian text,[5] as it likely contained fewer words familiar to the models. This approach served as a more rigorous test, evaluating the models' ability to accurately transcribe phonemes in audio files from new domains and thus show a low CER.

To collect this dataset, we followed two approaches, resulting in the VirgoolInformal and GPT-Informal datasets.

VirgoolInformal: To gather text for this dataset, we created a tool that distinguishes between formal and informal writing styles. Using this module, we then crawled the Persian blog post website, Virgool (vir), and gathered a set of informal text.[6] A subset of the collected text files was then recorded by a female speaker in a silent environment through different sessions.

The raw files of the dataset comprise 25 pairs of audio and text files from 25 informal-text posts. The total duration of the audio files is approximately 5.63 hours, with a vocabulary of 6840 unique words. The dataset is segmented into smaller audio and text chunks ranging from 2 to 12 seconds, encompassing up to about 24 words each.

GPTInformal: To further diversify the subjects of the dataset text, we prompted GPT-4o (OpenAI, 2023) to generate long texts on various topics in

[5] Persian language is spoken with slight variations between formal contexts and everyday use.

[6] The code for informal text detection is available at https://github.com/MahtaFetrat/Persian-Informal-Text-Detector

informal Persian. A female native speaker then recorded these texts in a quiet environment. This dataset covers a variety of subjects and includes over 6 hours of corresponding speech. It is published under the CC-0 1.0 license.[7]

B.2 Dataset Processing

We utilized the pre-processing component as same as our ManaTTS dataset preparation pipeline to obtain clean pairs of audio and text files. Given that the audio files in this dataset were meticulously recorded from the crawled text files without alteration, they remain an exact match. Consequently, there was no necessity for the start-end alignment process.

This precise correspondence also enables the use of the lighter forced alignment tool, Aeneas. The Aeneas forced alignment tool requires the text files to be split into sentences and then attempts to align the audio with the provided sentences. Therefore, we needed a sentence tokenization tool for the Persian language.

B.2.1 Sentence Tokenization Method

The most widely recognized and commonly used tool for this purpose in Persian is the Hazm sentence tokenizer. However, this tool primarily tokenizes based on punctuation, which can result in some very long sentences if the original text is not well-punctuated, a common occurrence in informal text. To address this issue, we integrate a part-of-speech (POS) model into Hazm tokenizer to get a customized sentence tokenization module. This module considers multiple criteria to split the text into more coherent and independent sentences.

The sentence splitting module requires an input minimum and maximum sentence length, along with the input text string. It utilizes the Hazm sentence tokenizer to segment the text into sentences, primarily separated by punctuation marks. Subsequently, it iterates through these sentences, dividing any that exceed the specified maximum length. Conversely, the minimum parameter is employed to avoid excessively short sentence fragments.

To achieve a meaningful split, this module employs the Perpos (Bashari, 2021) POS model to identify verbs within the text. Subsequently, it divides the string around these identified verb positions. Notably, it includes any adjacent symbols and the conjunction word '/vɑː/' (meaning 'and' in English) in the split with the verb. This is because symbols can influence the verb's intonation, and the word '/vɑː/' following the verb is typically phonetically integrated with it and pronounced as '/əʊ/'.

The Perpos model is also utilized to identify "Ezafe" tags. Words marked with this tag are pronounced in a manner that is linked to the preceding words. Therefore, it is not advisable to split sentences when encountering this tag, as it may result in the audio being interrupted in the middle of the vowel phoneme /e/. To address this consideration during sentence splitting, the module treats a word and all its subsequent Ezafe-tagged words as a single word group while iterating over the text tokens. The complete code for the processing steps of this dataset, including the sentence tokenization module, is publicly available.[8]

B.3 Evaluation

The audio-text chunks of VirgoolInformal dataset were employed to evaluate and compare the Persian ASR models. Each audio chunk underwent processing through all the ASR models, and the resulting transcripts were recorded. Following a lightweight text processing to eliminate irrelevant symbols and characters from both the hypothesis transcripts and the ground truth text, the CER between these two strings was computed. Subsequently, the average CER of each model on the dataset chunks was taken into account as the performance criterion (See the first column of Table 5).

It's noteworthy that while some ASR models encountered issues with truncated transcripts, they exhibited high-quality transcripts in other instances. Additionally, the transcription module effectively filters out truncated transcripts, alleviating concerns in this regard. These observations led us to first filter out truncated transcripts by excluding those with less than 80% of the length of the ground truth text. Subsequently, we calculated the average CER of the ASR models based on the remaining outputs. This metric offers insights into the quality of output transcripts independent of truncation issues. The results of this evaluation approach are presented in the second column of Table 5.

The first column of Table 5 also illustrates the ranking of ASR models' reliability utilized in the alignment tools, determined by evaluation results and initial experimental findings.

[7] https://huggingface.co/datasets/MahtaFetrat/GPTInformal-Persian

[8] https://github.com/MahtaFetrat/VirgoolInformal-Speech-Dataset

Table 5: Evaluation results of ASR models based on all output transcripts and transcripts after filtering out truncated instances.

ASR Model	Average CER (All Transcripts)	Average CER (Filtered Transcripts)
Vosk	0.1128	**0.1005**
Wav2vec-v3	**0.1090**	0.1053
Wav2vec-fa	0.1411	0.1147
Whisper-fa	0.1701	0.1616
Hezar	0.3703	0.3715

Considering metrics from the second column of Table 5 as the criterion to sort ASR models based on their reliability yields surprising results. As mentioned in the main body of the paper, utilizing the Vosk model as the most reliable ASR resulted in 71.46% of the chunks being accepted with a HIGH-quality match to the ground truth text. In contrast, if we selected Wav2vec-v3 as the most reliable ASR because of its smaller CER across all the transcripts, this ratio would reduce to 55.23%. This observation shows that the non-truncated transcripts from the Vosk model were a better match to the ground truth texts, and the second metric reflects this superiority better.

