SiTa - Sinhala and Tamil Speaker Diarization Dataset in the Wild

Uthayasanker Thayasivam, Thulasithan Gnanenthiram, Shamila Jeewantha, Upeksha Jayawickrama
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
{rtuthaya,thulasithan.20,shamila.20,kasuni.20}@cse.mrt.ac.lk

Abstract

The dynamic field of speaker diarization continues to present significant challenges, despite notable advancements in recent years and the rising focus on complex acoustic scenarios emphasizes the importance of sustained research efforts in this area. While speech resources for speaker diarization are expanding rapidly, aided by semi-automated techniques, many existing datasets remain outdated and lack authentic real-world conversational data. This challenge is particularly acute for low-resource South Asian languages, due to limited public media data and reduced research efforts. Sinhala and Tamil are two such languages with limited speaker diarization datasets. To address this gap, we introduce a new speaker diarization dataset for these languages and evaluate multiple existing models to assess their performance. This work provides essential resources, a novel dataset[1] and valuable insights from model benchmarks to advance speaker diarization for low-resource languages, particularly Sinhala and Tamil.

1 Introduction

The history of speaker diarization dates back to the early 1990s (Park et al., 2022), primarily driven by the need to enhance Automatic Speech Recognition (ASR) systems. Early diarization systems focused on tasks such as transcribing radio broadcasts, conference calls, and speech communication systems, with a major emphasis on improving the accuracy of ASR in multi-speaker environments (Jain et al., 1996; Padmanabhan et al., 1996; Gish et al., 1991). Since then, speaker diarization has been extensively researched over the last few decades, leading to significant advancements in its techniques and applications. Initially, basic clustering (Ng et al., 2001; Tung et al., 2010) and segmentation algorithms were used, but over time, more sophisticated methods, such as deep learning-based approaches, have emerged to perform single-module optimizations (Xie et al., 2016; Lin et al., 2020; Wang et al., 2020), as well as completely end-to-end neural diarization (EEND) systems (Fujita et al., 2019a,b; Horiguchi et al., 2022), pushing the boundaries of diarization performance.

When examining state-of-the-art diarization models, the requirement for large, well-developed diarization datasets can be identified as one of the major factors contributing to the performance of these models. While diarization datasets for languages such as English, Spanish, French, German, and other European languages, along with Asian languages such as Chinese, Japanese, and Korean, possess significantly larger volumes of content, the available datasets for low-resource languages frequently prove inadequate. This scarcity of data often leads to suboptimal performance in speaker diarization models for low-resource languages, where the lack of diversity and size in datasets becomes a major hindrance to model training and evaluation.

The South Asian region is home to many languages that are natively spoken but are often classified as low-resource, resulting in a significant lack of available datasets. Among these languages, Sinhala and Tamil are the two most widely spoken languages in Sri Lanka. Sinhala is an Indo-Aryan language, primarily used by the Sinhalese population of more than 15 million, and Tamil is a Dravidian language, spoken by more than 3 million Tamils in Sri Lanka (Department of Census and Statistics, Sri Lanka, 2012). Despite their large speaker bases, there is a significant lack of rich, diverse linguistic datasets for both languages, especially in fields like speech processing and speaker diarization. Developing diarization datasets for Sinhala and Tamil is essential to advancing speech recognition and diarization systems, which will, in

[1]The dataset is available at https://github.com/SiTa-SpeakerDiarization/SiTa

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 83–92
January 19, 2025.

turn, improve digital accessibility and linguistic resources for these populations.

In this paper, we introduce SiTa, a speech dataset specifically curated for speaker diarization tasks in Tamil and Sinhala languages. We detail the methodology utilized in developing this dataset, which includes data collection, preprocessing, and the annotation pipeline. Additionally, we present experimental results obtained from applying SiTa to state-of-the-art diarization models. Our efforts aim to address the scarcity of resources for low-resource languages like Sinhala and Tamil, thereby contributing to the advancement of multilingual speaker diarization research.

2 Related Work

Over the years, many speaker diarization datasets have been released, particularly in English language. The NIST SRE 2000 (Przybocki and Martin, 2001) or more commonly known as CALLHOME dataset has been the most commonly used speaker diarization dataset, especially for benchmarking purposes. This dataset consists of approximately 500 recordings of multilingual telephone conversations with each session containing two to seven speakers. The CHiME-5/6 challenge (Barker et al., 2018; Watanabe et al., 2020) contains a 50 hour dataset of casual conversations recorded in homes with multi-array microphones that focused on overlapping speech and noisy environments.

