Spoken Language Diarization using Source-Domain Aggregative Weakly-Supervised Language Labels

Spandan Dey, Hirak Mondal, Anmol Jain, Shirish Kumar, Premjeet Singh, Suyash Pandya
Samsung R&D Institute India-Bangalore
Bengaluru, Karnataka, India
{spandan.dey, hirak.mondal, anmol74.jain, shirish.k6, premjeet.sn, suyash.p1}@samsung.com

***Abstract*—Spoken language diarization (SLD) detects homogeneous language segments in utterances containing code-switching or code-mixing. Although SLD is influenced by speaker diarization (SD), unlike speaker labels, language labels belong to a finite and defined set, enabling more structured modeling. Leveraging this, we propose to improve SLD by developing a domain-specific spoken language identification (LID) front-end directly from source-domain development data, where only rich transcript time marked (RTTM) annotations are available. Focusing on Track-2 of the DISPLACE 2024 challenge, we introduce an aggregative RTTM to pseudo-language label mapping (ARPM) framework using the Whisper-large-v3 model. ARPM generates weakly supervised language labels by aggregating language change tags from RTTM ground-truths, filtering them with duration and entropy-based criteria, and then refining the labels using pivot language based refinement. The resulting LID model produces more reliable domain-specific language-aware representations for subsequent diarization stages. The proposed approach achieves a 4.58% absolute improvement in diarization error rate (DER) over the baseline. Despite using only 20 hours of development data and a standalone framework, we achieve competitive performance with top-performing submissions to the challenge, leveraging thousands of hours of training data and system fusion.**

***Index Terms*—Language diarization, code-switching, DISPLACE 2024, Whisper, weakly supervised learning.**

I. Introduction

Spoken language diarization (SLD) systems aim to detect homogeneous language segment boundaries in utterances where speakers frequently switch languages. These phenomena are linguistically termed *code-switching* (using interchanging languages at sentence or phrase level) and *code-mixing* (using interchanging languages at word level within a sentence) [1]. These shifts in speaking patterns are becoming a common part of today's growing cosmopolitan, multilingual societies. However, most spoken language identification (LID) systems used in multilingual human-to-computer interaction (HCI) are trained on homogeneous utterances and do not account for intra-utterance language changes. This limits their robustness in practical scenarios, motivating the need for SLD as a front-end component in modern speech systems [2], [3].

SLD has largely evolved by adapting methodologies from speaker diarization (SD), following a similar pipeline of voice activity detection (VAD), segment-level representation learning, clustering, and generating the predicted rich transcript time marked (RTTM) as output. Early SLD studies rely on synthetically creating code-switching datasets by stitching homogeneous language utterances from multilingual data [1], [4]. Gradually, natural conversational datasets for SLD research are evolving. In the past few years, the DISPLACE initiative [5], [6] has been one of the leading platforms providing naturally occurring multilingual conversation with code switching and mixing. From developmental aspects, initially, researchers focused on short-duration LID [7], [8] or language change point detection-based approaches for solving the diarization problem [9]. Later on, inspired by the speaker diarization, majority of the recent SLD approaches follow the multi-stage hybrid approach [10].

Our study focuses on the language diarization task (Track-2) of the DISPLACE 2024 [1] challenge (DISPLACE24) [6]. At first, we analyze the corresponding top-performing submissions and recent state-of-the-art approaches. The winning system by Kalada et al. [11] employs multi-stage LID fine-tuning of a wav2vec2-BERT model using large-scale corpora such as VoxLingua107 [12] (6628 hours) and NIST LRE [13] (3470 hours), followed by agglomerative hierarchical clustering (AHC) of 512-dimensional embeddings. The second-place system [14] uses wav2vec2.0-XLSR [15] with spectral clustering and Dover-lap [16] fusion. Similarly, [2] enhances Whisper [17] LID scores with accent modeling and multi-resolution fusion.

While effective, we observe that these frameworks relied heavily on large (even thousands of hours) amounts of multilingual training data and an aggressive fusion process across multiple complex systems. Contrary to that trend, we develop a standalone, less resource-intensive solution that performs comparably to these leading solutions. To develop such an effective SLD framework, we utilize the fundamental difference between speaker diarization and language diarization: Unlike speaker identities, which are open and indefinite, language labels belong to a finite and closed set. This enables more structured modeling and opens the possibility of learning domain-specific representations directly from the target domain, even though the language labels for them are not explicitly provided.

