TaLK-Corpus: A Regionally Diverse Evaluation Set for Sri Lankan Tamil Speech

Adsajan Thillainathan[1] Nishanthini Kanthakumar[1] Nivethiga Rasan[2] Kengatharaiyer Sarveswaran[1]

[1]Department of Computer Science, University of Jaffna
[2]Department of Linguistics and English, University of Jaffna
{2021sp146, 2020csc026, rnivethiga, sarves}@univ.jfn.ac.lk

Abstract

This paper introduces the TaLK Corpus, the first speech benchmark corpus for Sri Lankan Tamil Automatic Speech Recognition (ASR) covering speech from 22 administrative districts of Sri Lanka. The corpus contains 1 hour and 33 minutes of speech from 22 native speakers (one per district) and includes rich metadata on demographics, location history, recording conditions, and domain information, along with transcriptions in Tamil script and the International Phonetic Alphabet (IPA). Standardised preprocessing (16 kHz mono WAV format) and segmentation using Silero Voice Activity Detection (VAD) resulted in 1,214 utterances. All recordings were manually transcribed by trained linguists, and MD5-based file naming used to ensure data integrity and consistency. TaLK corpus enables district-wise benchmarking of ASR systems and supports dialect-sensitive evaluation. We establish baseline results for multilingual models (Whisper Large-V3 and Facebook's MMS) in zero-shot settings. The evaluation reveals substantial performance disparities across districts, highlighting the impact of regional phonological variation in low-resource Sri Lankan Tamil. Although Whisper Large-V3 outperforms MMS overall, it shows considerable variability, with mean Word Error Rates ranging from 0.672 to 0.903 across districts. These findings demonstrate strong regional effects even within a single model. By releasing TaLK-Corpus under the CC-BY-NC 4.0 licence, we aim to support dialect-robust ASR research and foster inclusive speech technologies for Sri Lankan Tamil-speaking communities.

Keywords: Sri Lankan Tamil, Speech Corpus, Dialectal Variation, Low-Resource Languages, ASR Benchmark Dataset

1. Introduction

Speech processing has become a critical area of research as voice interfaces power an increasing number of applications, from virtual assistants to real-time transcription and accessibility tools (Rakotomalala et al., 2021). In this context, Automatic Speech Recognition (ASR), which is the conversion of spoken language into text, and Text-to-Speech (TTS) have gained prominence. Transformer-based models such as Whisper (Radford et al., 2023), achieve high accuracy in controlled settings for high-resource languages. However, persistent challenges limit their robustness in real-world deployment, including noise and environmental variability, speaker and accent variability, multilingualism and code-switching, data scarcity in low-resource languages, and spontaneous or conversational speech (Ahlawat et al., 2025).

Dialectal processing also presents challenges for ASR. Dialectal variations introduce phonetic mismatches (e.g., vowel shifts, consonant changes), lexical differences, prosodic variations, and morphosyntactic deviations from standard training data (Palivela et al., 2025). Models trained primarily on mainstream varieties exhibit systematic biases, leading to substantially higher Word Error Rates (WER) for regional or minority dialects compared to standard forms. In low-resource languages, these issues are exacerbated by limited dialect-specific data, lack of standardized orthographies, and the absence of diverse training corpora, hindering the development of inclusive, robust systems (Dhasmana et al., 2026).

Benchmark datasets play a central role in advancing Automatic Speech Recognition (ASR) by enabling standardized and reproducible evaluation of model robustness across linguistic and acoustic variability.

Existing Tamil ASR benchmarks, (Mozilla Foundation, 2019; OpenSLR, 2019; Bharathi et al., 2022, 2024), mainly focus on Indian Tamil or demographic diversity without structured geographic coverage in Sri Lanka. While Sri Lankan resources like EmoTa (Thevakumar et al., 2025) exist, they target emotion recognition rather than ASR and do not provide district-level representation, leaving a gap for geographically structured Sri Lankan Tamil speech datasets.

To address this gap, we introduce a regionally di-

"TaLK" is inspired by the IETF BCP 47 language tag ("ta-LK") for Sri Lankan Tamil, which corresponds to "ta_LK" in the Unicode CLDR locale format.

Proceedings of SPEAKABLE @ LREC 2026, pages 194–200

verse evaluation set for Sri Lankan Tamil speech, called TaLK-Corpus[2], covering speakers from all 22 districts. The dataset is explicitly designed as a benchmarking resource, enabling reproducible district-wise evaluation of ASR systems. By incorporating Tamil script, Roman transliteration, and broad IPA annotation, TaLK supports conventional WER-based comparison, establishing a standardized benchmark for Sri Lankan Tamil dialectal ASR research.