The DIHARD Challenges (Ryant et al., 2018, 2019, 2020) are a series of annual events designed to tackle "hard" diarization scenarios where existing systems frequently underperform. The evaluation dataset spans complex domains, including clinical interviews, child language acquisition recordings, restaurant conversations, and online videos. In scoring, overlapped speech was accounted for, and no forgiveness collar was applied, making the challenge even more rigorous.

Following advancements in audio-only diarization, a new frontier opened in audio-visual diarization aimed at enhancing robustness and accuracy. This shift has been facilitated by the introduction of comprehensive multimodal datasets such as the AMI (Mccowan et al., 2005) and AVDIAR (Gebru et al., 2017) corpora. A common characteristic of these datasets is that they are recorded in controlled indoor environments, featuring scripted conversations performed by actors, which contrasts with more naturalistic settings typically encountered in real-world interactions. The AMI (Augmented Multi-party Interaction) Corpus contains 100 hours of meeting recordings captured across multiple locations, offering multi-microphone audio and multi camera video. This dataset also provides synchronized audio-visual streams and transcriptions, enabling the development of sophisticated Automatic Speech Recognition (ASR) systems integrated with speaker diarization. The AVDIAR dataset is designed to encompass a wide variety of multi-speaker scenarios, featuring diverse configurations such as static and moving participants, which facilitate the benchmarking of audio-visual diarization methods in complex interaction settings.

The REPERE (Giraudel et al., 2012) corpus is a multimodal French video dataset developed for advancing automatic people recognition systems, featuring annotated news and debate segments from French television with a focus on varied audio-visual conditions. The RTVE datasets (Lleida et al., 2019, 2020; Ortega et al., 2022) comprising 6 hours, 33 hours, and 25 hours of annotated Spanish speech data respectively, primarily focus on speaker diarization and include enrollment material for speaker identification. These datasets cover diverse accents, spontaneous speech, and overlapping dialogues across various broadcast scenarios.

Having utilized TV shows, meetings, and telephonic data, datasets began incorporating "in the wild" data, primarily from publicly available YouTube videos, to address challenges such as the limited diversity of speech patterns, low presence of overlapping conversations, and the limited variability in background noise that can compromise model performance in real-world applications. The first significant effort in this direction was the VoxCeleb Speaker Recognition Challenge (VoxSRC) series (Chung et al., 2019; Nagrani et al., 2020; Brown et al., 2022; Huh et al., 2023), which initially focused solely on speaker verification tasks; however, the diarization task was later introduced as Track 4 in subsequent iterations. Following this initiative, several novel "in the wild" speaker diarization datasets were created. Voxconverse (Chung et al., 2020) is one of the first dedicated large-scale diarization datasets, containing 64 hours of diverse YouTube videos which included a test set of 232 videos and a dev set of 216 videos. Its contributions also include a semi-automatic dataset creation pipeline, which significantly reduces the time required to annotate videos

with speaker labels. This innovation addresses a key reason for the scarcity of large-scale diarization datasets derived from natural conversations such as those in YouTube videos. Atomic Visual Action Audio-Visual Diarization (AVA-AVD) dataset (Xu et al., 2022) is another comparable multilingual speaker diarization dataset developed based on the AVA-Active Speaker dataset (Roth et al., 2020), which focuses on detecting active speakers in audiovisual contexts but excluding the videos with dubbed scenes, as dubbing can disrupt the audio-visual synchronization. The AVA-AVD dataset consists of 351 videos totaling 29 hours of content, extracted from movies produced worldwide with diversity in ethnicity, language, accents, dialects, and age. Similarly, MSDWild (Liu et al., 2022) is a large multilingual speaker diarization dataset, featuring 3,143 YouTube videos with an emphasis on daily conversations, totaling 80 annotated hours.

