The DISPLACE Challenge 2023 - DIarization of SPeaker and LAnguage in Conversational Environments

Shikha Baghel[1], Shreyas Ramoji[1], Sidharth[1], Ranjana H[1], Prachi Singh[1], Somil Jain[2], Pratik Roy Chowdhuri[2], Kaustubh Kulkarni[1], Swapnil Padhi[1], Deepu Vijayasenan[2], Sriram Ganapathy[1]

[1]LEAP Lab, Department of Electrical Engineering, Indian Institute of Science, Bengaluru, India
[2]Department of Electronics and Communication, National Institute of Technology Karnataka, Surathkal, India

shikhabaghel@iisc.ac.in

Abstract

In multilingual societies, social conversations often involve code-mixed speech. The current speech technology may not be well equipped to extract information from multi-lingual multi-speaker conversations. The DISPLACE challenge entails a first-of-kind task to benchmark speaker and language diarization on the same data, as the data contains multi-speaker conversations in multilingual code-mixed speech. The challenge attempts to highlight outstanding issues in speaker diarization (SD) in multilingual settings with code-mixing. Further, language diarization (LD) in multi-speaker settings also introduces new challenges, where the system has to disambiguate speaker switches with code switches. For this challenge, a natural multilingual, multi-speaker conversational dataset is distributed for development and evaluation purposes. The systems are evaluated on single-channel far-field recordings. We also release a baseline system and report the highlights of the system submissions.

Index Terms: Speaker diarization, language diarization, code-mixing, conversational speech, DISPLACE challenge.

1. Introduction

In multilingual communities, social conversations frequently involve code-mixed and code-switched speech with multiple speakers [1, 2]. *Code-mixing* is the scenario where words or morphemes from one language (secondary) are used within a sentence context of another language, and *code-switching* is the scenario where a language switch happens at the sentence or phrase level. Both code-mixing and code-switching are prevalent in multilingual communities like Asia, USA, and Europe [3, 4, 5]. As most of the current speech technologies are developed for single-speaker monolingual settings, the key first step in dealing with code-switched multi-speaker recordings would be the segregation of the input stream based on speaker and language.

The term *Diarization* was initially associated with the task of detecting and segmenting homogeneous audio regions based on speaker identity. This task, widely known as speaker diarization (SD), generates the answer for "who spoke when". In the past few years, the term diarization has also been used in linguistic context. Language diarization (LD) is the task of identifying "which language was spoken when", and aims to cluster segments of the same language [6, 7].

Most conventional speech processing systems, such as automatic speech recognition (ASR) and SD, are developed for monolingual scenarios. At the same time, language recognition systems are not benchmarked in multi-speaker settings. Further, most of the previous dataset building efforts focused on either of these challenges in isolation.

In this paper, we describe the details of the Interspeech DIarization of SPeaker and LAnguage in Conversational Environments (DISPLACE) challenge. The key contributions of this challenge are:

- Build a dataset resource of multi-lingual multi-speaker conversational speech, with code switching, natural overlaps, reverberation and noise, along with speech, speaker and language annotations.
- Benchmark speaker and language diarization on this dataset and announce an open call for system development.
- Provide a baseline system which is made available to the challenge participants, that performs speech activity detection, speaker diarization and language diarization.
- A leader-board style evaluation platform for technology development.

2. Related Work

Speaker Diarization - The majority of the early research in SD was driven by the National Institute of Standards and Technology-Rich Transcription (NIST-RT) [8, 9] evaluations on Broadcast News (BN) and conversational telephonic speech in English. The RT-05S [10], RT-06S, RT-07S [11] and RT-09 evaluations considered conference and lecture room meeting domain (in English) for the SD task [8]. The NIST-RT evaluations also contributed the Diarization Error Rate (DER), which remains the primary evaluation metric for the SD systems. Recently, there have been several evaluation challenges, namely, the DIHARD challenge [12, 13, 14], Fearless Steps Series [15, 16], the Iberspeech-RTVE [17], CHiME-6 [18] and VoxSRC-20 [19].