This paper makes two main contributions. First, we present a carefully manually curated speech corpus for Sri Lankan Tamil, comprising 1 hour and 33 minutes of high-quality, geographically representative audio–text pairs with verbatim transcriptions. Second, we establish a zero-shot performance baseline for this dialect by evaluating two widely used multilingual ASR models—Whisper V3 and Facebook's Massively Multilingual Speech (MMS).

2. Background and Motivation

Tamil is one of the world's oldest living classical languages, belonging to the Dravidian family, with a rich literary tradition spanning over 2,000 years (Newbigin, 2019). It is spoken by more than 86 million people (Zeidan, 2020), primarily in Tamil Nadu (India), Sri Lanka, Singapore, Malaysia, and diaspora communities, and holds official status in several regions. Despite its cultural and demographic significance, Tamil is a low-resource language in modern NLP and speech technologies (Sarveswaran et al., 2021), suffering from limited large-scale annotated digital corpora compared to high-resource languages like English. Key linguistic characteristics include agglutinative morphology, a rich phonemic inventory (notably retroflex consonants and vowel length distinctions) (Jain and Bhowmick, 2025), syllable-timed prosody (Thinakaran et al., 2025), diglossia (distinct literary and colloquial varieties), and frequent code-switching (Prasanna and Arora, 2024), especially with English.

Sri Lankan Tamil speech exhibits notable regional variation due to historical settlement patterns, prolonged contact with Sinhala and English, and population movements (Unjum et al., 2026; Yasmini, 2017). These factors lead to differences in pronunciation, accent, vocabulary, and prosody across districts and even within them. While Tamil ASR has advanced through corpora dominated by Indian Tamil (e.g., Common Voice Tamil, IISc-MILE) (Mozilla Foundation, 2019; OpenSLR, 2019), Sri Lankan-specific resources remain scarce and often task-specific, such as EmoTa for emotion recognition (Thevakumar et al., 2025), limiting the development of dialect-robust models for Sri Lankan contexts.

More importantly, there is no benchmark dataset available for Sri Lankan Tamil to evaluate the performance of ASR systems on this language variety. To address this gap, we present Version 1 of the TaLK speech corpus, a district-level Sri Lankan Tamil evaluation corpus comprising 22 speakers (one per district), with over one hour of total speech and rich metadata, annotated in the Tamil script.

3. Related Work

In Tamil ASR, several datasets have supported model development, including resources from Mozilla Common Voice Tamil (Mozilla Foundation, 2019) and OpenSLR (OpenSLR, 2019). While these corpora provide valuable training and testing material, they primarily represent Indian Tamil and largely consist of read or crowd-sourced speech, with limited structured metadata for dialect-aware benchmarking

The LT-EDI shared tasks on Speech Recognition for vulnerable Individuals in Tamil introduced evaluation datasets targeting elderly and transgender speakers in naturalistic settings (Bharathi et al., 2022, 2024; Nishanth et al., 2025). These initiatives represent an important step toward inclusive ASR benchmarking in Tamil by focusing on demographic diversity from the Indian region. Further, their design centers on speaker-group variability rather than geographically structured dialect variation, and they do not provide district-level dialectal coverage.

In the Sri Lankan context, EmoTa: A Tamil Emotional Speech Dataset (Thevakumar et al., 2025) contributed a valuable speech resource for emotion recognition research. Although important for affective computing, EmoTa is not structured as an ASR benchmark and does not support systematic evaluation of regional dialect robustness.

More broadly, dialect-focused benchmarks in multilingual settings such as IndicVoices-R (Javed et al., 2024)-have demonstrated the necessity of geographically diverse evaluation sets to measure accent and dialect sensitivity in ASR systems. These works highlight that models trained predominantly on standardised language varieties may exhibit systematic degradation when evaluated on regional speech.