During the early development of speaker diarization, publications focusing on low-resource speaker diarization datasets were limited, as initial research efforts primarily concentrated on larger corpora in English. Nonetheless, attempts to create non-English diarization datasets were made during the late 1990s and early 2000s, coinciding with the expanding interest in speech recognition. Among these early initiatives were the CALLHOME collections (Canavan and Zipperlen, 1996; Wheatley, 1996), which includes multilingual telephone conversational data for both diarization and speaker identification. In 2002, the Karlsruhe Institute of Technology introduced the GlobePhone corpus (Schultz, 2002), encompassing audio recordings and transcriptions in 20 languages, including Hausa, Swahili, and Vietnamese. Additionally, in 2005, the International Institute of Information Technology in India developed speech corpora for Tamil, Telugu, and Marathi languages (Chitturi et al., 2005). Released in 2021, AISHELL-4 (Fu et al., 2021) provided a Mandarin Chinese meeting corpus of 118 hours. In addition, AliMeeting (Yu et al., 2022) and RAMC (Yang et al., 2022) datasets feature meeting scenarios in distinct room environments and Mandarin phone call recordings respectively. The Corpus of Spontaneous Japanese (Maekawa, 2003) offered 12 hours of dialogue data captured using headset microphones, highlighting natural, spontaneous conversations between two speakers.

In the scope of South Asian languages, the DIarization of SPeaker and LAnguage in Conversational Environments (DISPLACE) challenge (Baghel et al., 2023) introduced a unique dataset for diarization tasks in multilingual, multi-speaker conversational contexts, highlighting the challenges posed by code-mixed and code-switched speech. In the second DISPLACE challenge (Kalluri et al., 2024), three tasks were introduced: speaker diarization (SD), identifying "who spoke when"; language diarization (LD), determining "which language was spoken when"; and automatic speech recognition (ASR), all of which are complicated by speaker overlaps and frequent language transitions. The dataset includes recorded conversations in various room settings, covering topics such as culture, lifestyle, entertainment, and sports. It spans nine Indian languages such as Hindi, Kannada, Bengali, Malayalam, Telugu, Tamil, and Indian English totaling 38 hours of conversational speech across 197 speakers. CONVURL (Zaheer et al., 2025) is a 24-hour dataset that consists of natural spontaneous conversations in Urdu, featuring 212 unique speakers. The dataset includes 38 clips sourced from YouTube videos, primarily encompassing scholarly debates and political talk shows. In 2022, another YouTube-based Tamil diarization dataset (Jarashanth et al., 2022) was published focusing more on overlapped speech. The development of speech corpora for Sinhala language has also gained attention, particularly within the domain of automatic speech recognition (ASR). Notable efforts include the publication of a 65-hour speech corpus in 2013 (Nadungodage et al., 2013), and the release of a 4.15-hour Sinhala speech corpus (Dinushika et al., 2019) in 2019. However, research literature on speaker diarization in Sinhala remains limited, indicating an insufficient focus in this domain. This study seeks to address this gap by introducing a speaker diarization corpus in the Sinhala and Tamil language to support further research and development in this area.

3 SiTa

3.1 Dataset Description

SiTa is an audio-only speaker diarization dataset for Sinhala and Tamil languages, comprising two distinct subsets: one for Sinhala speech and the other for Tamil. The Sinhala subset includes 60 videos totaling approximately 600 minutes (10 hours) of audio, while the Tamil subset contains 14 videos, accounting for around 120 minutes (2 hours) of audio. SiTa features multi-speaker conversations

set	# videos	# mins	# speakers	video durations (min)	speech %	overlap %
Sinhala	60	602	1 / 3.4 / 10	5.1 / 10.0 / 16.6	37.3 / 88.6 / 99.1	0 / 1.5 / 9.2
Tamil	14	121	2 / 3.0 / 6	5.0 / 8.7 / 14.5	78.6 / 92.7 / 97.4	0 / 3.3 / 14.21

Table 1: Statistics of the SiTa Dataset. Each entry with three values represents the minimum, mean, and maximum. speech %: is the proportion of audio time that contains speech. overlap %: is the proportion of speech occurring when two or more speakers are active simultaneously.

captured "in the wild" and encompasses a range of conversation types, from controlled settings with few speakers, slow-paced dialogue, minimal noise, and negligible speaker overlap such as panel discussions and interviews to more dynamic and challenging contexts. These include political debates, quiz programs, and lively discussions with numerous speaker turns, background noise, and frequent overlaps among a larger group of speakers. A unique aspect of SiTa is its inclusion of code-mixing, where English terms appear interspersed within native Tamil and Sinhala sentences. This phenomenon, common in South Asian languages, introduces an additional layer of complexity absent in most English speaker diarization datasets. Table 1 provides a comprehensive statistical overview of the Tamil and Sinhala subsets within the SiTa dataset.