Language Diarization - In recent years, LD has become one of the contemporary research issues in multilingual speech processing [7, 20]. In [5], Lyu et al. attempted the LD task on South-East-Asia Mandarin/English (SEAME) dataset [21]. Yilmaz et al. [2] used a radio broadcast dataset containing Frisian-Dutch code-switching for LD task. The first attempt for LD in the Indian languages was reported by Spoorthy et al. [4]. Recently, Shah et al. [22] organized a workshop that included the LID task in code-switched data. The data used in this workshop is referred to as WSTCSMC [22].
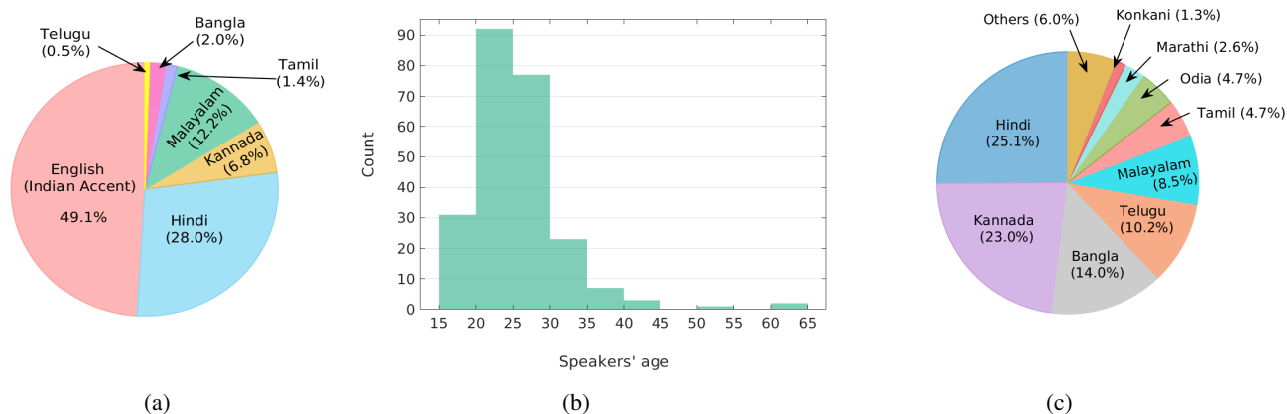
Figure 1: *The DISPLACE corpus statistics of (a) Spoken language, (b) Speaker age, and (c) Primary language (L1) of the speakers.*

3. DISPLACE Corpus

3.1. Recording Setup

The data is recorded at two different academic institutes. The recording rooms deployed are different in shape, size, and acoustic properties. These recording rooms did not have any specific soundproof settings and hence, the recordings contain natural noise and background speech seen in outdoor data collection settings. The data collection paradigm contains a close-talking microphone worn by each speaker and a common far-field desktop microphone. The close-talking audio was recorded using a lapel microphone connected to either an audio recorder or an Android phone. The participants were seated in either a circular or semi-circular setting and the far-field omni-directional microphone was used to collect the conversational speech. Speakers were approximately equi-distant from each other and the far-field microphone. All the recordings contain single-channel data. The worn microphone speech was only recorded for annotation purposes (for ease of human listening to determine the speech, speaker and language boundaries). All the system development and evaluation is carried out using the far-field microphone recordings.

3.2. Data Collection

Each conversation of duration 30-60 minutes comprises 3-5 participants. The participants had self-reported proficiency in at least one Indian language (L1) along with Indian accented English. The participants that came together for a conversation were selected based on the L1 languages. The topic of conversation was chosen among a diverse set of choices like climate change, culture, politics, sports, and entertainment. Before the recording, the participants also provided written consent and were informed of the guidelines. The participants also received monetary compensation for the recording. Finally, all the worn-mic recordings are time-aligned with the far-field audio, resampled at 16 kHz, and normalized to $[-1, 1]$ range.

3.3. Annotation

The data annotations were generated with professional annotators, who listened to the lapel microphone recordings. Each close-talking microphone is labeled based on the target speaker (participant wearing the microphone).

The first annotation task was to mark the speech activity of the target speaker (speaker activity). The rest of the regions (including silence, long-pause, and speech regions of non-target speakers) are marked as *non-speech*. The non-speech audio of the target speaker, such as laughing, coughing, tongue clicks, etc., are annotated. The prominent background sounds, such as telephone rings, car honk, etc are also marked separately.

The second annotation task was to mark language labels corresponding to the speaker activity regions. The language label was marked for every word to incorporate code-mixing of short linguistic content.