C Transcript Module Statistics

As detailed in the Transcription Module section, the typical process for aligning text with an audio chunk's transcript unfolds as follows:

1. The audio chunk undergoes processing by all ASR models, yielding a list of transcripts.

2. Any flawed transcripts, such as those exhibiting repetitive patterns, are identified and eliminated using regular expressions.

3. The longest transcript is singled out, and any transcripts shorter than 80% of its length are discarded.

4. The remaining transcripts are arranged in order of the reliability rates assigned to the ASR models during evaluation.

5. Sequentially, the transcripts undergo search algorithms until the earliest one meets the designated CER thresholds, at which point the process halts.

This method naturally leans towards utilizing the most reliable ASR for aligning audio chunks with text. As anticipated, the bulk of the chunks (96.46%, equating to 62542 chunks) are accepted by the Vosk transcript form. Figure 7 illustrates the acceptance ratios of the ASR models, with Vosk contributing the most and Whisper-fa the least.

It's intriguing to observe the effectiveness of the multiple ASR using scheme. This can be explored in two aspects, corresponding to the two strategies employed in the transcription module. Firstly, through majority voting on transcript length, and secondly, by attempting sequential transcript matching until a suitable fit is found.

Our primary focus lies in assessing the effectiveness of the length majority voting technique in recovering from truncated transcript errors. Upon analyzing our processed data chunks, we observed that transcripts generated by Vosk were excluded from analysis in 1646 audio segments due to their insufficient word count, possibly indicating an error in this particular ASR system. Notably, the truncation error was even more pronounced in less reliable ASRs like Whisper-fa; however, the majority voting technique effectively mitigated its impact. For a visual representation of the number and proportion of rejected transcripts due to the length filter, please refer to Figure 8.

Next, our aim is to assess whether alignment with alternative transcripts contributed to some audio chunks being successfully matched. Table 6 presents the number of transcripts that underwent the matching process until an audio chunk was successfully matched with the ground truth text. It's also worth noting that in 646 of the audio chunks, the transcript from Vosk couldn't be matched to the ground truth text, but it was matched by the transcript from subsequent ASRs.

Our transcription module employs a Majority of Experts (MoE) technique for forced alignment using multiple non-perfect ASR models to mitigate their individual errors. In one experiment, we aimed to assess the robustness of this forced alignment tool by determining how much error of the ASRs it could tolerate.

Before discussing the experiment, it's important

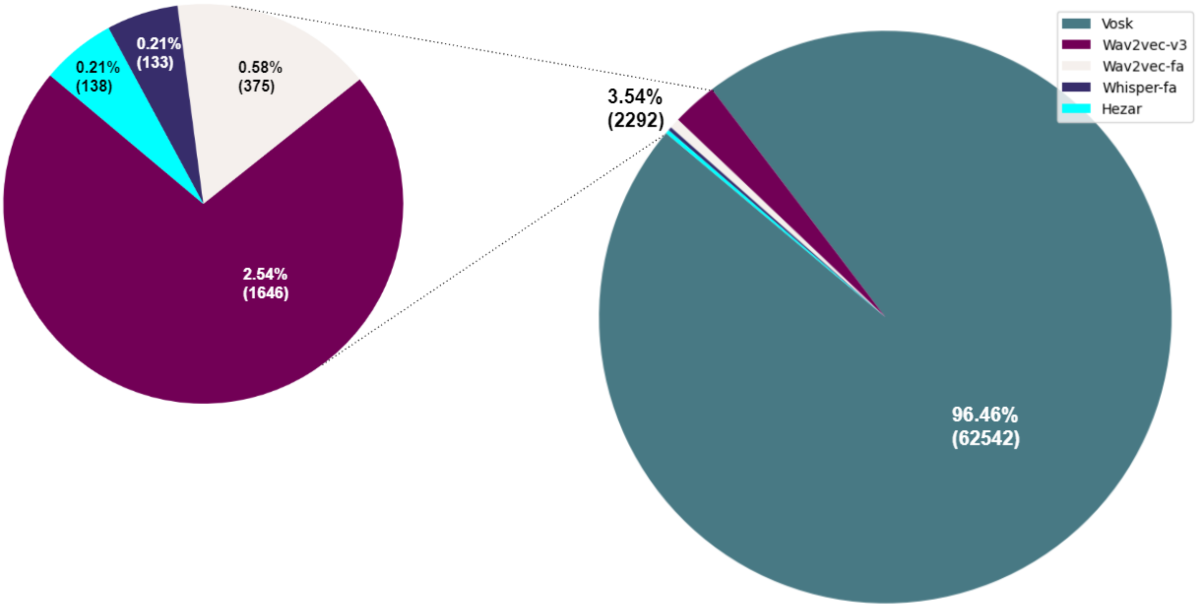


Figure 7: **Acceptance Ratio by ASR Model.** The values in parentheses represent the exact number of chunks.