3.2 Data Collection

In line with approaches used in in-the-wild datasets, YouTube videos were manually selected from diverse domains, including intellectual discussions, education, morning shows, celebrity interviews, political debates, and children's programs. Figure 1 illustrates the distribution percentages of video types selected for data collection in the Sinhala and Tamil languages. To ensure a broad selection, keywords from Sinhala, Tamil, and English were used in the search, as many content creators use English titles regardless of the content's spoken language. Videos were sourced from independent YouTube channels as well as the official YouTube channels of public television networks. Additionally, videos with extensive code-switching, those containing full sentences in English rather than isolated terms were avoided to allow for a focused study of language-specific dynamics in Sinhala and Tamil. Only one excerpt from each of these videos were carefully chosen for the dataset to maximize speech activity, speaker turns, and instances of overlapping speech. Videos with excessive noise or crowd laughter were avoided to ensure label clarity, although segments with applause were included without explicit labeling. All selected video excerpts were converted to a mono-channel waveform audio (WAV) format with a sampling rate of 16 kHz.

3.3 Annotation

A semi-automated, two-stage pipeline was implemented in the development of both Sinhala and Tamil subsets of SiTa. In the Initial stage, speaker labels and timestamps were generated using Pyannote 3.1[2], which were then manually reviewed and corrected using VGG Image Annotator (VIA) (Dutta and Zisserman, 2019), a standalone annotation tool for images, audio, and video during the second stage. VIA facilitates the addition of multiple timelines for different speakers, adjustable playback speeds, timeline zooming, selective playback of labeled segments, and the seamless insertion of new labels at the current playback position without manual adjustments. Figure 2 illustrates the user-friendly interface employed for the manual annotation of audio files.

3.3.1 Guidelines

Annotation guidelines and verification protocols were adopted from established datasets such as VoxConverse (Chung et al., 2020), MSDWild (Liu et al., 2022), and AVA-AVD (Xu et al., 2022). Minor non-verbal sounds, such as short utterances ("mmm", "ooo"), were ignored unless they were at least 100 ms in duration and clearly audible. Within a speech segment from the same speaker, pauses were not treated as split points unless the pause duration exceeded 250 ms. Additionally, annotators were instructed to ensure that label timestamp boundaries did not deviate from the actual speech boundaries by more than 100 ms to maintain precision. Music segments were excluded from labeling. In instances where speaker distinction was challenging based solely on audio, annotators referred to the corresponding YouTube video to accurately identify each speaker. The average time required for annotating an audio file, adhering to aforementioned guidelines, was approximately 8–10 times

[2]https://huggingface.co/pyannote/speaker-diarization-3.1

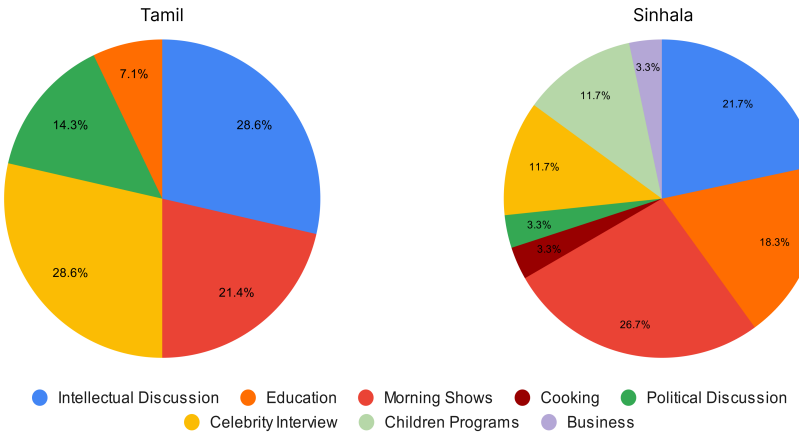

Figure 1: Percentage distribution of video types selected for data collection in Sinhala and Tamil languages.