Instead of relying on a generic language encoder, we estimate pseudo-language labels directly from segmented development data and use them to train a weakly supervised domain-specific LID model. Since standard diarization protocols provide only RTTM annotations with language change

[1]DISPLACE 2026 challenge is not concluded at the time of writing this manuscript. Also, it does not cover a dedicated SLD track for benchmarking.

indicators (e.g., L1, L2) and no explicit language labels, we propose an *aggregative RTTM to pseudo-language label mapping* (ARPM) framework using a pre-trained Whisper-large-v3 model [17]. While related ideas have been explored for speaker diarization [18], [19], such approaches remain underexplored in SLD. Our domain-specific language encoder is finetuned on fewer languages while capturing the non-lingual characteristics [20], which are likely to be present in the application domain (e.g., DISPLACE 24). This shift enables a lightweight and effective diarization framework that achieves competitive performance using only 20 hours of development data without system fusion. Following this, the major contributions of this work are summarized as follows:

- We propose a weakly supervised SLD framework to learn domain-specific language-discriminating representations directly from RTTM annotated development data, eliminating the need for external labeled data for domain-specific supervision.
- We introduce ARPM, a reliability-aware mapping strategy that establishes a one-to-one correspondence between RTTM language change tags and underlying spoken languages, enhanced with duration and entropy-based filtering.
- We propose a pivot language based label-space refinement to improve label consistency and reduce noise in weak supervision.
- We further suggest combining domain-specific and generic LID representations for adapting our ARPM-based SLD framework to any generic application domain.

II. Database and baseline framework description: DISPLACE 2024

Database: We use the DISPLACE24 Track-2 development set (Dev) data. It contains 35 multilingual single-channel conversational recordings, totaling 20 hours and 98 speakers. Each recording involves three to five participants conversing in their native languages with naturally occurring code-switching or code-mixing, predominantly with Indian-accented English. We evaluate the frameworks on the evaluation-set (Eval) of the Track-2. It contains 32 recordings in similar recording environments, totaling 18 hours of speech. All audio is provided in `.wav` format at a sampling rate of 16 kHz.

Baseline: The baseline follows a multi-stage hybrid diarization pipeline. First, Pyannote VAD [21] is applied to remove non-speech segments. The speech is then segmented into 10 s windows with 200 ms shifts and passed through the Whisper-base model [17] to obtain 99-dimensional LID posterior scores. Segment-level similarities are computed using cosine distance and clustered using agglomerative hierarchical clustering (AHC) with average linkage. Finally, cluster assignments are refined using variational Bayes (VBx) [22] to generate the output RTTMs. For further details, we refer readers to [6].

III. Methodology

We first optimize key hyperparameters of the baseline framework, and subsequently build the proposed ARPM-based weakly supervised SLD system on top of this optimized setup.

Baseline-optimized: We explore window lengths from 2 s to 12 s and shift lengths of 50 ms, 200 ms, and half of the window size. We also evaluate temperature scaling of the Whisper-base LID posterior scores in the range of 0.5 to 5.0. For the clustering stage, we tune key AHC hyperparameters, including GMM-based thresholds, linkage methods, and VBx settings to control cluster granularity. The final configuration is selected based on the lowest DER on the Dev set, resulting in a 5 s window, 400 ms shift, and temperature of 0.5. These optimizations provide a strong baseline for the subsequent ARPM-based improvements.

A. Aggregative RTTM to pseudo-language label mapping

Motivation: A key component of hybrid SLD pipelines is the LID model, which acts as a front-end language encoder before the clustering. In domain-specific scenarios, a tailored LID model can provide more reliable representations than a generic one. This can be achieved by adapting a large-scale pre-trained LID model using domain-specific data. However, a major challenge in SLD is the absence of explicit language labels in the development set; only RTTM annotations indicating language change boundaries are available. To address this, we propose an *aggregative RTTM to pseudo-language label mapping* (ARPM) framework that derives weak supervision directly from RTTMs. Unlike prior empirical mapping approaches [14], which concatenate segments with identical RTTM tags, ARPM accounts for (i) turn-level duration variability and (ii) distortions due to noise, overlap, or fillers [8], resulting in more reliable pseudo-label estimation.