4. Data Collection and Corpus Design

The dataset was collected from native Sri Lankan Tamil speakers across all 22 districts (out of 25)

[2] https://github.com/LTG-UoJ/TaLK-Corpus-Public

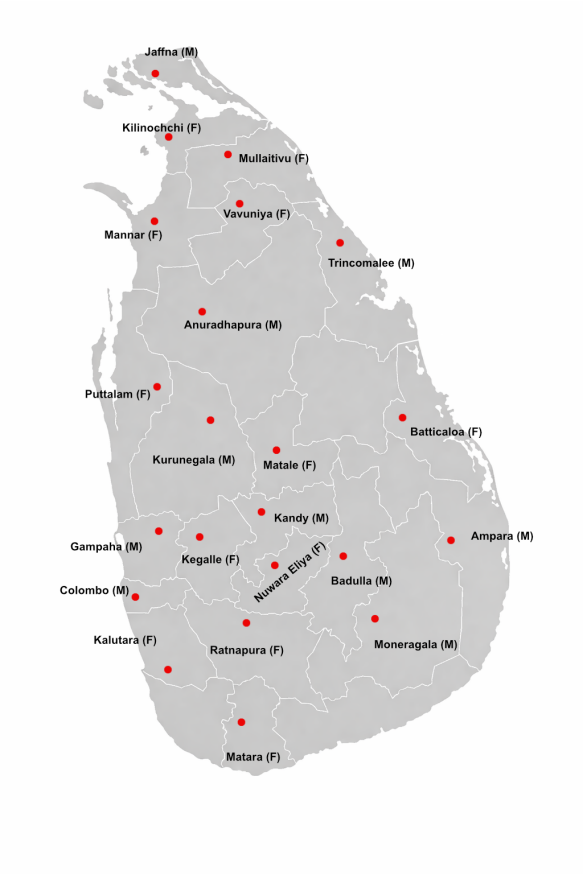

Figure 1: District-level distribution of speakers

of Sri Lanka to ensure comprehensive geographical coverage and dialectal diversity. The objective was to capture diverse variations in pronunciation, accent, and speaking style across different regions, making the corpus suitable for evaluation and diagnostic purposes in Automatic Speech Recognition (ASR) systems.

4.1. Speaker Selection

The corpus consists of 22 speakers (12 female and 10 male) representing diverse demographic backgrounds. The age distribution includes one speaker below 20 years, the majority between 20 and 60 years, and one speaker above 60 years. Participants were drawn from varied educational and occupational backgrounds to enhance representativeness for ASR development. All participants had previously sat for the GCE Ordinary Level (O/L) national examination in which mother tongue (Tamil) is an important subject.

4.2. Recording Protocol

Speech recordings were captured using the Sony PCM-A10 digital recorder and the DJI Mic Receiver connected to a mobile device. The DJI Mic features an Intelligent Noise Cancelling function, which was enabled to reduce background noise when recording with a smartphone. Recordings were conducted under controlled indoor environments with minimal background noise.

Participants were asked to speak spontaneously across multiple everyday domains, including school life, education, university experience, trips and travel, friends, family, places, festivals, funeral events, self-introduction, life experiences, and other common daily topics. These domains were selected to ensure lexical diversity and to reflect natural conversational contexts in Sri Lankan Tamil.

All recordings were standardised for ASR compatibility by converting the audio into mono-channel WAV format with a 16 kHz sampling rate. This configuration provides an optimal balance between speech quality and computational efficiency for modern speech recognition systems. The total duration of the collected speech data is approximately 1 hour and 33 minutes.

4.3. Metadata Schema

All metadata for the **TaLK** corpus is stored in a single CSV file, where each row corresponds to an audio file and its associated transcription. Speaker-level information is repeated for every utterance to ensure each row is self-contained. The metadata includes the district of birth, gender, age band, job, religion, and domain(theme). Utterance-level information includes the duration of the recording, the filename of the audio file, and its transcript in Tamil script. In addition, IPA annotations are captured for each utterance to enable further phonological analysis. For the Jaffna district, IPA transcriptions were generated using the tool called ThamizhIPA-Trans (Mahaganapathy et al., 2026) and validated by a linguist, while for other districts, annotations were manually created by the linguist. Table 1 provides a sample Metadata Entry from the TaLK Corpus. All audio files are stored in a dedicated folder, with filenames matching the entries in the CSV, allowing direct linking between metadata, audio, and transcripts. This format ensures compatibility with standard ASR processing pipelines and facilitates filtering or analysis by speaker attributes or recording characteristics.

4.4. File Naming Strategy and Anonymisation

An MD5-based hashing strategy was used to generate file names. Demographic attributes (Current Residence, Birth Place District, Age, and Gender) were combined into a unique input string and processed through MD5 so that in the future errors in file-naming can be easily tracked. A sequential

Field	Value
File name	0d2d12...9f850a_001.wav
Transcription	அங்க இருக்கும்போது...
IPA	*aŋga ɨrukkɨmbõðu...*
Birth Place (District)	Kilinochchi
Age Band	20-60
Gender	Feminine
Job	Student
Religion	Hindu
Theme	School
Duration	0:00:06

Table 1: Sample Metadata Entry from the TaLK Corpus

three-digit index was appended to maintain uniqueness across multiple files.