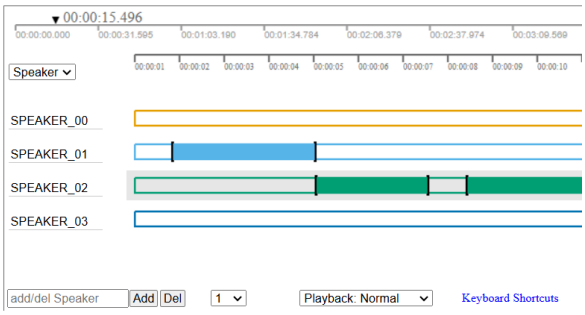

Figure 2: The user interface of the VGG Image Annotator (VIA), showing the audio playback timeline, speaker timelines, and temporal segments representing speaker labels.

the duration of the original audio. This duration could be reduced for audio clips with fewer speakers and minimal overlap.

3.3.2 Quality Assurance

To ensure consistency and maintain high-quality annotations, periodic double-annotations were conducted during which annotators independently annotated the same video, and the diarization error between the two annotations was calculated using one as the reference and the other as the hypothesis. This diarization error rate was maintained below 1%, and annotators were reminded to strictly follow the guidelines if this threshold was exceeded. Finally, a separate reviewer, distinct from the initial annotator, verified that the annotations accurately matched the WAV files and confirmed the overall quality and consistency of the annotations.

4 Experiments

4.1 Evaluation Metric

In this study, we used the Diarization Error Rate (DER) as our primary evaluation metric to assess the performance of speaker diarization. DER is defined as the sum of missed speech (MS), false alarm (FA), and speaker confusion (SC), normalized by the total duration of speech. Mathematically, it can be expressed as:

$$DER = \frac{MS + FA + SC}{Total\ Speech} \quad (1)$$

Consistent with prior research, the DER is calculated with a forgiveness collar of 0.25 seconds around each detected speech segment to account for slight temporal discrepancies in speaker labeling.

4.2 Baseline Models

We evaluated three baseline models to test our proposed SiTa dataset. These models include both modular and end-to-end (EEND) systems. The modular systems allow for distinct components to be developed and optimized separately, while the EEND systems aim to streamline the process by integrating all stages of speaker diarization into a single framework. This diversity in model architectures chosen enables a comprehensive assessment of our dataset's performance across different approaches to speaker diarization. All selected models utilized publicly available pre-trained versions, and their corresponding results on SiTa dataset are presented in Table 2.

Set	Model	MS	FA	SC	DER
Sinhala	End-to-End Segmentation (Bredin and Laurent, 2021)	2.5	12.9	9.5	**6.4**
	Powerset Cross-Entropy Diarization (Plaquet and Bredin, 2023)	1.6	12.0	9.7	**6.0**
	DiaPer (Landini et al., 2024)	19.4	6.8	42.4	**14.3**
Tamil	End-to-End Segmentation (Bredin and Laurent, 2021)	3.9	10.0	11.2	**6.0**
	Powerset Cross-Entropy Diarization (Plaquet and Bredin, 2023)	1.9	9.3	16.1	**6.6**
	DiaPer (Landini et al., 2024)	14.0	6.3	30.1	**11.5**

Table 2: Performance Metrics of Speaker Diarization Models on the SiTa Dataset. MS: Missed Speech; FA: False Alarms; SC: Speaker Confusion; DER: Diarization Error Rate.

4.2.1 End-to-End Speaker Segmentation with Overlap-Aware Re-segmentation

The End-to-End Speaker Segmentation model[3] (Bredin and Laurent, 2021) integrates voice activity, speaker change, and overlap detection into a single architecture by modeling it as a multi-label classification problem incorporating permutation-invariant training. The second contribution of this model is the incorporation of overlap-aware re-segmentation, which refines the initial segmentation output by based on contextual and temporal information across segments, ultimately enabling accurate assignment of overlapping speech segments to their corresponding speakers.

4.2.2 Powerset Cross-Entropy Diarization

The Powerset Cross-Entropy diarization model[4] (Plaquet and Bredin, 2023) uses the concept of assigning unique label combinations to overlapping speakers, using a powerset approach replacing the existing multi label classification approach. By representing all possible speaker combinations as distinct classes, this method allows for the removal of the detection threshold hyperparameter and thereby accurately handling overlaps in speech.

4.2.3 DiaPer: Perceiver-based Diarization

DiaPer[5] (Landini et al., 2024) introduces a novel architecture that replaces the encoder-decoder attractor component of the EEND-EDA (Horiguchi et al., 2022) with a Perceiver-based module. This new design, features a decoder that generates speaker attractors using a transformer-based Perceiver with fixed-size latent representations and cross-attention mechanisms, thereby eliminating the sequential processing typically associated with the LSTM-based EEND-EDA. Ultimately, DiaPer enhances speaker count estimation, accelerates inference speed, and improves diarization accuracy through this innovative module.

