



Ensemble of Incremental System Enhancements for Robust Speaker Diarization in Code-Switched Real-Life Audios

Raj Gohil, Ramya Viswanathan, Saurabh Agrawal, C. M. Vikram,
Madhu R. Kamble^(✉), Kamini Sabu, M. Ali Basha Shaik,
and Krishna K. S Rajesh

Samsung R & D Institute Bangalore, Bangalore, India
{raj.gohil,r.vishwanath,saurabh.a,vikram.cm,madhu.r,
kamini.sabu,m.shaik,ks.rajesh}@samsung.com

Abstract. Identifying individual speaker utterances in overlapped multi-speaker conversations pose a challenging problem in speaker diarization, specifically under multi-lingual scenarios. Standard speech diarization the system consists of a speech activity detector, a speaker-embedding extractor followed by clustering. We improve each of these components from the standard pipeline to enhance the speaker diarization in such complex cases. Our investigation focuses on addressing key sub-aspects of the task like the presence of noise variations, utterance duration variations, inclusion of enhanced ECAPA-TDNN embeddings for robustness etc. Finally, we use the DOVER-LAP approach to combine these system predictions so that complementary advantages of individual systems are efficiently incorporated. Our best-proposed systems outperform the baseline by achieving DER of 27.7% and 28.6% on Phase-1 and Phase-2 of Track-1 blind evaluation sets, respectively.

Keywords: Speaker diarization · ECAPA-TDNN · Spectral clustering

1 Introduction

In today's digital world, most of our communications and meetings tend to be online. In many applications like doctor visits, counsellor sessions, teacher-child interactions and customer support calls, it is necessary to know the time durations where each of the two parties are conversing. Precise time durations are one of the essential requirements in conversational scenarios to detect robust end-point detection [11], to generate high-quality transcription using automatic speech recognition [22] and to process using natural language understanding [24] and speech-to-speech translation [26]. In these cases, it is also important to label speech regions with the corresponding speakers to generate further enriched transcriptions. The segmentation of audio recordings by speaker labels, known as

speaker diarization, is the process of recognizing “*who spoke when*” [21]. Diarization is considered as a major task in conversational AI systems and has applications in the processing of telephone conversations, broadcast news, meetings, clinical recordings, etc. [4, 6, 21].

An overview of speaker diarization system is shown in Fig. 1. It consists of speech activity detector (SAD), speaker embedding extractor and clustering technique.

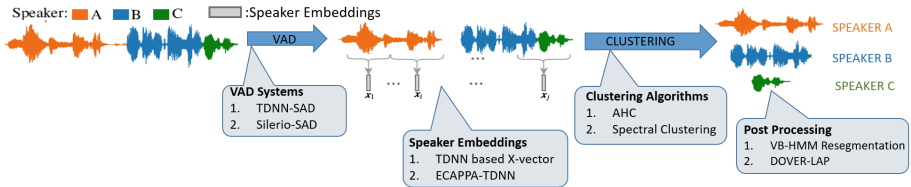


Fig. 1. Overview of Speaker Diarization System.

Recently, deep learning techniques are being widely used for speaker diarization tasks. [30] proposed a deep neural network (DNN) with fully-connected hidden layers to classify all speakers in the training set, and then use bottleneck features as a speaker representation. Later, D-vectors were improved by a long short-term memory (LSTM) [10] network with a triplet loss function [13]. An improved version of D-Vectors with TDNN architecture and a statistical pooling layer was proposed in [6] and this work was further improved by generating robust speaker representations as X-Vectors in [28]. Emphasized Channel Attention, Propagation and Aggregation Time delay Neural networks (ECAPA-TDNN) were proposed [7], which is an enhanced structure based on X-Vectors’ network. The basic TDNN layers are replaced with 1D-Convolutional Layers [9] and Res2Net-with-Squeeze-Excitation (SE-Res2Net) Blocks [9, 12, 14], while the basic statistical pooling layer is replaced with an Attentive Statistical Pooling. The ECAPA-TDNN system outperformed a strong X-Vectors baseline system as experimented in both speaker verification task and speaker diarization task [6, 7]. Although all these approaches tried to address speaker diarization in clean conditions, however, challenges remain open under noisy and speech-overlapping conditions. Recently, deep learning-based end-to-end speaker diarization approaches are also proposed to solve the issue of overlapping speech [21].

Alternatively, most of the speaker diarization systems in the literature are developed by considering monolingual recordings. When a speaker speaks in multiple languages then the diarization becomes more challenging than the monolingual cases. In code-switched conversational speech, it is trivial that a single speaker could speak in multiple languages [19]. In this case, diarization becomes more complex as both the language and speaker compete during clustering [32]. The higher variance among the languages along with the speakers also poses challenge for the speaker diarization task [32].

