Fine-tuning Whisper with Spontaneous Persian Speech (SPS)

Behnoosh Namdarzadeh, Nicolas Ballier
Université Paris Cité, ALTAE
Place Paul Ricoeur, F-75013 Paris, France
behnooshnamdar@gmail.com, nicolas.ballier@u-paris.fr

Abstract

This paper introduces the Spontaneous Persian Speech (SPS) dataset designed for automatic speech recognition (ASR) tasks and a methodology laying the groundwork for addressing the shortage of spontaneous speech data. The corpus aims to support research on natural and conversational Persian, which remains under-represented in current ASR resources. The dataset consists of 694 minutes of audio from a total of 65 speakers, including 34 male and 31 female speakers. It contains 526,585 tokens. The audio segmentation step produces intervals of 1.24 to 3.25 seconds, each containing 3 to 9 words. The recordings cover a variety of environments, from inside cars to homes and shopping areas, including both busy and quiet settings. We use the SPS dataset to fine-tune Whisper and the performance increases significantly for both the small and medium models based on Word Error Rate (WER). This could be an initiative toward building domain-oriented datasets for specific ASR tasks.

Keywords: Automatic Speech Recognition, Persian, Language Resource, OpenAI's Whisper

1. Introduction

Automatic Speech Recognition (ASR) systems have shown improvements in recent years, largely driven by the availability of large-scale, high-quality datasets. However, for Persian, existing corpora are mainly limited to read speech, which fails to capture the variability and disfluencies of natural, spontaneous conversation (Namdarzadeh et al., 2022). To address this gap, we introduce the Spontaneous Persian Speech (SPS) dataset, a spontaneous Persian speech dataset, derived from YouTube and exclusively intended for research use. The dataset aims to model an ASR model for spontaneous speech in Persian which would handle challenging features of spontaneous speech like hesitations, overlapping speech, and spontaneous-like discourse markers (Cihan et al., 2025; Tree and Schrock, 1999; Adda-Decker et al., 2008).

2. Review of Resources for Spoken Persian

This section highlights the main existing speech datasets for Persian and their characteristics and why we need more natural spoken datasets to have more robust ASR systems for real speech scenarios.

Persian is considered as an under-resourced language (Freihat and Abbas, 2021; Taghizadeh and Faili, 2016). Languages are classified as under-resourced when they lack the quantity of data necessary for training statistical and machine learning models (Liu et al., 2022). Furthermore, Persian is diglossic and resources are even rarer for informal spoken data (Kabiri et al., 2022). This can pose challenges in various fields like Neural Machine Translation (NMT) and ASR. Notably, although many corpora include both written and spoken data, written data usually dominates the training sets (Namdarzadeh and Ballier, 2022). As a result, NLP models often make more errors when handling spoken-language structures.

FARSDAT is the first Persian speech database, comprising read speech from 304 native speakers of diverse ages, genders, dialects, and educational backgrounds (Bijankhan et al., 1994). The male-to-female ratio is approximately 2:1 and each participant is to read sentences which were composed using 1,000 most common Persian words, sourced from newspapers. Followingly, **TFARSDAT** is the telephone Persian speech database, which contains approximately 8 hours of telephone speech, comprising 25,000 tokens. This material includes both spontaneous and read speech between 77 female and 125 male participants of varying ages, genders, education levels, and dialects (Bijankhan et al., 2003). The other attempt is **Mozilla Common Voice**, a large-scale multilingual collection of transcribed speech corpora, which is mostly read speech as well (Ardila et al., 2020). This free, downloadable language resource includes approximately 350 hours of Persian speech data from 4,100 speakers.[1] Nevertheless, most of the existing Persian speech corpora rely heavily on read speech. This means that the linguistic characteristics that are concerned with spoken data may be ignored. In the newly introduced dataset, we aim to bridge this gap and try to make ASR systems more robust and reliable in

[1] https://voice.mozilla.org/fa/datasets

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 263–269
11–12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

spontaneous and real speech settings.

3. Introducing the Spontaneous Persian Speech (SPS) Dataset

This section explains our contribution to the creation of a spontaneous Persian speech dataset and describes the pipeline for this process, which leverages the ASR model, OpenAI's Whisper (Radford et al., 2023), to create the dataset. Whisper was chosen for its multilingual performance, robustness to real-world audio variation, and ability to work well with minimal training (Radford et al., 2023). Its open-source availability, and easy deployment also make it ideal for reproducible experiments. Whisper is easy to reproduce and fine-tune for detailed analysis, which is why we chose it to build the baseline for our dataset. To ensure transparency and reproducibility while enabling model inspection and adaptation, we prioritized open-source solutions. Among them, Whisper is a well-validated, high-performance choice for Persian ASR and other low-resource languages like Urdu (Sedghiyeh et al., 2025; Sehar et al., 2025).