4.3 Results and Discussion

For both modular systems (End-to-End Segmentation and Powerset Cross-Entropy Diarization), the Diarization Error Rate (DER) for the Sinhala and Tamil sets in our SiTa dataset closely aligns with the values reported by the original authors of each model for the VoxConverse dataset. The lower DER observed for the SiTa dataset, compared to the reported DER for VoxConverse by the authors, in the EEND model, DiaPer, can be attributed to the training approach. While DiaPer was primarily trained on simulated speech data for its evaluation on VoxConverse, the version used for evaluating our dataset was fine-tuned on the MSDWild dataset. Additionally, when compared to the other two modular systems, DiaPer exhibits a relatively higher DER. This can be attributed to its lack of Voice Activity Detection (VAD) or other pre-processing steps, making it more susceptible to noise. Moreover, EEND models, including DiaPer, are empirically known to overfit more than other diarization approaches, further contributing to the higher DER observed.

Apart from that, the DER is closely influenced by the percentage of overlapping speech within each dataset. According to the statistics, the Sinhala set has an average overlap of 1.5%, while the Tamil set exhibits a higher mean overlap of 3.3%. This difference results in higher speaker confusion in the two modular diarization systems on the Tamil set, as increased overlap poses challenges in accurately distinguishing between speakers. This suggests that overlap handling is crucial for improving DER in multi-speaker environments. Code-mixing, however, shows minimal impact on DER across our dataset. Common issues observed in the generated diarization results included excessive speaker la-

[3]https://huggingface.co/pyannote/segmentation

[4]https://github.com/FrenchKrab/IS2023-powerset-diarization

[5]https://github.com/BUTSpeechFIT/DiaPer

bels in non-speech segments, missing labels for speech segments, omitted speaker timelines resulting in fewer identified speakers, and instances of swapped speaker labels. Specifically, male speakers with similarly rough voices were frequently assigned identical labels. Female speakers were often mislabeled, with extreme cases where all female speech segments were assigned the same speaker label. In some cases, even though the labeling was mostly correct, short utterances such as 'mmmm's and 'aaah's were mislabeled. Non-speech sounds, such as chimes in quiz programs, were occasionally mislabeled as speakers. In children's speech segments, the initial utterances of female child voices were often misattributed to the wrong female child speaker, similar to cases in clips with only adult female speakers. Male child voices were sometimes misattributed to female child speakers or even to adult female speakers if present. Additionally, segments of speech, which preceded or followed music, exhibited slight misalignments in timestamps. Furthermore, the staccato rhythm of the Sinhala language was particularly prominent in intellectual discussions. In such cases, continuous speech from a single speaker was often fragmented into multiple segments.

5 Limitations

The size of the Sinhala and Tamil subsets in our SiTa dataset is limited to around 12 hours, which restricts its use for reliably training speaker diarization models from scratch. As a result, these subsets were only used as test sets to evaluate the models' performance. Furthermore, the amount of overlapping speech in the larger Sinhala subset is relatively low, as the annotation of such segments is an error-prone and time-intensive process. While the dataset does encompass a range of domains, the number of audio samples from each domain remains limited, which could affect the robustness of model evaluation across different contexts. We did not intentionally include all regional dialects present in Sinhala and Tamil in the SiTa dataset and the selected YouTube videos primarily consisted of the standard language typically spoken in TV broadcasts, as well as in formal and urban settings. Even in more casual videos, this was largely the case. As a result, the dataset lacks representation of regional dialectal diversity, and we are unable to evaluate the potential dynamics or impacts of it on speaker diarization performance.

6 Conclusion

In this study, we introduced SiTa, a speaker diarization dataset comprising subsets in Sinhala and Tamil languages. Through evaluations using diverse baseline models, including modular and end-to-end (EEND) approaches, we observed the effects of language-specific characteristics, overlap levels, and continuous speech patterns on diarization performance. Our results reveal that overlapping speech significantly impacts DER, highlighting the need for effective overlap handling in speaker diarization. For the Sinhala subset, we found that increasing gap tolerance before segmenting speech can mitigate over-segmentation, thereby enhancing diarization accuracy.