In this paper, we propose a system for speaker diarization in multilingual code-switched scenarios for Track-1 of the DISPLACE 2023 Challenge. We start with the baseline architecture and improve each of its components as more robust substitutes. We use Silero VAD for improving performance in noisy and reverberant conditions. We validate the robustness of improved ECAPA-TDNN embeddings over X-vector variations for speaker diarization in the presence of multilingual code-switched data. We also observe that speaker clustering works much better than AHC for speaker diarization tasks. Observing that these incremental system enhancements improve the overall system performance for individual key aspects of the task, we combine these system outputs for final predictions on the evaluation set.

The remainder of this paper is organized as follows. Section 2 describes the Track-1 DISPLACE challenge dataset and the evaluation metric used during the system development. In Sect. 3, the technical details of our system are discussed. Experimental results and discussion are detailed in Sect. 4 along with case-by-case analysis. Finally, Sect. 5 provides the main conclusions of this work along with the future work directions.

2 DISPLACE Challenge Overview

In this section, we briefly describe the Diarization of Speaker and Language in Conversational Environments (DISPLACE) Challenge [2] details. The challenge aims to detect and label all speaker or language segments automatically in each conversation. It features two tracks: Track-1 focuses on speaker diarization in multilingual scenarios, while Track-2 focuses on language diarization in multi-speaker settings.

Track-1 aims to perform speaker diarization (*“who spoke when”*) in multilingual conversational audio data, where the same speaker speaks in multiple code-mixed and/or code-switched languages. On the other hand, track-2 aims to perform language diarization (*“which language was spoken when”*) in multi-speaker conversational audio data, where the same speaker speaks in multiple languages within the same recording. We participated in the Track-1 speaker diarization challenge.

2.1 Challenge Dataset

The development set provided by the challenge was recorded in far-field conditions. The development and evaluation set consist of real-life multilingual, multi-speaker conversations. Each conversation is around 30 to 60 min long involving 3 to 5 participants. The participants show good proficiency in Indian languages along with English (though English is often observed to use the L1 accent). The development and evaluation set consists of approximately 15.5 h (27 recordings) and 16 h (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 h, and the Phase-2 evaluation set consisted of the full evaluation set.

The data was collected using a close-talking microphone worn by each speaker as well as a far-field microphone. The latter was provided to the participants for working on the challenge, while the organizers marked the ground truth using the close-talking microphone. The data contains natural code-mixing, code-switching, a variety of language dialects, reverberation, far-field effects, speaker overlaps, short turns, and short pauses.

The evaluation set features unseen languages as well. Participants were encouraged to use any publicly available datasets for training and developing the diarization systems.

2.2 Evaluation Metric

The performance metric is the diarization error rate (DER) calculated with overlap (the speech segments with multiple speakers speaking simultaneously are included during the evaluation) and without collar (tolerance around the actual speaker boundaries). Only the speech-based speaker activity regions are considered for evaluation. DER consists of three components: false alarm (FA), missed detection (Miss), and speaker confusion, among which FA and Miss are mostly caused by VAD errors. DER is defined as:

$$DER = \frac{D_{FA} + D_{miss} + D_{error}}{D_{total}} \quad (1)$$

where, D_{FA} is the total duration of wrongly detected non-speech, D_{miss} refers to the duration of wrongly detected speech, D_{error} refers to the duration of wrong speaker labeling, while D_{total} refers to the total speech duration in the given utterance.

3 Speaker Diarization System

This section explains the baseline system and the proposed system architectures in detail.

3.1 Core System

The core of the speaker diarization baseline is largely similar to the Third DIHARD Speech Diarization Challenge [23]. It uses basic components: speech activity detection, front-end feature extraction, X-vector extraction, and PLDA scoring followed by AHC. SAD is a TDNN model based on the Kaldi Aspire recipe (“egs/aspire/s5”). The speech intervals detected by the SAD are split into 1.5-sec windows with 0.25-sec shifts. For every window, 30-dimensional Mel Frequency Cepstral Coefficients (MFCCs) are computed with 25 ms window length and 10 ms hop. These are used to extract X-vectors at every 0.25 sec. The network used for X-vector extraction is the BigDNN architecture reported

in [31] instead of the DNN network used in [23]. The X-vectors are centred and whitened every 3-sec using statistics estimated from the DISPLACE Dev set part 1. These vectors are then grouped into different speaker clusters using AHC (Agglomerative Hierarchical Clustering) and a similarity matrix produced by scoring with a Gaussian PLDA (Probabilistic Linear Discriminant Analysis) model. Finally, the speaker and non-speech labels are aligned temporally with the utterance waveform. The labels are further refined using Variational Bayes Hidden Markov Model (VB-HMM) and as Universal Background model-Gaussian Mixture Model (UBM-GMM). X-vector extractor as well as UBM-GMM and total variability matrix used for resegmentation are trained on VoxCeleb I and II [5, 20] augmented with additive noise and reverberation.