It is true that collecting spontaneous speech requires energy, time, and supervision. Thus, this project aims to collect spontaneous data from YouTubers who have been recorded and published on YouTube in various settings. This has already been done for languages like English (Coats, 2019), Turkish (Safaya and Erzin, 2022), and multi-lingual datasets (Valk and Alumäe, 2021; Li et al., 2024). YouTube videos may have some spoken-related features that could not be found in telephone conversations, especially when people know that they are being recorded (Labov, 1972). This means that we do not have noise, interruptions, or other speech-related features that are useful in real ASR scenarios, which we try to include in the SPS dataset. The selected videos mainly consist of informal conversations between two or at most three people. The topics are primarily related to entertainment, and the recording settings of the videos range from studio environments with high-quality microphones to cars, supermarkets, and other busy places. [2]

We present our semi-automatic creation of the SPS dataset as an initiative to build more data which are close to speech in real time for Persian. The SPS dataset is created by leveraging an audio Large Language Model (LLM), Whisper (Radford et al., 2023). The aim is to build a linguistic corpus in a context of technological changes, replicating what has been done for Hebrew (Marmor et al., 2023) and Mandarin (Sun et al., 2024). The audio data were transcribed using OpenAI's Whisper. Transcriptions follow orthographic conventions, with special attention to disfluencies, fillers, and etc. Each file includes metadata such as duration, gender, and etc..

Figure 1 illustrates the complete pipeline used to create the SPS dataset using Whisper. We first extracted the audio files from YouTube videos. The criteria for selecting the videos are as follows: (a) 2 to 3 speakers spontaneously interacting; (b) balanced representation of female and male speakers; (c) preference for the accent, spoken in Tehran (the capital of Iran); (d) an informal interaction setting to maximize informal lexical choices and youth slang; (e) speakers aged 20 to 30 years; and (f) the selection of indoor and outdoor settings to have the variety of background sounds and noises.

The process involves feeding an audio file into the system, which automatically detects the language and proceeds with transcription, starting from a small model and progressing to the largest available model (large-v3 for Persian). Whisper takes raw waveform data as an input, which is then processed by an encoder-decoder Transformer to directly generate the corresponding text transcription. The encoder maps the audio file to a sequence of embeddings, while the decoder attends to these embeddings and generates the text token by token. The resulting transcriptions are produced in various formats such as .txt, .srt, .vtt. We used the.srt files generated by Whisper, as they include the start and end times of each utterance. We then converted the .srt files into TextGrid format using the `pysrt` Python library [3] to facilitate readability and correction in Praat (Boersma and Weenink, 2025). The correction process involved adjusting timestamps and correcting the orthography of the sentences, along with the punctuations, which has been done by a Persian native speaker. The segmentation of the chunks which is based on human judgment, along with the xmin and xmax values of the .wav files are provided in .TextGrid format. This can be considered a step forward in fine-tuning Whisper to potentially achieve better results for ASR tasks which will be discussed in Section 6.

[2] The SPS dataset was compiled from publicly accessible YouTube videos strictly for internal, non-commercial research use, and it will not be publicly released, redistributed, or shared, unless we receive consent attestations from the owners of the YouTube channels. Personal data is not really involved in the conversations, but speakers do use their first names and their family names so that anonymization would be required for full public disclosure of the dataset. Ethically, we contacted the owners of the channels, but we have not received responses from them yet.