A third annotation task was undertaken to generate the multilingual transcripts of the spoken content. These transcripts are not released as part of the DISPLACE challenge and will be used for future evaluations.

The participants use a lot of fillers and back-channel words, such as *ahh*, *umm*, and *ohh*. For such cases, past and future contexts across the non-lexical words or sounds were considered for assigning a language label. The annotation process was a tedious one involving speech, speaker and language annotations. We also employed multiple levels of quality checks before arriving at the final annotations used in the challenge. The annotations obtained for all the participants' lapel microphones are combined to generate a single Rich Transcription Time Marked (RTTM) annotation file for each conversation.

3.4. Development and Evaluation set

For the challenge, systems are evaluated on single channel far-field audio. Only the development and evaluation sets were released. There was no training data provided, and the participants were free to use any proprietary and/or public resource for training the diarization systems. The Development (Dev) and Evaluation (Eval) set consists of ≈ 15.5 hours (27 recordings) and 16 hours (29 recordings) of multilingual conversations, respectively. The evaluation was done in two phases, namely, Phase-1 and Phase-2. The Phase-1 evaluation set consists of a subset of the full evaluation set with 20 recordings spanning 11.5 hours, and the Phase-2 evaluation set consisted of the full Eval set. The Phase-1 evaluation ran for 7 weeks (from Jan-15-2023 till March-4-2023), while Phase-2 was open till May-22-2023.

The speakers in the evaluation and development sets are mutually exclusive. Also, the Eval recordings contain new languages not seen in the Dev set. The DISPLACE corpus (DEV and Eval Phase-1) contains conversations in Hindi, Kannada, Bengali, Malayalam, Telugu, Tamil, and Indian English. The
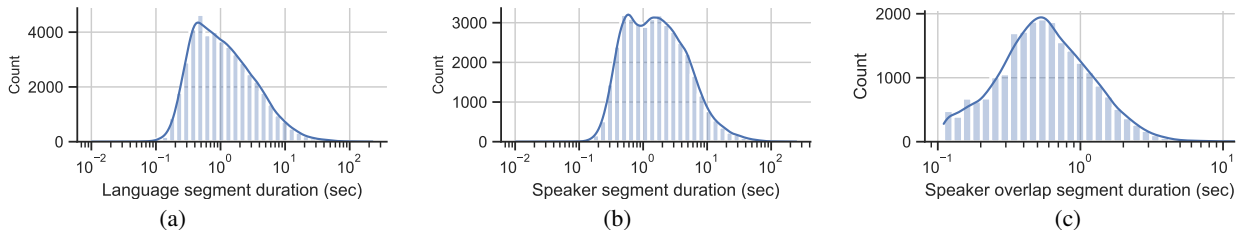
Figure 2: *Histograms of (a) language, (b) speaker, and (c) overlap segment durations obtained from the DISPLACE corpus.*

distribution of these languages is illustrated in Figure 1 (a). The corpus also contains a significant amount (16.56%) of over-lapped speech. Majority of the speakers' age ranges from 17 to 40 years (Figure 1 (b)). The dataset contains 33% female and 67% male participants. The speakers in the DISPLACE corpus had different L1 languages, such as Bengali, Hindi, Kannada, Konkani, Malayalam, Marathi, etc. Figure 1 (c) displays the distribution of L1 languages spoken by speakers in the DISPLACE dataset (Dev and Eval Phase-1). The *Others* category includes L1 languages like Assamese, Gujarati, Kashmiri, Maithili, Nepali, Punjabi and Tulu. The corpus has a variety of speech accents as well.

Figure 2 demonstrates the distribution of language, speaker, and overlap segment durations. A language/speaker segment represents a homogeneous speech region containing a single language/speaker. A segment containing simultaneous speech of multiple speakers is considered as a speaker overlap segment. In Figures 2 (a), (b), and (c), x-axes represent the segment duration (s) and y-axes denote segment count. In Figure 2 (a), the majority (99.87%) of the language turns have a duration in the range of 0.10s to 100s. Similarly, 99.96% of the speaker turn durations vary from 0.1s to 100s. As shown in Figure 2 (c), overlap segment durations vary from 0.10s to 10s. Approximately 90.60% overlap segments lie between 0.2s to 4s duration. These language, speaker, and overlap turn statistics show that frequent speaker and language switching is common in informal social conversations.