4.5. Ethics and Consent

All participants provided informed consent before recording, and their identities were anonymised to protect privacy. The data is licensed for research use, with restrictions preventing commercial exploitation, ensuring ethical compliance throughout data collection and usage.

5. Audio segmentation and Annotation

Speech segmentation was performed using Silero VAD v4.0[3] with a threshold of 0.5, a minimum speech duration of 250 ms, a minimum silence duration of 100 ms, a 512-sample window size, and 30 ms padding to produce clean speech segments. This process resulted in 1,214 utterances. Transcription was carried out manually by two trained linguists in dialect-preserving Tamil script and aligned to the segments. Text preprocessing and normalisation included Unicode normalisation, removal of punctuation, special characters, and extra or leading/trailing spaces. Quality control was conducted through manual checks, with double annotation considered for future validation.

6. Model benchmarking

The primary task supported by the TaLK corpus is the benchmarking of ASR systems for Sri Lankan Tamil. In addition, the detailed geographic and demographic metadata enables optional auxiliary tasks such as dialect or regional classification. In this section we report the performance of two widely used models Whisper V3 and MMS.

[3] https://github.com/aosfatos/silero-vad-v4

Model	WER	CER
Whisper v3	0.807	0.425
facebook/mms-1b-all	0.845	0.342

Table 2: ASR evaluation results on the TaLK dataset.

Model	TaLK	IndicVoices-Ta
Whisper v3	0.807	0.784
MMS	0.845	0.754

Table 3: WER comparison between TaLK (ours) and IndicVoices Tamil.

6.1. Metrics

Word Error Rate (WER) and Character Error Rate (CER) were used as the primary evaluation metrics. Scoring follows Tamil-aware tokenisation that respects the language's agglutinative morphology and script conventions. Punctuation marks and special characters are ignored during evaluation, and numerals are normalised consistently.

7. Baseline results

We present evaluation results for multilingual ASR models on Sri Lankan Tamil in zero-shot settings, assessing their performance without any Sri Lankan Tamil specific adaptation.

All experiments use the same audio preprocessing (mono, 16 kHz), normalisation, and scoring pipeline to ensure fair comparison across districts. We compute WER on the Tamil script transcripts and additionally report WER on Roman transliterations to separate orthographic effects from acoustic errors. Decoding parameters and any language-model usage are held constant across districts and documented for reproducibility.

8. Results and Analysis

We evaluate on 1,214 utterances from 1 hour 33 minutes TaLK-Corpus. Table 2 reports WER and CER for the two zero-shot baselines. Whisper V3 outperforms Facebook's MMS under this evaluation. For comparison, we also report the performance of Whisper V3 and Facebook's MMS from the IndicVoices (Javed et al., 2024) study in Table 3, which shows that the models perform poorly for Sri Lankan Tamil compare to the Indian Tamil (or the major variety included in language models).

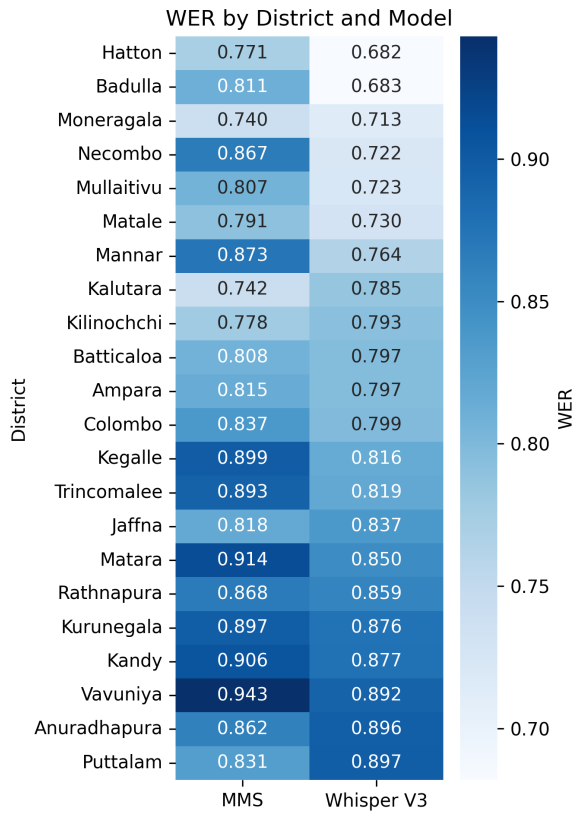

Figure 2: WER per district for MMS and Whisper V3.