Stage-1: Basic ARPM - Let $i = 1 : N$ denote utterances in the development set, where the i-th utterance contains N_i homogeneous segments indexed by j. Utilizing the time-stamp information from the RTTM, we can extract each homogeneous segment s_{ij} and compute its duration d_{ij}. The corresponding RTTM hidden-language tags (h_{ij}, usually mentioned as $L1$, $L2$, $\cdots$) for the i-th utterance is collectively denoted with the set $\{\mathcal{H}_i\}$. We can feed s_{ij} to a pre-trained LID model $\mathcal{F}$ (recognizing N_L number of languages) to get the prediction score $p_{ij} \in \mathbb{R}^{N_L}$ and the corresponding predicted language label l_{ij}. Considering p_{ij} $\forall j$, let the set of unique predicted languages in utterance i be $\{L_i\}$. We then define an *association matrix* $\mathbf{C}_i \in \mathbb{R}^{|\mathcal{H}_i| \times |L_i|}$ whose each element can be computed as:

$$\mathbf{C}_i(m,n) = \sum_{j=1}^{N_i} \mathbb{1}\left\{h_{ij} = \mathcal{H}_i[m], l_{ij} = L_i[n]\right\} \tag{1}$$

To mitigate turn imbalance, we further normalize $\mathbf{C}_i$ row-wise:

$$\tilde{\mathbf{C}}_i(m,n) = \mathbf{C}_i(m,n) / \sum_{n=1}^{|L_i|} \mathbf{C}_i(m,n) \tag{2}$$

We then formulate a mapping function (f_i) that estimates the maximum-likelihood pseudo-language label $\hat{l}_{ij}$ from the association matrix and corresponding segment's RTTM label h_{ij}:

$$f_i(h) = \arg\max_n \tilde{\mathbf{C}}_i(h,n) \tag{3}$$

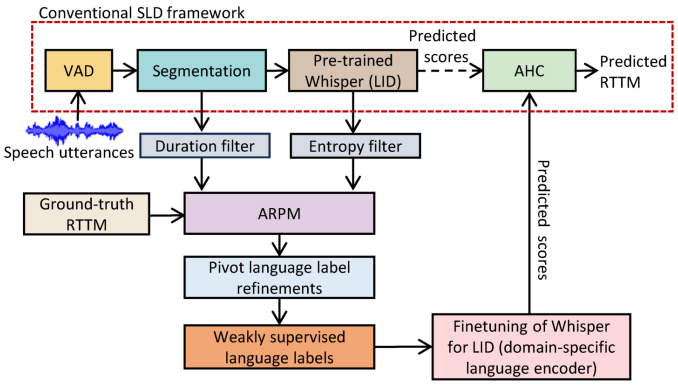

Fig. 1. Overview of the ARPM-based diarization framework. Here, AHC stands for agglomerative hierarchical clustering. The dotted arrow after Pre-trained Whisper (LID) indicates optional concatenation of generic LID prediction scores with domain-specific LID prediction scores (cf. Section V-D).

$$\hat{l}_{ij} = L_i[f_i(h_{ij})] \tag{4}$$

Eq. 3 and Eq. 4 can be interpreted as estimating the conditional mapping $p(l \mid \mathcal{H})$ under scenarios when the explicit language labels are not available. Finally, considering all the N Dev utterances, the ARPM is used to generate a weakly supervised LID dataset with $|\hat{L}|$ estimated languages, where $\hat{L} = \bigcup_{i=1}^{N} \bigcup_{j=1}^{N_i} \left\{\hat{l}_{ij}\right\}$.