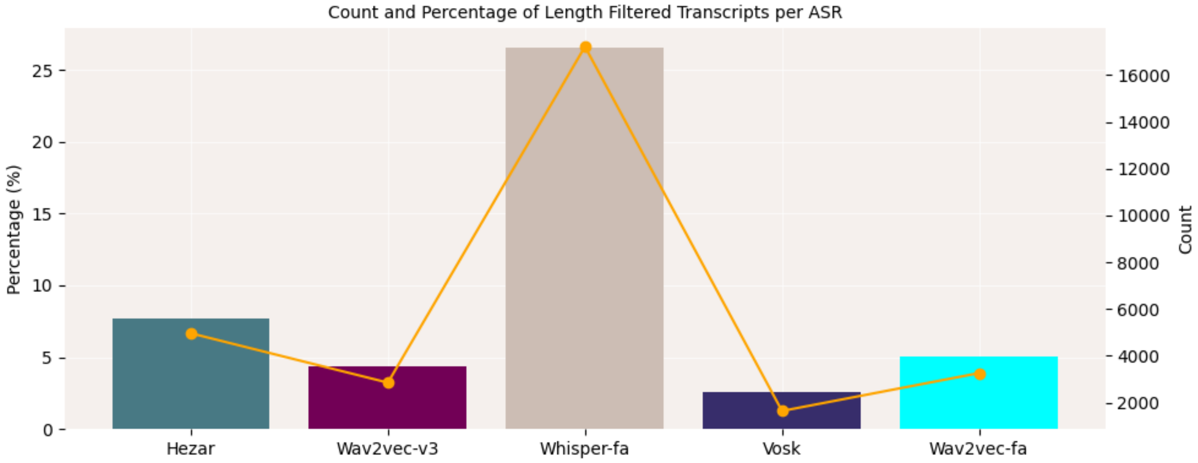


Figure 8: Distribution of transcripts filtered out due to inadequate length.

to note that regardless of ASR weaknesses, the quality of audio-text chunks from the pipeline remains high. This is because chunks are only accepted if they meet strict CER thresholds, ensuring they uphold a high-quality standard. The primary impact of weaker ASRs is on the number of accepted chunks, not their quality. As ASR errors increase, their transcripts become less similar to the ground truth, resulting in fewer chunks passing the CER thresholds.

To evaluate this, we introduced artificial errors into the ASR outputs, randomly flipping characters in the transcripts. The error rates were uniformly chosen from the ranges $[0, 0.1]$, $[0, 0.2]$, $[0, 0.3]$, $[0, 0.4]$, and $[0, 0.5]$. This produced average CER increases of approximately 5%, 10%, 15%, 20%, and 25% respectively. These were substantial increases, especially considering the ASRs already had baseline CERs of 10-30% on the VirgoolInformal dataset.

Using these modified ASRs, we performed forced alignment on an audio file that had previously been segmented into 151 chunks without any rejections. The results were impressive, demonstrating the resilience of the MoE approach to ASR degradation. The number of rejected chunks for each error level, as shown in Table 7, highlights this robustness.

Table 6: Distribution of transcripts processed to match individual audio chunks. The table shows the number of chunks aligned using different numbers of transcripts. For example, 64119 chunks were aligned using the transcript from a single ASR, 592 chunks required processing into the second transcript, and so forth.

Number of Chunks	Number of Processed Transcripts
1	64119
2	592
3	74
4	39
5	10

Table 7: Number of rejected chunks at different ASR error levels. The columns represent the range of additional error introduced into the ASR outputs, while the rows compare the performance of our method (which combines multiple ASRs) with that of using a single ASR. The numbers in the cells indicate how many out of 151 chunks were rejected, with lower values indicating greater robustness and effectiveness.

	$\mathbf{[0, 0.1]}$	$\mathbf{[0, 0.2]}$	$\mathbf{[0, 0.3]}$	$\mathbf{[0, 0.4]}$	$\mathbf{[0, 0.5]}$
Our Method (Multiple ASRs)	0	2	1	3	9
Single ASR	1	4	17	53	65

D TTS Model Evaluation Details

In this section, we elaborate on the method used to evaluate the TTS model trained on the MansTTS dataset.

D.1 MOS Score

As described in Section 5.1.2, the MOS test was conducted using five utterances, with the order of the models shuffled for each utterance to minimize bias. The resulting MOS scores for each utterance and source are presented in Table 8, along with the position of each source in the playback sequence.

The MOS across all samples and utterances, along with their variability, are presented in Table 2 and visualized in Figure 9. The standard deviation of the scores was calculated using the numpy std function on the aggregated scores from all samples of the five utterances. The primary sources of variation in the scores are as follows:

- Specific sources may appear more natural in some utterances and perform worse in others.
- Subjects have varying understandings and expectations regarding the naturalness of a speech sample.
- The random shuffling of the sources' order in each utterance affects the scores given by subjects due to the relative naturalness of the different sources.

In addition to examining the model's overall performance, it is insightful to analyze the distribution of Mean Opinion Score (MOS) ratings given to our model by individual subjects. This provides valuable insight into how opinions varied among respondents regarding our model. Figure 10 presents this distribution, shedding light on how the average scores assigned to our model are spread among respondents.

D.2 Objective Scores

In addition to the subjective MOS score, we conducted a more comprehensive evaluation of the trained TTS model using several objective methods. We selected a subset of 100 audio and text chunks, generating audio from these text chunks using our TTS model and two baseline models (VITS and Glow). Additionally, we regenerated the audio from their spectrograms using the vocoder employed by our TTS system. The evaluation metrics included PESQ (Rix et al., 2001), MCD (Kubichek, 1993), and APTD (Average Predicted Time Difference in seconds). The results of these metrics are presented in Table 9.

We also evaluated intelligibility using two ASR models: 1) Google Speech Recognition API (Zhang, 2017), and 2) Vosk. The same 100 randomly selected audio chunks generated by the TTS models and HiFi-GAN vocoder were transcribed by these ASR systems. We computed the Character Error Rate (CER) by comparing the ASR-generated

Table 8: Subjective assessment of outcomes of the different speech sources per utterance. **GT Spec** refers to the utterances with ground truth spectrograms but HiFi-GAN-synthesized waveforms, and **GT Waveform** refers to the natural speech samples.