[3] https://pypi.org/project/pysrt/

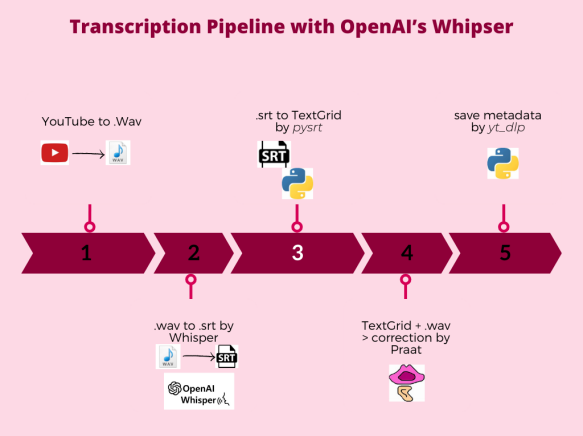

Figure 1: Transcription Pipeline using Whisper

4. Characteristics of Spontaneous Persian Speech

This section explains the characteristics of the SPS dataset. The data collected comprises 694 minutes of speech (approximately 11 hours) and 526,585 tokens. Although this is not a large dataset compared to (Saraji et al., 2025) and (Marmor et al., 2023), it is based on authentic and fluid interaction between dyads with real linguistic data. The backgrounds of the videos are diverse, ranging from quiet locations such as studios to noisy places environments such as cars and shopping areas. Among the videos selected for the construction of this dataset, there are 34 male speakers and 31 female speakers. All speakers are native Persian speakers with a Tehrani accent, which is the standard accent spoken in Tehran, the capital of Iran.

The average interval duration ranges from 1.24 s to 3.25 s. Each segment contains between 3.16 and 8.77 words, with an overall average of 5.01 words per interval across all the files. The total number of tokens across all TextGrid files is 526,585. The dataset exhibits an average lexical diversity (TTR) of 0.352, meaning that roughly one in three word tokens corresponds to a distinct word type. This reflects a relatively rich and varied vocabulary across the collected speech samples.

5. Error Analysis

This section presents a comprehensive examination of the errors generated by Whisper's audio language model in 70 minutes of the SPS dataset. The findings underscore the need to incorporate spontaneous speech data to ensure the robustness and generalizability of the performance of the audio models when it comes to spontaneous scenarios. The transcription performance was evaluated using Word Error Rate (WER). The WER quantifies errors in the hypothesis (insertions, deletions, substitutions) and is expressed as a percentage of errors normalized by the reference transcript's word count (N) (Levenshtein, 1966).

The Whisper models achieved the WER of 88.72%, 83.10%, and 49.04%, respectively, for small, medium, and large models. This indicates a considerable level of transcription error, including substitutions, deletions, and insertions, particularly at the word level, and highlighting the challenges of transcribing spontaneous Persian speech accurately.

The errors produced by Whisper (Radford et al., 2023), when handling spontaneous speech for a language with low resources like Persian, indicate the absence of an informal register in the training data of the models. Data collection is usually initiated in the most accessible and cost-effective way, usually meaning the use of written resources from the internet. This supports the findings on the limitations of available resources for spoken linguistic features in neural machine translation models as well (Namdarzadeh et al., 2022). Thus, we observed that when the test register for one model, (in our example, the informal register used to test Whisper) includes unique challenges like noise, disfluencies, interruptions, and repetitions, the transcriptions of the ASR systems deteriorate. We found a wide variety of errors at the phonological level, as indicated in Table 1. The errors at the phonological level are mostly related to sounds with the same place and manner of articulation, which are wrongly captured by Whisper, whether due to noise or the speakers' fast articulation in the informal register.

Gold word	Gold IPA	Whisper Word	Whisper IPA	Notes
کلید	kelid	کلیت	kelit	/d/→/t/, final consonant change
واژه	vaʒe	واجر	vadʒer	/ʒ/→/dʒ/, vowel + final consonant change
رو	ro	–	–	word omitted
دزدیدم	dozdidam	دستیده	dastide	/o/→/a/, /z/→/s/, /d/→/t/

Table 1: phoneme-level alignment between gold transcription and Whisper transcriptions (large model)

The different outputs generated by the small, medium, large models of Whisper differ significantly. We classify them into three distinct categories:

- **Effects of Multilinguality on Model Output:** Corresponding errors can be attributed to the multilingual nature of the Whisper model. This means that we observed subtokens not only from English, but also from Russian, Chinese, Urdu, and other languages. This is the price to pay for multilingual models, as there is always a risk of data leakage from other languages in the output. Although we did not examine all subtokens from other languages present in

Whisper's training data, by analyzing a subset of them, we assume that their phonological features are similar to those expected in Persian. A more subtle form of hallucination can be observed when the subtokens predicted by Whisper for Persian data do belong to the Perso-Arabic script, but correspond to Arabic, not to Persian. This is probably explained by the discrepancy of the size of the training data (24 hours for Persian vs. 739 hours for Arabic for the transcription task).