Figure 2 provides a per-district breakdown. Whisper V3 shows notable variability across districts, with mean WER ranging from 0.672 to 0.903, indicating strong regional effects even under a single model. Hatton, Moneragala, and Badulla districts show comparatively lower WER for both MMS and Whisper V3. This is likely because the speech in these districts is closer to Indian Tamil, which the models were primarily trained on, so the ASR systems can recognise it more accurately than the more distinct Sri Lankan dialects in other districts.

Error analysis. Whisper V3 frequently code-switches to English for numerals and educational terms, which increases WER against Tamil-script references, for instance, as shown below, the ASR output contains code-switching and Roman script. Interestingly, the case markers present in the Tamil text are also reflected in the romanised text, but in Tamil script. For instance, கு *ku* (DAT marker) follows "lecture" in the ASR output below.

Ground truth: இப்ப இங்க வந்து பார்த்-தோம்னா இங்கிலீஷ் மீடியம் ரொம்பவே டிஃபிகல்ட் தான் என்ன சம்ரைம்ஸ் லெக்சருக்கு போனா தூங்கிருன்ட்டு வர மாதிரி தான் இருக்கும் ஏன் சொன்னா

ASR output: இங்க வந்து பார்த்தாம்-னா, **English Medium**, ரொம்பவே **Diffi-cult** தான் **Sometimes**, **lecture** கு போ-னா, தூங்கிருந்து வர மாது தான் இருக்-கும் ஏன் சொன்னாம்?.

These patterns highlight the importance of handling code-switching and numerals consistently in evaluation. The IPA annotation layer in TaLK enables future analysis of dialectal phonological patterns beyond orthographic WER.

9. Availability, Licensing, and Reproducibility

The TaLK dataset has been made publicly available to support research in dialect-aware Sri Lankan Tamil speech processing. It includes clean, segmented 16 kHz mono WAV audio files and Tamil script and IPA transcriptions. The dataset is released under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. This permits sharing, adaptation, and use for non-commercial research, specifically for evaluation, and academic purposes, provided appropriate credit is given to the creators (citation required).

10. Conclusion

We introduce the TaLK Corpus, a Sri Lankan Tamil speech corpus comprising district-level data for benchmarking ASR applications for Sri Lankan Tamil, along with rich metadata that can be used for in-depth dialectal studies and phoneme-level benchmarking. The benchmark enables reproducible, dialect-aware evaluation of ASR systems across all districts, highlighting meaningful regional variability in zero-shot performance. Our baseline results and error analysis show that current multilingual ASR systems struggle with dialectal variation and code-switching, reinforcing the need for geographically diverse evaluation and richer Sri Lankan Tamil resources. TaLK-Corpus provides a concrete starting point for more inclusive ASR research and for future expansions in speaker coverage and recording conditions. The TaLK corpus repository is available at https://github.com/LTG-UoJ/TaLK-Corpus-Public.

11. Limitations and Future Work

The current dataset is small to support model training; therefore, model evaluation was conducted in a zero-shot setting. In addition, the present version of TaLK includes only one speaker per district. Consequently, it is not yet possible to

fully disentangle district-level dialect effects from speaker-specific characteristics such as speaking rate, articulation style, or recording variability. The reported district-wise WER differences should therefore be interpreted as preliminary and indicative rather than definitive evidence of systematic dialect-level model performance differences.

WER evaluation is further affected by orthographic or spacing differences in the manually created ground truth. For example, when the model predicts "இருக்கும் போது" as two tokens while the ground truth is "இருக்கும்போது" is one token, the WER metric counts this as an error despite the prediction being linguistically correct. Such discrepancies are inherent to human annotation and can slightly inflate reported error rates, especially in low-resource languages with flexible orthographic conventions. Therefore, WER should be interpreted as indicative of overall performance rather than exact linguistic correctness.

This work is part of a broader initiative to develop a Sri Lankan Tamil speech corpus. We are currently expanding the dataset by including multiple speakers per district with balanced demographic representation, increasing the total number of recording hours through longer and more natural conversations, and incorporating diverse recording conditions (e.g., varying noise levels and channels). In addition to direct speech recordings, we plan to compile speech data from publicly available sources, such as YouTube, with appropriate consent and ethical compliance, to further enhance district-level coverage.