		VITS	**Glow**	**Ours**	**GT Spec**	**GT Waveform**
utterance 1	**position**	3	2	4	1	5
	MOS	1.78	1.44	4.24	3.69	**4.42**
utterance 2	**position**	5	2	1	4	3
	MOS	1.65	1.26	**3.10**	2.96	3.03
utterance 3	**position**	3	2	4	5	1
	MOS	1.34	1.30	3.92	**4.20**	3.96
utterance 4	**position**	4	3	1	5	2
	MOS	2.03	1.39	3.48	**4.16**	4.05
utterance 5	**position**	4	1	5	3	2
	MOS	1.61	1.32	4.08	4.25	**4.57**

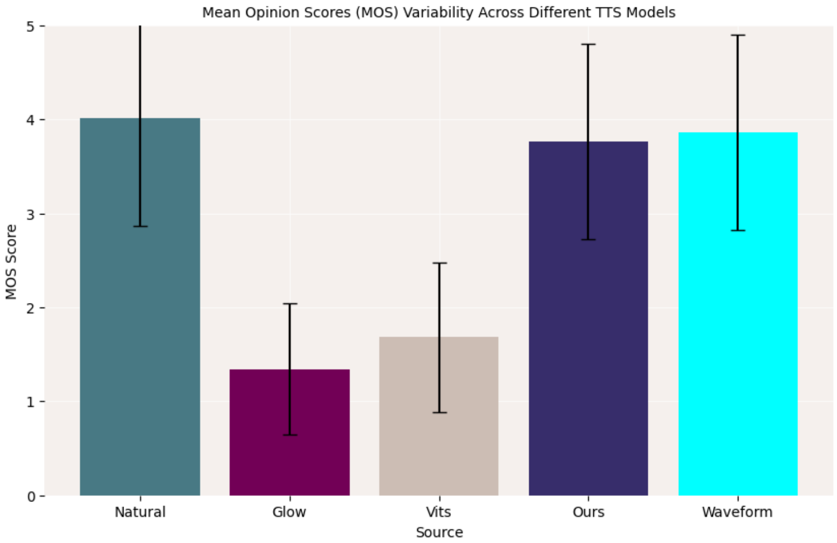


Figure 9: MOS of different sources and their variability.

transcripts with the ground truth transcripts. Additionally, we computed the CER for the ground truth audio to account for the inherent error rates of the ASR models. The results are summarized in Table 10.

E Supplementary Figures

Figure 11 shows the flowchart of the forced alignment algorithm used in our processing pipeline.

F Supplementary Tables

Table 11 summarizes the tools used in our data processing pipelines.

G Datasheet Questions

In this section we present the datasheet for the Man-aTTS dataset, adhering to the guidelines outlined by Gebru et al. (Gebru et al., 2021).

G.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

For what purpose was the dataset created? Was there a specific task in mind? Was there a

Table 9: Objective assessment of outcomes of the TTS models. **GT Spec** refers to the utterances with ground truth spectrograms but HiFi-GAN-synthesized waveforms.

	Baseline 1 (Vits)	**Baseline 2 (Glow)**	**Ours**	**GT Spec**
PESQ	1.05	1.06	1.11	2.89
APTD	1.4185	0.2781	0.5783	0.0064
MCD	15.1611	18.5218	18.5682	7.1069

Table 10: Average CER of outcomes of TTS models. **GT Spec** refers to the utterances with ground truth spectrograms but HiFi-GAN-synthesized waveforms, and **GT Waveform** refers to the natural speech samples.

	Baseline 1 (Vits)	**Baseline 2 (Glow)**	**Ours**	**GT Spec**	**GT Waveform**
Google API	0.1259	0.2325	0.0956	0.0533	0.0482
Vosk	0.2095	0.2762	0.1506	0.1406	0.1372

Table 11: The list of tools used in the dataset preparation code; all with open source licenses.

Tool Name	**Usage**	**Repository Page**	**License**
Spleeter (2020)	Source separation (remove background music)	Github	MIT
Parsi.io (2023)	Number extraction & number to text conversion	Github	Apache-2.0
Hazm (2024)	Text normalization	Github	MIT
Pydub (2022)	Silence detection/removal	Github	MIT
Perpos (2021)	Part of speech tagging for sentence tokenization See appendix.	Github	MIT
Vosk (2024)	Forced alignment	Github	Apache-2.0
Whisper-fa (2023; 2021)	Forced alignment	HuggingFace	Apache-2.0
Wav2vec2-v3 (2021)	Forced alignment	HuggingFace	-
Wav2vec2-fa (2021)	Forced alignment	Github	Apache-3.0
Hezar (2023)	Forced alignment	Github	Apache-2.0
JiWER (2024)	CER calculation	Github	Apache-2.0

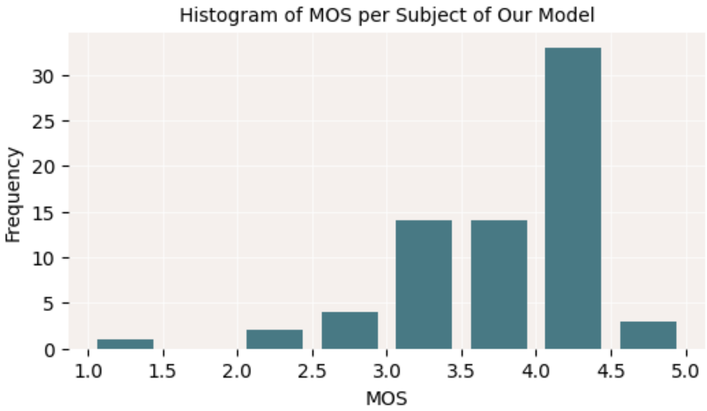


Figure 10: Distribution of Mean Opinion Score (MOS) ratings given to our model by individual subjects.

specific gap that needed to be filled? Please provide a description.