Stage-2: ARPM with reliability-aware filtering - We further improve the reliability of the ARPM-generated weakly supervised language labels. During a conversation, certain turns (homogeneous language segments) can be of very short duration, making their predicted language label unreliable [23]. Similarly, we need to carefully discard certain outlier turns heavily contaminated by dominant transient noise or non-linguistic events. For such segments, often the LID model produces less confident high entropy predictions [20]. Even the RTTM tags labeled by human annotators may not always be reliable in such cases. To ensure the mapping is reliable, it is important to identify and discard all such segments during association matrix aggregation. Concerning the minimum input duration required for reliable LID performance [23], we experiment with applying different duration filters with thresholds (d_{th}) of 0.5 s, 1 s, 2 s, and 3 s. Segments shorter than the chosen threshold are discarded when computing $\mathbf{C}$.:

$$\mathbf{C}_i(m,n) = \sum_{j=1}^{N_i} \mathbb{1}\left[\{h_{ij} = \mathcal{H}_i[m], l_{ij} = L_i[n]\} \;\&\&\; (d_{ij} > d_{\text{th}})\right] \tag{5}$$

Simultaneously, to address the noisy turns, if the entropy (e_{ij}) of a segment's LID prediction score is higher than an empirically selected threshold (e_{th}), we do not consider such segments in ARPM. The computation of e_{ij} is presented in Eq. 6. Further, when $N_L >> |\hat{L}|$, then the entropy-based filtering can also help in discarding utterances with false positives to spurious language labels.

$$e_{ij} = -\sum_{k=1}^{N_L} p_{ij}[k]\, log[p_{ij}[k]] \tag{6}$$

Algorithm 1 Aggregative RTTM to pseudo-language label mapping (ARPM) with pivot-based label refinement.

Input: Utterances $\{U_i\}_{i=1}^{N}$, RTTMs $\{\mathcal{H}_i\}_{i=1}^{N}$, LID model $\mathcal{F}$.
Output: Weakly supervised labels $\hat{L}$ and $\hat{L}_{\text{pivot}}$.

for $i = 1 \rightarrow N$ **do**
  for $j = 1 \rightarrow N_i$ **do**
    Extract segments s_{ij} with duration d_{ij} using $\mathcal{H}_i$.
    $p_{ij} = \mathcal{F}(s_{ij})$ ▷ Posterior score computation
    Compute entropy e_{ij}.
    if $(d_{ij} < d_{\text{th}}) \;\vee\; (e_{ij} > e_{\text{th}})$ **then**
      discard segment and continue.
    else
      $l_{ij} = \arg\max p_{ij}$
  Define $L_i \leftarrow$ unique $\{l_{ij}\}\ \forall\ j$.
  Build $\mathbf{C}_i(m,n) = \sum_{j=1}^{N_i} \mathbb{1}\{h_{ij} = \mathcal{H}_i[m], l_{ij} = L_i[n]\}$
  Normalize: $\tilde{\mathbf{C}}_i(m,n) = \mathbf{C}_i(m,n) / \sum_n \mathbf{C}_i(m,n)$
  Map: $\hat{l}_{ij} = L_i\left[\arg\max_n \tilde{\mathbf{C}}_i(h_{ij}, n)\right] \quad \forall j$
Resolve conflict: If two RTTM tags map to the same language, assign the dominant RTTM tag (higher turn count) to the shared language, and map the other to its next most likely language.
Weakly supervised labels: $\hat{L} = \bigcup_{i,j}\{\hat{l}_{ij}\}$.
Select pivot languages $\{l_p\}$ and define $\hat{L}_{\text{pivot}} = \{l_p\} \cup \{\text{other}\}$.

B. Pivot language based ARPM refinements

Upon analyzing the histogram of the weakly supervised language labels $\hat{L}$, we observe that they typically exhibit a long-tail distribution, where a few dominant languages account for most segments, while there exist several languages with strikingly low frequency of segments. We suspect these spurious languages often arise from the misclassification of the backbone LID model $\mathcal{F}$ (for our case Whisper-large-v3). Hence, considering all the estimated classes in $\hat{L}$ for language encoder finetuning can lead to heavy class imbalance.

Following this, we make further refinements in the ARPM-generated labels using *pivot language based label-space refinement*. Based on the histogram of $\hat{L}$, we select the top-1 or top-2 dominant languages as pivot set $\{l_p\}$. All remaining labels are grouped into an "other" category, resulting in a reduced label set $\hat{L}_{\text{pivot}}$ with cardinality $|l_p| + 1$. This refinement improves label consistency and stabilizes representation learning.