4. Challenge Tasks

The challenge tasks entail diarization of each conversational audio based on speaker and language information. The speech activity detection is not provided and the participants are encouraged to develop their own system. Submissions to both tracks are evaluated only on speech-based speaker activity regions, including voiced back-channels and fillers. However, non-speech speaker activities, such as laughing, clapping, sneezing, etc., are excluded as non-speech. The DISPLACE challenge contains two tracks,

- Track-1: Speaker Diarization in multilingual scenarios
- Track-2: Language Diarization in multi-speaker settings

The metric used for both tracks is the diarization error rate (DER).

$$DER = \left(\frac{D_{FA} + D_{miss} + D_{error}}{D_{total}} \right) \times 100 \qquad (1)$$

where, D_{FA} represents the total speech duration which is not attributed to a reference speaker/language, D_{miss} denotes the total reference speaker/language duration, which is not attributed to a system speaker/language, D_{error} represents the

Table 1: *Baseline speaker diarization results (in terms of DER %) for the DISPLACE development (Dev.) and evaluation (Eval) data (Phase-1 and Phase-2) using AHC and spectral clustering, followed by VB-HMM re-segmentation.*

	Method	DER (%)		
		Dev	Eval	
			Phase1	Phase2
B1	AHC	33.97	40.98	38.86
	AHC+VB-HMM	32.57	40.08	37.96
B2	SC	31.66	33.58	32.51
	SC+VB-HMM	31.17	32.94	32.18

total system speaker/language duration attributed to the wrong reference speaker/language, and D_{total} denotes the total reference speaker/language duration. For the DISPLACE challenge, DER is computed with overlap and without collar.

5. Baseline Systems

5.1. Speech Activity Detection (SAD)

The SAD system from the DIHARD-III baseline [14] is used. The model contains 5 Time Delay Neural Network (TDNN) layers [23] followed by 2 statistics pooling layers [24]. The DNN was trained on DIHARD-III development set to perform speech and non-speech classification and this model is trained for 40 epochs.

5.2. Speaker Diarization (SD)

Speaker diarization system, based on DIHARD-III [14], is also adapted for this purpose. This modeling framework involves segmenting the audio into short overlapping segments of 1.5s with 0.25s shift and extracting speaker embeddings (x-vectors). These are then used to perform probabilistic linear discriminant analysis (PLDA) scoring followed by agglomerative hierarchical clustering (AHC). For refining speaker boundaries, re-segmentation is performed using VB-HMM with posterior scaling [25, 26].

The x-vector extractor model is a 13-layer Extended-Time Delay Neural Network (ETDNN) [27], which takes the 40D mel-spectrogram features as input to extract 512D x-vectors. The ETDNN model is trained on the VoxCeleb1 [28] and Vox-Celeb2 [29] datasets, for speaker identification task, to discriminate among the 7, 146 speakers. A PLDA model is trained using x-vectors extracted from a subset of VoxCeleb1 and VoxCeleb2. The x-vectors are centered and whitened using the statistics estimated from the DISPLACE Dev data. The x-vectors are also
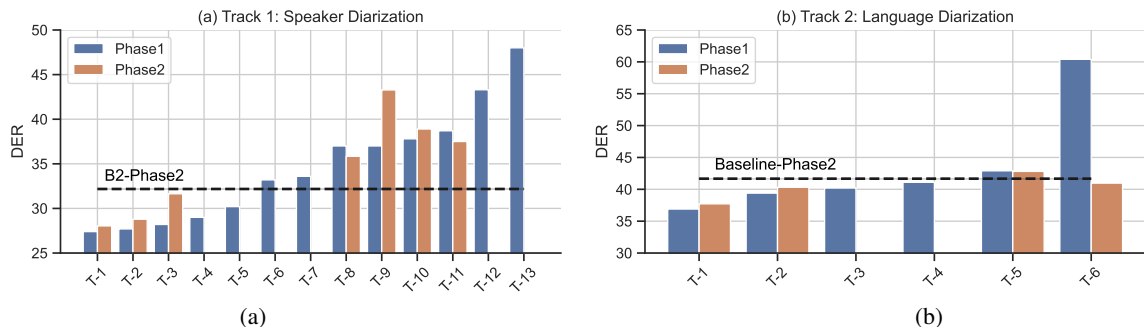
Figure 3: *Eval Phase1 (blue bars) and Phase2 (orange bars) performance (DER %) of all the participating teams for (a) speaker and (b) language diarization tracks.*

Table 2: *Baseline language diarization results for the DISPLACE development and evaluation sets (Phase-1 and Phase-2) using AHC and spectral clustering.*

Method	DER (%)		
	Dev	**Eval**	
		Phase1	**Phase2**
AHC	48.61	41.76	41.92
Spec. Clustering	48.48	41.48	41.67

unit-length normalized for PLDA training.