- **Different prosodic boundaries:** There are notable discrepancies in the prosodic boundaries encoded by the small Whisper model. The scope of what Whisper calls "segment" (the utterance) varies across models, as already noted for English (Ballier et al., 2023). The utterances differ in terms of the number of tokens, and this variation does not appear to be systematic throughout the 70 minutes of the audio file. In contrast, the medium and large models produce more utterance-like and consistent token counts. It appears that larger Whisper models more effectively account for the prosodic boundaries of the audio, which results in shorter utterances.
- **Linguistic errors:** The different models for the whisper multilingual version share the same inventory of subtokens, as the same multilingual tokenizer was used, whatever the number of parameters of that model. Nevertheless, the acoustic mapping to the subtoken varies across models. The tiny model fails to produce any results and the small model transcription consists of random combinations of subtokens from different languages, aligned with the acoustic characteristics of the audio file. Due to the presence of nonsensical words and a large number of subtokens from other languages, the transcription remains unreadable. Another issue contributing to the lack of readability is hallucination. While this phenomenon is present in the small model, it is, perhaps surprisingly, more frequent in the medium Whisper model for one long sound file (80 minutes). In the latter case, from approximately the midpoint of the audio file onward, the transcriptions consist of repeated occurrences of a single word with no meaningful interpretation in Persian.

6. Fine-tune Whisper with the SPS Dataset

This section explains the fine-tuning procedures and results of using the SPS dataset. Fine-tuning Whisper can be used with the aim of increasing the performance of the model for under-resourced languages (Gete et al., 2025; Sehar et al., 2025), or it may help for domain-specific purposes in order to increase the performance of the ASR system (Pawlowicz et al., 2025). Basically, the aim of fine-tuning is to target the error-prone features that we need to improve the system. In our study, we used the SPS dataset to better predict transcriptions when the text is in informal Persian and to fine-tune the models. In simple terms, it is to add new data to the model that has already been created to see if there are any improvements.

We explain the data preparation process for fine-tuning and then present the results of fine-tuning using SPS.

- **Audio standardization** The audio files need to be resampled to 16 kHz, so the conversion of all the audio files is of high importance. converting stereo to mono (average channels) is obligatory as well.
- **Segmentation** This aligns with having a training item which is short, intelligible, and aligned. We keep the segments typically 2–15 seconds long. We also filter for min duration $\geq$ 1s and max duration $\leq$ 15–20 s. We make sure that there is non-empty text in the data, as well as no very short/long segments.
- **Persian text normalization** This is one of the crucial pre-processing steps in fine-tuning for Persian. NFKC is applied for Unicode normalization. The other important normalization in this regard is to replace Arabic ك by Persian character ک. Diacritics are omitted, and whitespace is also normalized.

We used 80 % of the SPS to train the Whisper model, then we used 10 % for validation. We picked the best epoch and used it on the test set (10 % of the SPS) and compared it to the baseline small and medium models.

Table 2: Baseline vs. fine-tuned Whisper small and medium models on the test split

Model	Mean WER ↓	Mean CER ↓
Whisper-small (baseline)	2.5767	2.6269
Whisper-small (fine-tuned, 2 epochs)	0.8046	0.5513
Whisper-medium (baseline)	2.3679	1.8567
Whisper-medium (fine-tuned, 2 epochs)	0.7675	0.5061

Table 2 shows that the WER and CER of the test set decrease significantly with the addition of only 11 hours of spontaneous data. A detailed insight into the improvement of the transcriptions predicted by the fine-tuned model is not within the scope of the current work, but we could suggest that fine-tuned models rarely produce hallucinations (Guerreiro et al., 2023) in the transcriptions.

7. Discussion

This section discusses the results of the fine-tuning experiment with the SPS dataset in order to enrich the model with the characteristics of spoken language. In comparison to other spoken-oriented datasets, the SPS dataset features fluid and authentic Tehrani Persian. Having observed issues in the transcription of spontaneous data in Section 5, enriching the systems with such high-quality data may improve the outputs. The use of this type of data appears to be rare, and fine-tuning systems may enhance the performance.

Unlike previous Persian speech datasets (Bijankhan et al., 2003; Saraji et al., 2025), SPS implements a 1:1 male-to-female speaker ratio. Since there are significant gender biases in ASR models, incorporating such data for fine-tuning could impact performance (Harris et al., 2024; Zanon Boito et al., 2022) and help neutralize gender bias in the outputs.

Although relatively small, SPS represents nearly half of the Persian training data (24 hours) used for Whisper (Radford et al., 2023). Incorporating SPS into the training data increases existing Persian data and potentially improves the results of ASR models. This means that the integration of high-quality data, even if it is small, may increase the performance of the models and make them more robust to different linguistic registers.