12. Acknowledgements

This research is part of the Sri Lankan Tamil Corpus (TaLK Corpus) project[4] at the University of Jaffna and is supported by a Google Research Scholar Award to Kengatharaiyer Sarveswaran. The authors thank Ms Sumirtha Karunakaran for her assistance with data collection. We also thank all participants and collaborators who contributed to the development of the corpus.

References

Harsh Ahlawat, Naveen Aggarwal, and Deepti Gupta. 2025. Automatic Speech Recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering*, 6:201–237.

B. Bharathi, Bharathi Raja Chakravarthi, et al. 2022. Findings of the Shared Task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*.

B. Bharathi, Bharathi Raja Chakravarthi, et al. 2024. Overview of the Third Shared Task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*.

Akriti Dhasmana, Aarohi Srivastava, and David Chiang. 2026. Dialect Matters: Cross-Lingual ASR Transfer for Low-Resource Indic Language Varieties. *arXiv preprint arXiv:2601.04373*.

P. Jain and A. Bhowmick. 2025. Comparative performance analysis of end-to-end ASR models on Indo-Aryan and Dravidian languages within India's linguistic landscape. *Journal of Audio, Speech, and Music Processing*, 10(2025).

Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vaijayanthi, Krishnan Karunganni, Pratyush Kumar, and Mitesh Khapra. 2024. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782, Bangkok, Thailand. Association for Computational Linguistics.

Ahrane Mahaganapathy, Sumirtha Karunakaran, Kavitha Navakulan, and Kengatharaiyer Sarveswaran. 2026. Bridging dialectal variation: A phonetic transcription tool for Tamil. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 234–241, Rabat, Morocco. Association for Computational Linguistics.

Mozilla Foundation. 2019. Mozilla Common Voice. https://commonvoice.mozilla.org.

JILL CR Newbigin. 2019. Evolution of Tamil Language: A Historical Study. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 6(3):115–123. © 2019 JETIR1903O19.

S. Nishanth, Shruthi Rengarajan, Burugu Rahul, and G. Jyothish Lal. 2025. NSR@LT-EDI-2025: Automatic Speech Recognition in Tamil. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*.

OpenSLR. 2019. OpenSLR: Free Speech and Language Resources. http://www.openslr.org.

[4] https://sites.google.com/univ.jfn.ac.lk/talkcorpus/home

Hemant Palivela, Meera Narvekar, David Asirvatham, Shashi Bhushan, Vinay Rishiwal, and Udit Agarwal. 2025. Code-switching asr for low-resource indic languages: A hindi-marathi case study. *IEEE Access*, 13:9171–9198.

Kabilan Prasanna and Aryaman Arora. 2024. Irumozhi: Automatically classifying diglossia in Tamil. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3096–3103.

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

Francis Rakotomalala, Hasindraibe Randriatsarafara, Hajalalaina Aimé Richard, and Ravonimanantsoa Ndaohialy Manda Vy. 2021. Voice User Interface: Literature Review, Challenges and Future Directions. *SYSTEM THEORY, CONTROL AND COMPUTING JOURNAL*, 1:65–89.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2021. ThamizhiMorph: A morphological parser for the Tamil language. *Machine Translation*, 35(1):37–70.

Jubeerathan Thevakumar, Luxshan Thavarasa, Thanikan Sivatheepan, Sajeev Kugarajah, and Uthayasanker Thayasivam. 2025. EmoTa: A Tamil Emotional Speech Dataset. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 193–201, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Preethi Thinakaran, Malarvizhi Muthuramalingam, Anushiya Rachel Gladston, P Vijayalakshmi, Hema A Murthy, T Nagarajan, et al. 2025. SIToBI–A Speech Prosody Annotation Tool for Indian Languages. *arXiv preprint arXiv:2502.09661*.

Naveed Unjum, Stephanie Evert, Kengatharaiyer Sarveswaran, Ruvan Weerasinghe, and Nevidu Jayatilleke. 2026. Llms for low-resource languages: A survey.

P. Yasmini. 2017. The contrast between jaffna tamil and upcountry tamil: A dialectological study. In *The Third International Conference on Linguistics in Sri Lanka, ICLSL 2017*, page 143. Department of Linguistics, University of Kelaniya, Sri Lanka. Conference proceedings.

A. Zeidan. 2020. Languages by Total Number of Speakers. https://www.britannica.com/topic/languages-by-total-number-of-speakers-2228881. Accessed: 2024-12-29.