ANS: The dataset was developed to address the scarcity of open-source datasets and models for speech tasks, particularly text-to-speech (TTS), in Persian. We hope that these resources will foster the development of open tools, improving accessibility for Persian-speaking individuals, including the visually impaired.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

ANS: The dataset was created by the speech processing team of the Data Science and Machine Learning (DML) Laboratory at Sharif University of Technology.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

ANS: The dataset creation received no external funding and is provided free of charge.

Any other comments?

ANS: No.

G.2 Composition

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

ANS: The instances in the dataset consist of pairs of small audio and transcript chunks, derived from larger audio and text files that approximately match each other.

How many instances are there in total (of each type, if appropriate)?

ANS: There are a total of 64,834 pairs of audio-text files in the dataset.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

ANS: The dataset is derived from the archive of the Nasl-e-Mana magazine (NasleMana) and includes all issues published up to the date of this paper's submission. However, the dataset doesn't contain all the content from these issues, as a small proportion (about 2%) of the data is discarded. This is because we only accept audio and text chunks that meet a specified level of quality. As outlined

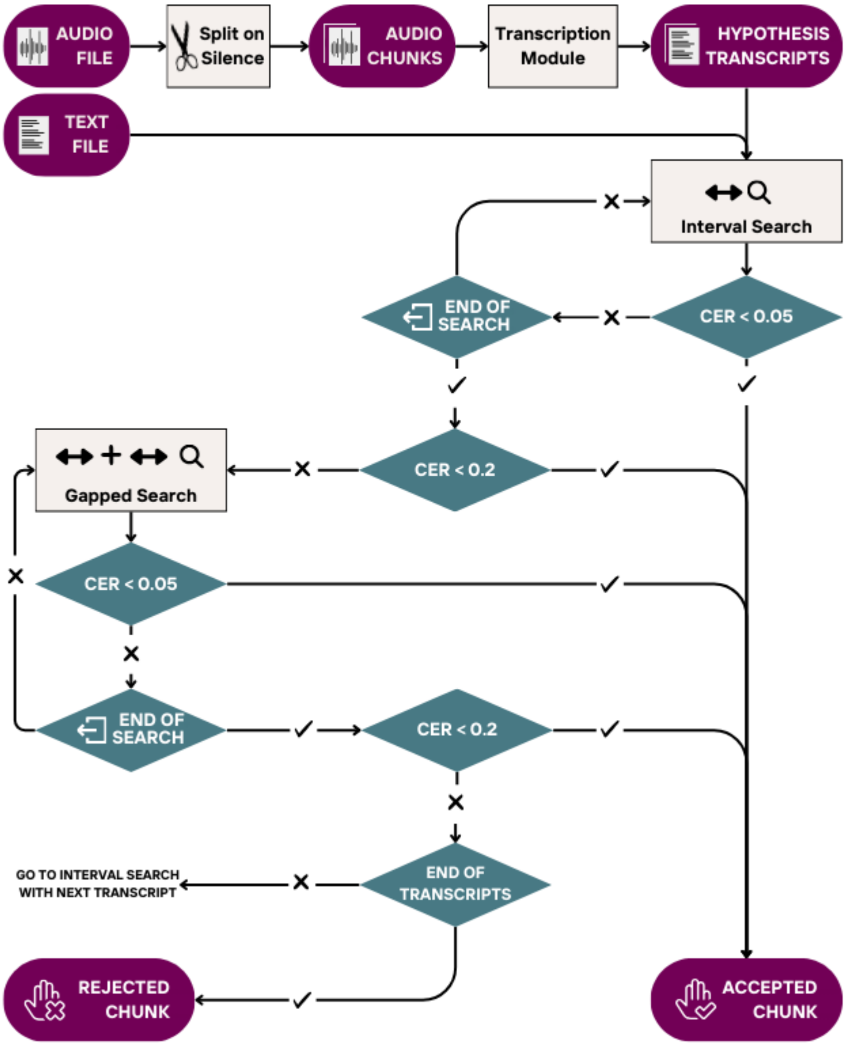


Figure 11: Flowchart of the forced alignment algorithm.

in the original paper, mismatches between audio and text typically arise due to three reasons: 1) English words in the text that cannot be matched with their Persian spoken form in the transcript; 2) Specifically formatted numbers or symbols with differing spoken and written forms; and 3) Underlying mismatches in the original audio and text files stemming from factors such as censorship, speaker mistakes, different text versions, etc. Therefore, while the dataset provides a representative sample of the Nasl-e-Mana magazine archive, it may not include all issues, and the selection process ensures quality and relevance to the research objectives.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

ANS: Each instance comprises an audio file, typically lasting a few seconds, paired with a corresponding text file containing a transcript of the audio.

Is there a label or target associated with each instance? If so, please provide a description.

ANS: Yes, depending on the dataset's purpose, either the transcripts or the audio files serve as the label. For instance, in a text-to-speech (TTS) task, the text files are the input, and the audio files act as labels.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

ANS: No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social

network links)? If so, please describe how these relationships are made explicit.