Dialect variation could be one of the additional challenges to be encountered in such systems. Thus, target-oriented data collection for the special purpose could be a way to enhance the performance of the systems. In spontaneous interactions between Tehrani speakers, we observed that the large Whisper model cannot handle them very well based on the reported WER in Section 5. The SPS dataset focuses exclusively on the Tehrani dialect, intentionally excluding other varieties. However, this methodology can be replicated for other accents and dialects spoken across Iran, allowing for future comparative linguistic analyses.

8. Conclusion and Further Research

This paper presented the creation of the SPS dataset as an initiative to expand the language resources of an under-resourced language, Persian. We have shown the limitations of current systems trained with insufficient spontaneous spoken data and that are more dependent on written data for ASR tasks. This is the reason why the models cannot capture spoken-related characteristics such as repetitions, delays, laughter, false starts, and disfluencies.

Our pipeline for the creation of the dataset leverages the Whisper (Radford et al., 2023) model to speed up the production of resources in a human-in-the-loop perspective, since we experienced the need for transcription correction. Using synthetic data for models has been shown to increase the risks (Nadăș et al., 2025) in training models (insufficient stylistic realism, bias amplification, etc.). The SPS dataset exhibits good quality and features spontaneous conversational exchanges involving a mix of genders and variety of topics, all of which are valuable characteristics for a linguistic dataset. Register plays a crucial role in building this dataset, which is why the collection process emphasizes an efficient and time-effective approach to gathering authentic and simultaneous linguistic data.

To test the performance of the SPS dataset, we fine-tuned the small and medium Whisper models, and with 2 epochs of learning from the SPS dataset, the baseline small model's WER mean changes from 2.57 to 0.80 and the medium model WER mean changes from 2.36 to 0.76, which is a significant improvement. This indicates that a small, but domain-oriented dataset could improve the performance of ASR systems. In this study, for Persian, with only the addition of half (11 hours) of the training data of Whisper (24 hours), the evaluation metrics showed improvement in the transcription task. This could be extended to different domains, such as fine-tuning Whisper with specialized domains and different dialects in order to have better performance based on the target. We chose Whisper as a strong open-source multilingual ASR model with proven robustness and competitive performance for Persian and other low-resource languages, while remaining reproducible and easy to fine-tune. Alternative open-source and commercial systems were considered, but either underperformed or lacked transparency for control (ASR et al., 2025), released after our experiments, are promising, and we leave their evaluation on Persian ASR for future work. The code for transcribing audio using Whisper is available on GitHub[4]. Depending on the training data for the language of choice, one could begin building a dataset for that language. Fine-tuning is also reproducible, and the relevant code— inspired by https://huggingface.co/blog/fine-tune-whisper.

Acknowledgements

We thank the three anonymous LREC reviewers for their comments on a previous version of this paper. This publication has emanated from research supported in part by a research equipment grant from the Scientific Platforms and Equipment Committee (PAPTAN project) under ANR Grant Number ANR-18-IDEX-0001 (Financement IdEx Université de Paris).

[4]https://github.com/openai/whisper

9. Bibliographical References

Martine Adda-Decker, Claude Barras, Gilles Adda, Patrick Paroubek, Philippe Boula de Mareüil, and Benoit Habert. 2008. Annotation and analysis of overlapping speech in political interviews. In *LREC 2008*, pages 3105–3111.

Nicolas Ballier, Behnoosh Namdarzadeh, Maria Zimina, and Jean-Baptiste Yunès. 2023. Translating dislocations or parentheticals: Investigating the role of prosodic boundaries for spoken language translation of French into English. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 119–131.

Paul Boersma and David Weenink. 2025. Praat: doing phonetics by computer. https://www.praat.org. Computer program.

Ruixing Cihan, Zhang Xiangyu, Mehmet Emre, Veronica Lavanya, Rong Ethan, Emmanuel Sanjeev, and Leibny Paola Garcia. 2025. Casper: A large scale spontaneous speech dataset. *arXiv preprint arXiv:2506.00267v1*.

Steven Coats. 2019. A Corpus of Regional American Language from YouTube. In *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference (DHN 2019)*, pages 79–91, Copenhagen, Denmark.

Abed Alhakim Freihat and Mourad Abbas, editors. 2021. *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Association for Computational Linguistics, Trento, Italy.