The overall ARPM procedure, including reliability filtering and pivot refinement, is summarized in Algorithm 1.

IV. Experiment Details

Framework: We follow a hybrid strategy inspired by the DISPLACE24 baseline with modified hyperparameters (cf. Fig. 1). Here, the VAD-processed audio is segmented into 5 s windows with 0.4 s shifts. These segments are passed through our ARPM-based weakly supervised LID encoder to

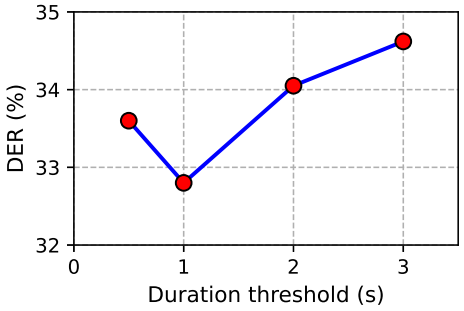

Fig. 2. Dev set diarization performances (DER shown in range of 32-35%) for different choices of duration thresholds in ARPM.

generate prediction scores. These prediction scores are used for the AHC stage. Performance is evaluated using diarization error rate (DER) and Jaccard error rate (JER), following the challenge protocol.

LID encoder training: We finetune the last two encoder and the classification layers of the *Whisper-large-v3* model [17] for developing the LID encoder. The Dev set utterances are first segmented into 3 s audio chunks. These chunks are then split into utterance-disjoint training and validation parts using 90:10 ratio. After discarding the non-speech and silence segments using VAD, we use the remaining segments' ARPM-generated weakly supervised language labels for LID finetuning. The model is finetuned for 50 epochs with a batch size of 16 and learning rate of 1×10^{-4} using cross-entropy loss and the Adam optimizer. The best checkpoint, selected based on validation loss, is used to generate prediction scores for the clustering stage.

V. Experiment results

A. Different aspects of developing the ARPM-based SLD framework: Progressive performance analysis

Using the DISPLACE24 Dev set, we first obtain ARPM-generated language labels $\hat{L}$ corresponding to: English, Hindi, Kannada, Malayalam, and Telugu. These labels are used to finetune the Whisper-large-v3 model, producing $p \in \mathbb{R}^{1\times5}$ score vectors for AHC clustering. Similarly, we also train another language encoder using pivot-refined labels $\hat{L}_{\text{pivot}}$.

Optimal choice of duration threshold in ARPM: To mitigate the effect of unreliable short segments, we evaluate duration thresholds of 0.5s, 1s, 2s, and 3s, discarding segments shorter than the threshold during label generation. To illustrate the impact of this decision threshold, in Fig. 2, we show the Dev set DERs for the respective diarization frameworks. To note, except for the final performance reported in Table II, we do not disclose the Eval set when selecting different hyperparameters. From Fig. 2, we observe that a threshold of 1s yields the best performance. Lower thresholds retain ambiguous short segments, while higher thresholds (2s–3s) discard a significant portion of useful data, degrading performance.

Effect of pivot language based label refinement: Based on dataset characteristics [6] and label histograms, English emerges as the dominant pivot language, with Hindi as a secondary candidate. We evaluate both dichotomy (En-rest, Hi-rest) and trichotomy (En-Hi-rest) settings for LID encoder development. As shown in Table I, the English-based dichotomy provides the best performance, indicating that a dominant pivot improves label consistency and clustering reliability.

TABLE I
Dev set diarization performances with different pivot languages (without ARPM filters): English (En), Hindi (Hi).

Metric	En - rest	Hi - rest	En - Hi - rest
DER (%)	33.82	36.11	34.46
JER (%)	52.80	57.68	53.82

TABLE II
Overall diarization performances of the proposed approaches.