ANS: No.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

ANS: We don't recommend a particular data split for this dataset. The samples are uniform, with no overlap or redundant audio. Therefore, users can make an arbitrary split that suits their needs.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

ANS: The dataset's transcripts are extracted from ground truth text files using hypothesis transcripts from ASR systems, aiming for approximate matches. A match is determined by comparing the Character Error Rate (CER) of the hypothesis to the selected part of the ground truth text, with matches below a specified threshold considered acceptable. Although a transcription module utilizing various ASR models was implemented to enhance confidence in the transcripts, this process isn't entirely error-proof. Unpredictable ASR errors and the insensitivity of CER to minor discrepancies may still lead to inaccuracies in the transcripts. However, the dataset provides confidence score (in term of the CER) for each hypothesis transcript with the corresponding ground truth transcript for every audio-text pair. Users have the flexibility to filter the data to include only pairs with CER values below a desired threshold. It's worth noting that the thresholds in the processing pipeline were carefully selected to ensure that all accepted chunks meet an acceptable quality standard.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

ANS: The dataset is self-contained and doesn't rely on external resources.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

ANS: The dataset does not contain confidential or personal data as it is derived from the public magazine of Nasl-e-Mana, dedicated to the blind community.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

ANS: The dataset originates from the trusted Nasl-e-Mana magazine, dedicated to the blind community, which also has a certificate from the Integrated Media System of the country.[9] While there isn't an explicit mechanism to check for offensive content, we believe that the source material does not contain such data.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

ANS: The dataset does not identify any subpopulations; it consists entirely of audio from a single speaker.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

ANS: The dataset includes audio from a single speaker, allowing for potential identification based on their voice.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or

[9] https://e-rasaneh.ir/Certificate/89184

health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

ANS: The dataset is not considered to contain sensitive data, as it is sourced from the publicly online free available Nasl-e-Mana magazine, which is believed not to include such information.

Any other comments?

ANS: No.

G.3 Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners to create alternative datasets with similar characteristics.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

ANS: The raw data, including pairs of approximately matching audio and text files, was downloaded from the Nasl-e-Mana magazine website (naslemana.com).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

ANS: The majority of the data was collected using a crawling script to download content from the Nasl-e-Mana magazine website (naslemana.com). However, a subset of issues hosted on the ibngo.ir domain couldn't be accessed via automated requests and were therefore downloaded manually.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

ANS: The dataset is not a sample from a larger set. It encompasses all issues of the Nasl-e-Mana magazine archive to date that could be matched with the text files using our forced alignment method. We do not claim that it is representative of all audio-transcript pairs available on the internet.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

ANS: No human subjects were involved in the data collection process, and no compensation was provided.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

ANS: The raw data was collected from 41 issues of the Nasl-e-Mana magazine over a span of more than three years. The first issue was published on January 19, 2021, and the most recent issue was released on May 20, 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

ANS: No ethical review processes were conducted.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

ANS: The data was obtained via third parties.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

ANS: Yes, we communicated directly with the Nasl-e-Mana magazine owners, and they provided their consent to publish the dataset publicly, requesting a formal letter. Figure 12 shows a screenshot of this letter.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

ANS: Consent was provided through a formal response from the data owners. Figure 13 shows the response received from them.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

ANS: While formal consent mechanisms were not established, ongoing communication with the data owners allows for consent revocation. If the data owners request it, we will remove the processed material from the public domain.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

ANS: No such analysis has been conducted.

Any other comments?

ANS: No.

G.4 Preprocessing/cleaning/labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

ANS: Yes, preprocessing was conducted on the data. Initially, the audio and text files were cleaned and standardized. Subsequently, the data underwent alignment and forced alignment, resulting in fine-grained audio-text pairs. Finally, postprocessing was performed to remove any unnecessary parts of the audio. For more detailed information, please refer to the paper.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

ANS: No, the raw data was not saved separately. It remains accessible on the Nasl-e-Mana website through the provided crawling scripts.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

ANS: Yes, all the code/scripts used for crawling to processing are publicly available under the MIT license. You can access them via this Repository.

Any other comments?

ANS: No.

G.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

Has the dataset been used for any tasks already? If so, please provide a description.

ANS: Yes, it has been employed to train a Text-to-Speech (TTS) model in our research, which is used to evaluate data efficiency.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

ANS: No there isn't.

What (other) tasks could the dataset be used for?

ANS: The dataset is well-suited to meet the demand for high-quality, open, and large-scale data for Persian TTS model development. It also has the potential to be used in various tasks (with some adjustments) that require speech and transcript pairs.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

ANS: We do not believe that the dataset carries such risks.

Are there tasks for which the dataset should not be used? If so, please provide a description.

ANS: The dataset should not be utilized for replicating or imitating the speaker's voice for malicious purposes or unethical activities, including voice cloning for malicious intent.

Any other comments?

ANS: No.

G.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

ANS: Yes, the dataset is available to the public under a CC-0 license.

How will the dataset be distributed (e.g., tarball on website, API,)? Does the dataset have a digital object identifier (DOI)?

ANS: The dataset is distributed through the link provided in the paper. The dataset does have a DOI in that link.

When will the dataset be distributed?

ANS: The dataset is already publicly available and can be accessed via the link provided in the paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

ANS: The dataset is shared under the CC-0 license, allowing free use, but it prohibits harmful activities like voice cloning for malicious purposes.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

ANS: No, there are no IP-based or other restrictions imposed on the data associated with the instances.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

ANS: No, there are no export controls or other regulatory restrictions applicable to the dataset or individual instances.

Any other comments?

ANS: No.

G.7 Maintenance

The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

Who will be supporting/hosting/maintaining the dataset?

ANS: The dataset is stored on a public data repository (Huggingface) and maintained by the authors for updates.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

ANS: You can contact the authors via the following email addresses:

- Mahta Fetrat: m.fetrat@sharif.edu
- Zahra Dehghanian: zahra.dehghanian97@sharif.edu
- Hamid R. Rabiee: rabiee@sharif.edu

Is there an erratum? If so, please provide a link or other access point.

ANS: There is currently no erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list,)?