Dawit Ketema Gete, Bedru Yimam Ahmed, Tadesse Destaw Belay, Yohannes Ayana Ejigu, Sukairaj Hafiz Imam, Alemu Belay Tessema, Mohammed Oumer Adem, Tadesse Amare Belay, Robert Geislinger, Umma Aliyu Musa, Martin Semmann, Shamsuddeen Hassan Muhammad, Henning Schreiber, and Seid Muhie Yimam. 2025. Whispering in Amharic: Fine-tuning Whisper for Low-resource Language.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166–15184, Miami, Florida, USA. Association for Computational Linguistics.

Roya Kabiri, Simin Karimi, and Mihai Surdeanu. 2022. Informal Persian Universal Dependency treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7096–7105, Marseille, France. European Language Resources Association.

William Labov. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Song Li, Yongbin You, Xuezhi Wang, Zhengkun Tian, Ke Ding, and Guanglu Wan. 2024. Msr-86k: An evolving, multilingual corpus with 86,300 hours of transcribed audio for speech recognition research. pages 1245–1249.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944. Association for Computational Linguistics.

Yanir Marmor, Kinneret Misgav, and Yair Lifshitz. 2023. ivrit.ai: A Comprehensive Dataset of Hebrew Speech for AI Research and Development. ArXiv2307.08720.

Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, pages 134615–134633.

Behnoosh Namdarzadeh and Nicolas Ballier. 2022. The neural machine translation of dislocations. *ExLing 2022*, 28:127–131.

Behnoosh Namdarzadeh, Nicolas Ballier, Lichao Zhu, Guillaume Wisniewski, and Jean-Baptiste Yunès. 2022. Toward a test set of dislocations in Persian for neural machine translation. In *Proceedings of the Third International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2022) co-located with ICNLSP 2022*, pages 14–21, Trento, Italy. Association for Computational Linguistics.

Daniel Pawlowicz, Jule Weber, and Claudia Dukino. 2025. Effectiveness of Whisper's Fine-Tuning for Domain-Specific Use Cases in the

Industry. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025)*, pages 1398–1405. SCITEPRESS – Science and Technology Publications, Lda.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Ali Safaya and Engin Erzin. 2022. Experiments on Turkish ASR with Self-Supervised Speech Representation Learning. ArXiv2210.07323.

Mohammad Azim Saraji, Abdullah Khalili, and Ahmad Hatam. 2025. PAZHVAK: a Word-Level Farsi Speech Corpus by University of Hormozgan. *SN Computer Science*, 6(7):865.

Nima Sedghiyeh, Sara Sadeghi, Reza Khodadadi, Farzin Kashani, Omid Aghdaei, Somayeh Rahimi, and Mohammad Sadegh Safari. 2025. PSRB: A Comprehensive Benchmark for Evaluating Persian ASR Systems. *arXiv preprint arXiv:2505.21230*.

Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking Whisper for Low-Resource Speech Recognition: An N-Shot Evaluation on Pashto, Punjabi, and Urdu. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 202–207, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Jingyi Sun, Yaru Wu, Nicolas Audibert, and Martine Adda-Decker. 2024. Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé. In *Actes de JEP-TALN-RÉCITAL 2024 (35 Journées d'Études sur la Parole)*, pages 291–300, Toulouse, France. Actes des 35 Journées d'Études sur la Parole.

Nasrin Taghizadeh and Heshaam Faili. 2016. Automatic wordnet development for low-resource languages using cross-lingual WSD. *Journal of Artificial Intelligence Research*, 56:61–87.

Jean E Fox Tree and Josef C Schrock. 1999. Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language*, 40(2):280–295.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève. 2022. A study of gender impact in self-supervised models for speech-to-text systems. In *Proceedings of Interspeech 2022*, pages 1278–1282, Incheon, Korea.

10. Language Resource References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4218–4222, Marseille, France. European Language Resources Association.

Team Omnilingual ASR, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages.

Mahmood Bijankhan, Javad Sheikhzadegan, and Mahmood R Roohani. 1994. FARSDAT- The speech database of Farsi spoken language. In *Proceedings of the Australian Conference on Speech Science and Technology*, volume 2, pages 826–830.

Mahmood Bijankhan, Javad Sheykhzadegan, Mahmood R Roohani, Rahman Zarrintare, Seyyed Z Ghasemi, and Mohammad E Ghasedi. 2003. Tfarsdat-the telephone Farsi speech database. In *Proc. Eurospeech 2003*, pages 1525–1528.