Though the dataset size limits its use to testing rather than training, SiTa serves as a valuable benchmark for evaluating diarization in low-resource languages and presents a first step towards more comprehensive multilingual diarization datasets. Future work could focus on expanding SiTa and improving overlap annotation to enable training applications and further insights into language-specific diarization challenges.

References

Shikha Baghel, Shreyas Ramoji, Sidharth, Ranjana H, Prachi Singh, Somil Jain, Pratik Roy Chowdhuri, Kaustubh Kulkarni, Swapnil Padhi, Deepu Vijayasenan, and Sriram Ganapathy. 2023. The displace challenge 2023 - diarization of speaker and language in conversational environments. In *INTERSPEECH 2023*, pages 3562–3566.

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *Interspeech 2018*, pages 1561–1565.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. *arXiv preprint arXiv:2104.04045*.

Andrew Brown, Jaesung Huh, Joon Son Chung, Arsha Nagrani, Daniel Garcia-Romero, and Andrew Zisserman. 2022. Voxsrc 2021: The third voxceleb speaker recognition challenge. *arXiv preprint arXiv:2201.04583*.

Alexandra Canavan and George Zipperlen. 1996. Callhome japanese speech. Web Download.

Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinderpal Singh, R. N. V. Sitaram, and S. P. Kishore. 2005. Development of indian language speech databases for large vocabulary speech recognition systems.

Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. 2020. Spot the conversation: speaker diarisation in the wild. *arXiv preprint arXiv:2007.01216*.

Joon Son Chung, Arsha Nagrani, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman. 2019. Voxsrc 2019: The first voxceleb speaker recognition challenge. *arXiv preprint arXiv:1912.02522*.

Department of Census and Statistics, Sri Lanka. 2012. Census of population and housing of sri lanka, 2012 – table a3: Population by district, ethnic group and sex. http://203.94.94.83:8041/Pages/Activities/Reports/SriLanka.pdf. Additional data available at https://www.statistics.gov.lk/PopHouSat/CPH2012Visualization/htdocs/index.php?action=Map&indId=10&usecase=indicator and https://www.statistics.gov.lk/pophousat/cph2012visualization/htdocs/index.php?usecase=indicator&action=Data&indId=10.

Thilini Dinushika, Lakshika Kavmini, Pamoda Abeyawardhana, Uthayasanker Thayasivam, and Sanath Jayasena. 2019. Speech command classification system for sinhala language based on automatic speech recognition. *2019 International Conference on Asian Language Processing (IALP)*, pages 205–210.

Abhishek Dutta and Andrew Zisserman. 2019. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA. ACM.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al. 2021. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019a. End-to-end neural speaker diarization with permutation-free objectives. *arXiv preprint arXiv:1909.05952*.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019b. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303.

Israel D Gebru, Sileye Ba, Xiaofei Li, and Radu Horaud. 2017. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1086–1099.

Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The repere corpus : a multimodal corpus for person recognition. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

H. Gish, M.-H. Siu, and R. Rohlicek. 1991. Segregation of speakers for speech recognition and speaker identification. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 873–876 vol.2.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Paola Garcia. 2022. Encoder-decoder based attractors for end-to-end neural diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1493–1507.

Jaesung Huh, Andrew Brown, Jee-weon Jung, Joon Son Chung, Arsha Nagrani, Daniel Garcia-Romero, and Andrew Zisserman. 2023. Voxsrc 2022: The fourth voxceleb speaker recognition challenge. *arXiv preprint arXiv:2302.10248*.

Udeeta Jain, Matthew Siegler, Sam-Joo Doh, Evandro Gouvea, Juan Huerta, Pedro Moreno, Bhiksha Raj, and Richard Stern. 1996. Recognition of continuous broadcast news with multiple unknown speakers and environments.

S.T. Jarashanth, K. Ahilan, R. Valluvan, T. Thiruvaran, and A. Kaneswaran. 2022. Overlapped speech detection for improved speaker diarization on tamil dataset. In *2022 6th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, pages 1–5.

Shareef Babu Kalluri, Prachi Singh, Pratik Roy Chowdhuri, Apoorva Kulkarni, Shikha Baghel, Pradyoth Hegde, Swapnil Sontakke, Deepak K T, S.R. Mahadeva Prasanna, Deepu Vijayasenan, and Sriram Ganapathy. 2024. The second displace challenge: Diarization of speaker and language in conversational environments. In *Interspeech 2024*, pages 1630–1634.