ANS: We do not have a formal update plan yet.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

ANS: There are no retention limits specified for the dataset.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

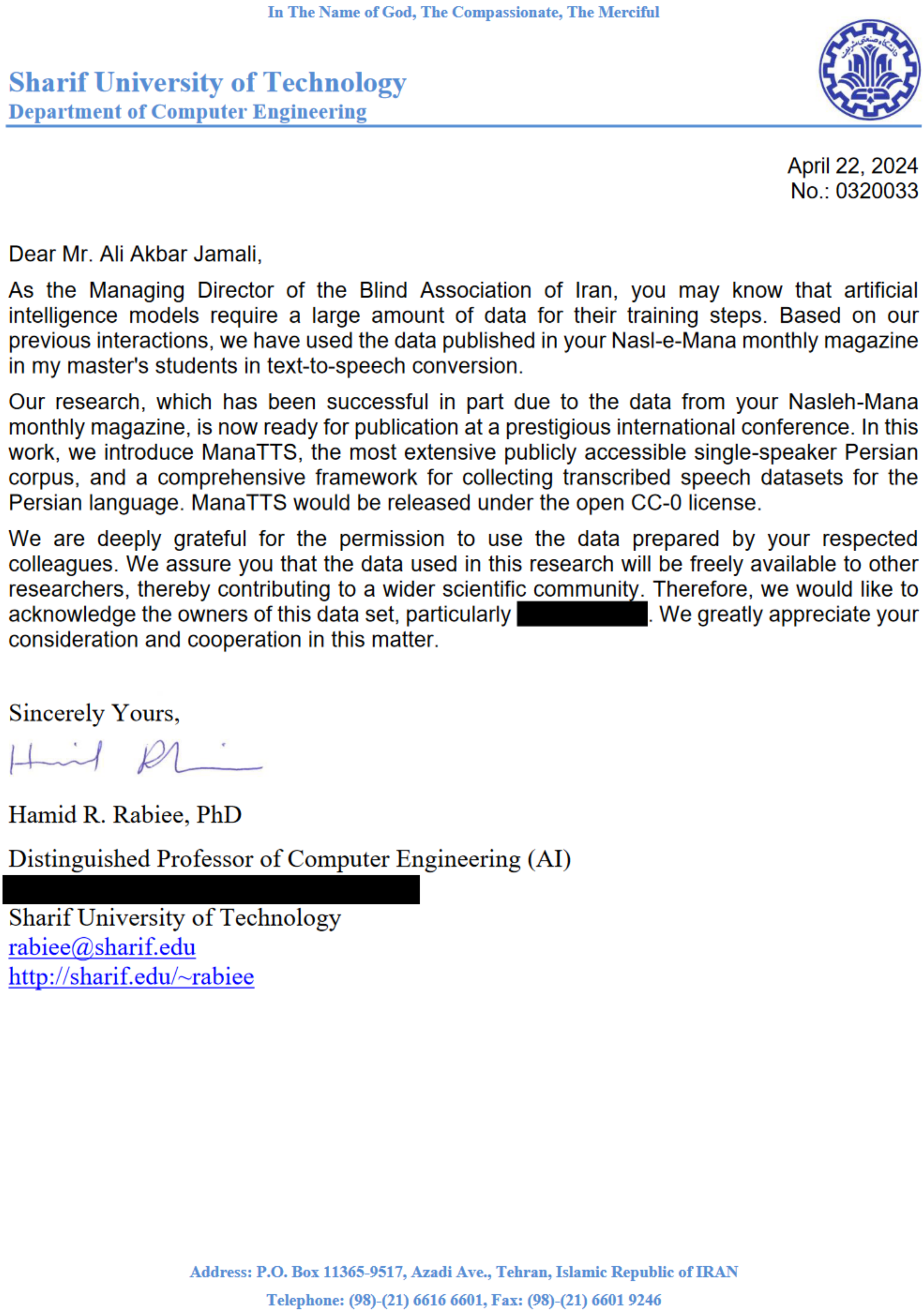
ANS: We do not have plans to support or maintain older versions of the dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

ANS: Contributions are very welcome. Contributors can open issues or submit pull requests on , or contact the authors directly for error reports or improvements. We plan to address serious issues and explicit improvements directly, while leaving the validation of other enhancements to the community.

Any other comments?

ANS: No.



In The Name of God, The Compassionate, The Merciful

Sharif University of Technology
Department of Computer Engineering

April 22, 2024
No.: 0320033

Dear Mr. Ali Akbar Jamali,

As the Managing Director of the Blind Association of Iran, you may know that artificial intelligence models require a large amount of data for their training steps. Based on our previous interactions, we have used the data published in your Nasl-e-Mana monthly magazine in my master's students in text-to-speech conversion.

Our research, which has been successful in part due to the data from your Nasleh-Mana monthly magazine, is now ready for publication at a prestigious international conference. In this work, we introduce ManaTTS, the most extensive publicly accessible single-speaker Persian corpus, and a comprehensive framework for collecting transcribed speech datasets for the Persian language. ManaTTS would be released under the open CC-0 license.

We are deeply grateful for the permission to use the data prepared by your respected colleagues. We assure you that the data used in this research will be freely available to other researchers, thereby contributing to a wider scientific community. Therefore, we would like to acknowledge the owners of this data set, particularly [redacted]. We greatly appreciate your consideration and cooperation in this matter.