Approach	Framework	Dev set		Eval set	
		DER (%)	JER (%)	DER (%)	JER (%)
Baseline	DISPLACE24 Baseline	41.10	61.88	34.32	63.75
	Baseline-optimized	40.05	60.95	34.33	61.85
ARPM	No filter	33.60	54.00	31.72	56.47
	With filters	32.80	52.29	32.05	56.96
	With filters, Pivot: En	32.60	52.09	30.44	55.06
Concatenate LID scores ARPM-filter, Pivot: En		**31.44**	**50.59**	**29.74**	**51.58**

B. Overall performance of the ARPM-based SLD framework

Table II presents the final performance on both Dev and Eval sets. Compared to the DISPLACE24 baseline and its optimized variant, the proposed ARPM-based framework achieves consistent performance improvements. Applying filtering to basic ARPM improves diarization performance by potentially eliminating challenging turn segments during the language encoder training. Pivot language based refinements further improve the diarization performance. The best performance is obtained by concatenating prediction scores from ARPM-filtered ($p \in \mathbb{R}^{1\times5}$) and pivot-refined ($p \in \mathbb{R}^{1\times2}$) language encoders, resulting in a combined representation of $p \in \mathbb{R}^{1\times7}$.

C. Comparison with related literature

We compare our approach with the top-2 submissions to the DISPLACE 2024 language diarization track. These systems rely on large-scale multilingual training (thousands of hours) and, in some cases, multi-system fusion using Dover-lap [16]. In contrast, we develop a significantly less resource-demanding standalone diarization framework using only 20 hours of development data, learning a domain-specific language encoder via ARPM-based weak supervision. The results in Table III demonstrate that ARPM can effectively extract reliable supervision from RTTM annotations, enabling competitive diarization performance to the top submissions without extensive data or system complexity.

To further isolate the contribution of ARPM from the merits of using Whisper-large-v3 as LID backbone, we perform diarization with the pre-trained Whisper-large-v3 model without adaptation, using its native LID predictions ($p \in \mathbb{R}^{1\times99}$) for clustering. As reported in Table III, the results clearly distinguish the contribution of ARPM-based weak supervision alone, independent from the choice of LID backbone.

TABLE III
COMPARISON OF OUR ACHIEVED PERFORMANCES WITH THE DISPLACE24 TOP-SUBMISSIONS.

Approach	Development data (h)	Fusion applied	Eval set DER (%)
Baseline	20	No	34.32
Winner [11]	> 3000	No	27.60
Runners' up [14]	> 3000	Dover-lap	30.99
Whisper-large-v3 encoder	NA	No	32.68
Proposed	20	No	29.74

TABLE IV
EVAL SET DIARIZATION PERFORMANCES WITH DIFFERENT DIMENSIONS OF LANGUAGE ENCODER PREDICTION SCORES AS INPUTS TO CLUSTERING STAGE. FOR NON-TARGET DOMAINS, WE SUGGEST CONCATENATION OF THE PROPOSED ARPM+PIVOT (EN) REFINED PREDICTION SCORES WITH THE PRE-TRAINED WHISPER-LARGE-V3 LID SCORES.

Systems	Pre-trained Whisper	Proposed	Concatenated
Score dimension	99	7	106
DER (%)	32.68	**29.74**	32.45

D. Utilizing our framework for generic domains

While ARPM is designed for domain-specific adaptation, it can be extended to general scenarios by combining domain-specific and generic representations. Specifically, we concatenate the 99-dimensional pre-trained Whisper-large-v3 LID scores with the 7-dimensional ARPM-based LID scores (with filtering and pivot refinement), resulting in a 106-dimensional representation for clustering. Table IV shows that this hybrid approach achieves competitive performance, indicating that ARPM provides complementary information to generic LID models and can generalize beyond the target domain.

VI. CONCLUSIONS

This work presents a weakly supervised framework for developing a domain-specific language encoder to improve spoken language diarization, evaluated on Track-2 of the DISPLACE 2024 challenge. To estimate language labels from RTTM annotations, we propose the *aggregative RTTM to pseudo-language label mapping* (ARPM) approach, incorporating duration and entropy-based filtering along with pivot language based label refinement. Despite being significantly less resource-intensive, the proposed method achieves competitive performance with top challenge submissions.

While the current study is restricted by the limited availability of large-scale, annotated code-switching corpora, the results consistently demonstrate the effectiveness of leveraging RTTM-based weak supervision for domain-specific adaptation, along with provisions for potential extension to generic domains. Future work will focus on extending this framework to more diverse multilingual settings and exploring its applicability to broader real-world conversational scenarios.