We have also experimented with spectral clustering (SC) [30] algorithm along with AHC. The input to the SC are processed PLDA scores, obtained by applying sigmoid with temperature scaling of 0.1. The VB-HMM is initialized separately for each recording from the clustering output. The hyper-parameter values for clustering algorithms and VB-HMM are set based on the performance on the DISPLACE Dev set. Thus, we consider two baseline systems - AHC+VB-HMM (B1) and SC+VB-HMM (B2). The DER results are reported in Table 1. It can be observed that baseline system B2 outperforms B1 on Eval Phase1 and Phase2 data relatively by 18% and 15%, respectively.

5.3. Language Diarization (LD)

Existing models for language diarization assume that the test audio recordings contain two known languages [2], as opposed to the DISPLACE dataset, which deals with a more generic case involving unknown languages. Hence, our baseline systems were redesigned for this task and used language embedding extraction at the segment level, followed by clustering.

We use the SpeechBrain [31] language identification model[1] based on the ECAPA-TDNN architecture [32] for embedding extraction. The model is trained on the Voxlingua107 dataset [33], which contains 6628 hours of speech from 107 languages, extracted from YouTube videos. We begin by segmenting the audio, containing only speech regions, into short overlapping segments of 0.4s with 0.2s shift. We extract 256 dimensional embeddings at the segment level, compute the pairwise cosine similarity score matrix, and apply a clustering algorithm to obtain the segment-level language predictions. We

---

[1] https://huggingface.co/speechbrain/lang-id-voxlingua107-ecapa

present the results of two clustering algorithms: agglomerative hierarchical clustering (AHC) and spectral clustering. This is given in Table 2.

6. Challenge Results

Figure 3 shows the Eval Phase-1 and Phase-2 DER performance of all the participating teams for speaker and language diarization tracks. For SD track, a total of 13 and 7 teams participated in the Phase-1 and Phase-2 evaluations, respectively. The top three teams for SD track have reported significant improvements over baseline for both evaluation phases (as shown in Figure 3 (a)). For Phase-1, the lowest DER obtained for the SD task is 27.4% (T-1) followed by 27.7% (T-2) and 28.2% (T-3). For Phase-2, the bottom three DER (for SD track) is 28.04% (T-1) followed by 28.79% (T-2) and 31.64% (T-3). The LD track (Track-2) received six submissions in Phase-1, out of which four teams outperformed baseline DER. In Phase-2, only four teams participated in this track. For the LD track, DER for the top three performing teams are 36.9 (T-1), 39.4 (T-2) and 40.2 (T-3), respectively for Phase-1. For Phase-2, the lowest DER obtained for the LD track is 37.72 (T-1) followed by 40.32 (T-2) and 41.67 (Baseline).

7. Summary

In this paper, we have detailed the Interspeech DISPLACE challenge aimed at fostering research on processing multi-lingual multi-speaker conversational audio. The challenge entails two tracks, i) speaker diarization and ii) language diarization. The dataset is common for the evaluation of both the tasks. The paper describes the data collection and annotation process.

Most of the previous SD challenges and existing works focus on monolingual scenarios, where a recording contains only one language. Similarly. most of the previous works on LD used single speaker recordings. To the best of our knowledge, no other publicly available dataset contains the plethora of diversity observed in the DISPLACE dataset in terms of multi-lingual, code-mixed, multi-speaker, natural conversational speech.

In this paper, we also describe the baseline system and summarize the Phase-1 and Phase-2 submissions made by the challenge participants. Even with the global scale of system development efforts, the dataset is challenging with the best results reporting a DER of 28.04% (Phase-2) for SD and 37.72% (Phase-2) for LD. These results highlight the need for more future work to enable speech technology development in natural multi-lingual conversations.