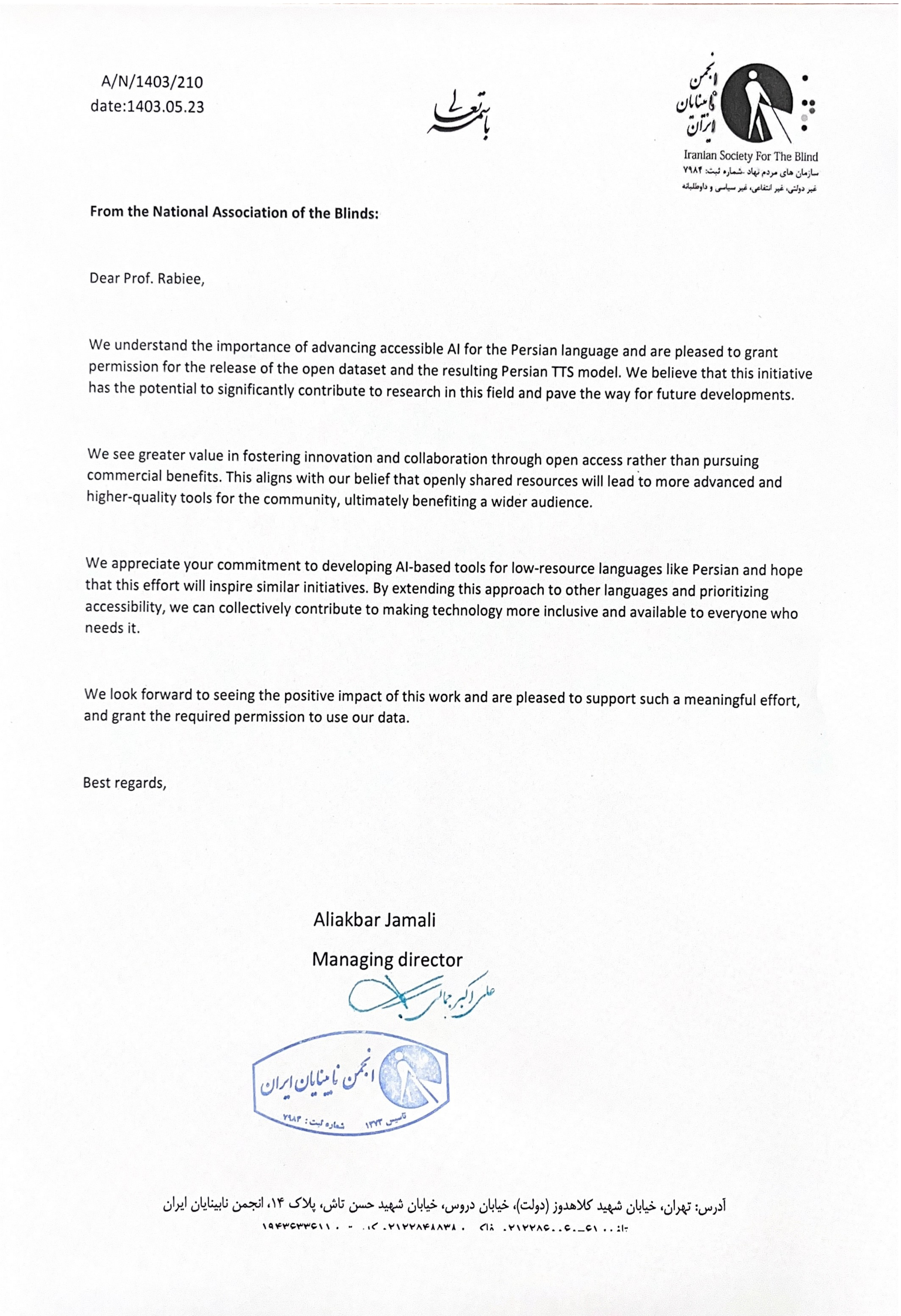
Sincerely Yours,

Hamid R. Rabiee, PhD

Distinguished Professor of Computer Engineering (AI)
[redacted]
Sharif University of Technology
rabiee@sharif.edu
http://sharif.edu/~rabiee

Address: P.O. Box 11365-9517, Azadi Ave., Tehran, Islamic Republic of IRAN
Telephone: (98)-(21) 6616 6601, Fax: (98)-(21) 6601 9246

Figure 12: Screenshot of the formal letter informing Nasl-e-Mana magazine owners about the data collection and public release of the dataset. The speaker's name is hidden for more privacy.



A/N/1403/210
date:1403.05.23

باسمه تعالی

انجمن نابینایان ایران
Iranian Society For The Blind
سازمان های مردم نهاد -شماره ثبت: ۷۹۸۴
غیر دولتی، غیر انتفاعی، غیر سیاسی و داوطلبانه

From the National Association of the Blinds:

Dear Prof. Rabiee,

We understand the importance of advancing accessible AI for the Persian language and are pleased to grant permission for the release of the open dataset and the resulting Persian TTS model. We believe that this initiative has the potential to significantly contribute to research in this field and pave the way for future developments.

We see greater value in fostering innovation and collaboration through open access rather than pursuing commercial benefits. This aligns with our belief that openly shared resources will lead to more advanced and higher-quality tools for the community, ultimately benefiting a wider audience.

We appreciate your commitment to developing AI-based tools for low-resource languages like Persian and hope that this effort will inspire similar initiatives. By extending this approach to other languages and prioritizing accessibility, we can collectively contribute to making technology more inclusive and available to everyone who needs it.

We look forward to seeing the positive impact of this work and are pleased to support such a meaningful effort, and grant the required permission to use our data.

Best regards,

Aliakbar Jamali

Managing director

علی اکبر جمالی

انجمن نابینایان ایران
تاسیس ۱۳۷۳ شماره ثبت: ۷۹۸۴

آدرس: تهران، خیابان شهید کلاهدوز (دولت)، خیابان دروس، خیابان شهید حسن تاش، پلاک ۱۴، انجمن نابینایان ایران
تلفن: ۰۲۱۲۲۸۶۰۰۶۰_۶۱ فکس: ۰۲۱۲۲۸۴۸۸۳۸ کد پستی: ۱۹۴۳۶۳۳۶۱۱

Figure 13: Screenshot of the formal response from Nasl-e-Mana magazine owners about the public release of the dataset.