Federico Landini, Themos Stafylakis, Lukáš Burget, et al. 2024. Diaper: End-to-end neural diarization with perceiver-based attractors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Qingjian Lin, Yu Hou, and Ming Li. 2020. Self-attentive similarity measurement strategies in speaker diarization. In *Interspeech 2020*, pages 284–288.

Tao Liu, Shuai Fan, Xu Xiang, Hongbo Song, Shaoxiong Lin, Jiaqi Sun, Tianyuan Han, Siyuan Chen, Binwei Yao, Sen Liu, Yifei Wu, Yanmin Qian, and Kai Yu. 2022. Msdwild: Multi-modal speaker diarization dataset in the wild. In *Interspeech 2022*, pages 1476–1480.

Eduardo Lleida, Alfonso Ortega, Antonio Miguel, Virginia Bazán, Carmen Pérez, Manuel Gómez, and Alberto de Prada. 2020. Albayzin evaluation: Iberspeech-rtve 2020 multimodal diarization and scene description challenge.

Eduardo Lleida, Alfonso Ortega, Antonio Miguel, Virginia Bazán-Gil, Carmen Pérez, Manuel Gómez, and Alberto de Prada. 2019. Albayzin 2018 evaluation: The iberspeech-rtve challenge on speech technologies for spanish broadcast media. *Applied Sciences*, 9(24).

Kikuo Maekawa. 2003. Corpus of spontaneous japanese : Its design and evaluation.

Iain Mccowan, J Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, V Karaiskos, M Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P Wellner. 2005. The ami meeting corpus. *Int'l. Conf. on Methods and Techniques in Behavioral Research*.

Thilini Nadungodage, Viraj Welgama, and Ruvan Weerasinghe. 2013. Developing a speech corpus for sinhala speech recognition.

Arsha Nagrani, Joon Son Chung, Jaesung Huh, Andrew Brown, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman. 2020. Voxsrc 2020: The second voxceleb speaker recognition challenge. *arXiv preprint arXiv:2012.06867*.

Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Alfonso Ortega, Antonio Miguel, Eduardo Lleida, Virginia Bazán, Carmen Pérez, and Alberto de Prada. 2022. Iberspeech-rtve 2022 speaker diarization and identity assignment. In *Albayzin Evaluation*. Accessed: 2024-11-05.

M. Padmanabhan, L.R. Bahl, D. Nahamoo, and M.A. Picheny. 1996. Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 701–704 vol. 2.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. *arXiv preprint arXiv:2310.13025*.

Mark Przybocki and Alvin Martin. 2001. 2000 nist speaker recognition evaluation. Web Download. LDC2001S97.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. 2020. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE.

Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2019. The second dihard diarization challenge: Dataset, task, and baselines. *arXiv preprint arXiv:1906.07839*.

Neville Ryant, Mark Liberman, James Fiumara, and Christopher Cieri. 2018. First dihard challenge evaluation - nine sources. Web Download. LDC Catalog No. LDC2019S12, DOI: 10.35111/1bsf-4c55.

Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2020. The third dihard diarization challenge. *arXiv preprint arXiv:2012.01477*.

Tanja Schultz. 2002. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 345–348.

Frederick Tung, Alexander Wong, and David A. Clausi. 2010. Enabling scalable spectral clustering for image segmentation. *Pattern Recognition*, 43(12):4069–4076.

Jixuan Wang, Xiong Xiao, Jian Wu, Ranjani Ramamurthy, Frank Rudzicz, and Michael Brudno. 2020. Speaker diarization with session-level speaker embedding refinement using graph neural networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7109–7113. IEEE.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, and Neville Ryant. 2020. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, pages 1–7.

Barbara Wheatley. 1996. Callhome mandarin chinese transcripts. Web Download.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR.

Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. 2022. Ava-avd: Audio-visual speaker diarization in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3838–3847.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. 2022. Open source magicdata-ramc: A rich annotated mandarin conversational(ramc) speech dataset. pages 1736–1740.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 2022. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In *Proc. ICASSP*. IEEE.

Nimra Zaheer, Agha Ali Raza, and Mudassir Shabbir. 2025. Conversations in the wild: Data collection, automatic generation and evaluation. *Computer Speech Language*, 89:101699.