REFERENCES

[1] J. Mishra, J. N. Patil, A. Chowdhury, and M. Prasanna, "End to end spoken language diarization with wav2vec embeddings," in *INTERSPEECH*. ISCA, 2023, pp. 501–505.

[2] B. Vachhani, D. Singh, and R. Lawyer, "Multi-resolution Approach to Identification of Spoken Languages and To Improve Overall Language Diarization System Using Whisper Model," in *INTERSPEECH*. ISCA, 2023, pp. 1993–1997.

[3] J. Mishra and S. M. Prasanna, "Implicit self-supervised language representation for spoken language diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3393–3407, 2024.

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

[5] S. Baghel, S. Ramoji, Sidharth, R. H, P. Singh, S. Jain, P. Roy Chowdhuri, K. Kulkarni, S. Padhi, D. Vijayasenan, and S. Ganapathy, "The DISPLACE challenge 2023 - DIarization of SPeaker and LAnguage in Conversational Environments," in *INTERSPEECH*. ISCA, 2023, pp. 3562–3566.

[6] S. B. Kalluri, P. Singh, P. Roy Chowdhuri, A. Kulkarni, S. Baghel, P. Hegde, S. Sontakke, D. K T, S. M. Prasanna, D. Vijayasenan, and S. Ganapathy, "The Second DISPLACE Challenge: DIarization of SPeaker and LAnguage in Conversational Environments," in *INTERSPEECH*. ISCA, 2024, pp. 1630–1634.

[7] H. Liu, L. P. G. Perera, A. W. Khong, E. S. Chng, S. J. Styles, and S. Khudanpur, "Efficient self-supervised learning representations for spoken language identification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1296–1307, 2022.

[8] S. Dey, H. Mondal, and S. K. Kurmi, "Teacher-free knowledge distillation for improving short-utterance spoken language identification," in *INTERSPEECH*. ISCA, 2025, pp. 1483–1487.

[9] D.-C. Lyu, E.-S. Chng, and H. Li, "Language diarization for code-switch conversational speech," in *ICASSP*. IEEE, 2013, pp. 7314–7318.

[10] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[11] J. Kalda, T. Alumae, M. Lebourdais, H. Bredin, S. Baroudi, and R. Marxer, "TalTech-IRIT-LIS Speaker and Language Diarization Systems for DISPLACE 2024," in *INTERSPEECH*. ISCA, 2024, pp. 1635–1639.

[12] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[13] Y. Lee, C. Greenberg, E. Godard, A. A. Butt, E. Singer, T. Nguyen, L. Mason, and D. Reynolds, "The 2022 NIST language recognition evaluation," in *INTERSPEECH*, 2023, pp. 1928–1932.

[14] E. Ausev, V. Volokhov, S. Novoselov, V. Marchevskiy, E. Shangina, and A. Logunov, "ITMO language diarization and identification systems for the DISPLACE 2024 challenge," in *ICASSP*. IEEE, 2025, pp. 1–5.

[15] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[16] D. Raj, L. P. Garcia-Perera, Z. Huang, S. Watanabe, D. Povey, A. Stolcke, and S. Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 881–888.

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[18] J. Ao, M. S. Yıldırım, R. Tao, M. Ge, S. Wang, Y. Qian, and H. Li, "USED: Universal speaker extraction and diarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 96–110, 2025.

[19] S. Wu, J. Du, M.-K. He, S. Niu, H. Chen, H. Tang, and C.-H. Lee, "Semi-supervised multi-channel speaker diarization with cross-channel attention," in *ASRU*. IEEE, 2023, pp. 1–8.

[20] S. Dey, M. Sahidullah, and G. Saha, "Towards cross-corpora generalization for low-resource spoken language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 5040–5050, 2024.

[21] H. Bredin *et al.*, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP*. IEEE, 2020, pp. 7124–7128.

[22] D. Klement, M. Diez, F. Landini, L. Burget, A. Silnova, M. Delcroix, and N. Tawara, "Discriminative training of VBx diarization," in *ICASSP*. IEEE, 2024, pp. 11 871–11 875.

[23] P. Shen, X. Lu, S. Li, and H. Kawai, "Knowledge distillation-based representation learning for short-utterance spoken language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2674–2683, 2020.