8. References

[1] B. E. Bullock and A. J. E. Toribio, *The Cambridge handbook of linguistic code-switching*. Cambridge University Press, 2009.

[2] E. Yilmaz, M. McLaren, H. van den Heuvel, and D. A. van Leeuwen, "Language diarization for semi-supervised bilingual acoustic model training," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 91–96.

[3] P. Auer, *Code-switching in conversation: Language, interaction and identity*. London: Routledge, 1998.

[4] S. V, V. Thenkanidiyoor, and D. A.D, "SVM based language diarization for code-switched bilingual indian speech using bottleneck features," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 132–136.

[5] D.-C. Lyu, E.-S. Chng, and H. Li, "Language diarization for conversational code-switch speech with pronunciation dictionary adaptation," in *2013 IEEE China Summit and International Conference on Signal and Information Processing*, 2013, pp. 147–150.

[6] ——, "Language diarization for code-switch conversational speech," in *IEEE ICASSP*, 2013, pp. 7314–7318.

[7] H. Liu, L. P. G. Perera, X. Zhang, J. Dauwels, A. W. Khong, S. Khudanpur, and S. J. Styles, "End-to-end language diarization for bilingual code-switching speech," in *Proc. Interspeech 2021*, 2021, pp. 1489–1493.

[8] NIST, "Rich transcription evaluation," Available at https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation [Online; accessed February-2023].

[9] J. S. Garofolo, C. D. Laprun, J. G. Fiscus *et al.*, *The rich transcription 2004 spring meeting recognition evaluation*. US Department of Commerce, National Institute of Standards and Technology, 2004.

[10] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, "The rich transcription 2005 spring meeting recognition evaluation," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 369–389.

[11] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*, R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 373–389.

[12] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, *First DIHARD Challenge Evaluation Plan*, 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1199638

[13] ——, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Proc. Interspeech 2019*, 2019, pp. 978–982.

[14] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "The Third DIHARD Diarization Challenge," in *Proc. INTERSPEECH*, 2021, pp. 3570–3574.

[15] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural Fearless Steps challenge: A giant leap for naturalistic audio," in *Proc. Interspeech*, 2019, pp. 1851–1855.

[16] A. Joglekar, J. H. Hansen, M. C. Shekar, and A. Sangwan, "Fearless steps challenge (fs-2): Supervised learning with massive naturalistic apollo data," in *Proc. Interspeech 2020*, 2020, pp. 2617–2621.

[17] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: The iberspeech-rtve challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, 2019.

[18] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. The 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[19] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "Voxsrc 2020: The second voxceleb speaker recognition challenge," *arXiv preprint arXiv:2012.06867*, 2020.

[20] H. Liu, H. Xu, L. P. Garcia, A. W. Khong, Y. He, and S. Khudanpur, "Reducing language confusion for code-switching speech recognition with token-level language diarization," *arXiv preprint arXiv:2210.14567*, 2022.

[21] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "SEAME: a mandarin-english code-switching speech corpus in south-east asia," in *Proc. Interspeech 2010*, 2021, pp. 1986–1989.

[22] S. Shah, S. Sitaram, and R. Mehta, "First workshop on speech processing for code-switching in multilingual communities: Shared task on codeswitched spoken language identification," in *WSTC-SMC 2020*, 2020, p. 24.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015.

[24] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs." in *Proc. Interspeech*, 2016, pp. 3434–3438.

[25] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on Bayesian HMM with eigenvoice priors," in *Proc. Odyssey*, 2018, pp. 147–154.

[26] P. Singh, H. V. M.A., S. Ganapathy, and A. Kanagasundaram, "LEAP Diarization System for the Second DIHARD Challenge," in *Proc. Interspeech*, 2019, pp. 983–987.

[27] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *IEEE ICASSP*, 2019, pp. 5796–5800.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. of INTERSPEECH*, pp. 2616–2620, 2017.

[29] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. of INTERSPEECH*, 2018, pp. 1086–1090.

[30] Q. Lin, R. Yin, M. Li, H. Bredin, and C. Barras, "LSTM Based Similarity Measurement with Spectral Clustering for Speaker Diarization," in *Proc. of Interspeech*, 2019, pp. 366–370.

[31] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[32] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[33] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.