3.2 Speech Activity Detection (SAD)

In our experiments, we investigate the use of TDNN-based SAD used in the baseline system [23], Silero VAD [29] and LSTM-based VAD [25]. The open-source Silero VAD [29] is trained on a large amount of data from over 100 languages and various background noises and reverberation conditions. It uses CNNs and transformers. It has been known to perform better than conventional VAD approaches in challenging noisy conditions both in terms of both precision and recall [29]. The model is trained using 30 ms frames and can also handle short frames without performance degradation.

Furthermore, we also evaluate our performance using the 2-layer LSTM VAD [25] system that predicts speech or non-speech decisions at frame-level. The system uses 20ms long frames to compute the input features: log energies of six frequency bands in the range 80 Hz to 4 kHz. The decisions may indicate some spurious unlikely short spurts of speech/silence. These are removed through post-processing where every speech region is expected to be at least 100 ms and every silence region is expected to be at least 200 ms.

3.3 Speaker Embeddings

Besides the X-vector used in the baseline, we try a different variation of the X-vector reported in [28] which is trained for the speaker verification task. In particular, we use the improved ECAPA-TDNN embeddings inspired by our previous work [6].

X-Vector. We used the X-vector described in the baseline system. This has been trained on the Voxceleb dataset augmented with additive noise and reverberation. The RIR dataset from [15] has been used to generate reverberation samples, while the additive noise sampled are taken from MUSAN [27], a corpus of music, speech and noise. The X-vectors are 512-dimensional vector embeddings computed every 1.5-s segments with a shift of 0.25 s.

VoxCeleb SID. We used Speaker Identification (SID) X-vector system in our experiments. It uses a smaller DNN network than a regular X-vector specifically trained for speaker recognition task [28]. The initial few layers use temporal context such that every frame sees a total context of 15 frames. The features are 24-dimensional filter banks with a frame length of 25 ms, mean-normalized over a sliding window of up to 3 s. The model is trained on VoxCeleb I and II [5, 20] datasets augmented with additive noise and reverberation from Room Impulse Response and Noise Database [3] and MUSAN [27] datasets.

ECAPA-TDNN. We use the ECAPA-TDNN model inspired by our previous work [6] to extract enhanced speaker embeddings. It is an X-vector model improved to include Res2 blocks and channel- and context-dependent attention pooling. Multi-layer Feature Aggregation (MFA) is also used to merge complementary information before the statistics pooling. It has been trained on data with different augmentation strategies like waveform dropout, frequency dropout, speech perturbation, reverberation, addition noise, and noise with reverberation augmentation techniques. The data augmentation is applied on-the-fly to every speech utterance during training. This helps us more variety of data. The ECAPA-TDNN is trained using VoxCeleb I and VoxCeleb II [5, 20] database with Room Impulse Response and Noise Database [3] and MUSAN [27] datasets used for the augmentation. 80-dimensional log Mel-filterbank energies mean-normalized across an input segment forms the input to the ECAPA-TDNN model. For every speech segment, 192-dimensional embeddings are extracted with a sliding window of size 1.5 s. In this work, we empirically try different hop sizes while computing the embeddings. Best performing hop sizes are 0.75 sec and 0.25 sec, and we refer to them as ECAPA-TDNN-1 and ECAPA-TDNN-2, respectively.

3.4 Clustering Algorithms

We tried different types of cluttering techniques in our experiments. In addition to using AHC from baseline setup, we also tried spectral clustering from [17] which has been shown to give high performance [6]. Spectral clustering is a graph-based clustering technique that uses an affinity matrix calculated using the cosine similarity metric. The affinity matrix is then enhanced and the Eigenvectors are computed. The Eigen-values are thresholded to get the number of speaker clusters k . The top Eigenvectors give the spectral embeddings which are more separable and give quite distinct speaker clusters through k-means clustering. We observed that AHC is better if there is hierarchy in the clusters while spectral clustering is useful if the data has connected clusters that do not form a globe.

4 Results and Discussion

The challenge provided baseline results on development dataset. Even though the challenge paper [1] reports DER to be 32.60%, we observe DER of 40.24%

in our implementation. We treat the latter as the baseline for all comparisons as indicated in Table 1. The baseline had UB-GMM and VB-HMM-based resegmentation modules as optional elements. We try modifying this module by using the default speaker shift probability as 0.45. The first two systems S1 and S2 in Table 1 show that adding resegmentation helps improve the DER of the baseline system. This holds not only for the baseline system but also for other systems as can be seen in Table 1.

Table 1. Diarization error rate on development set for different combinations of SAD, speaker embedding vectors and clustering methods.

System No	System Description	DER (%)
Baseline System		
S1	BL SAD+X-vector+PLDA+AHC [1]	40.24
S2	BL SAD+X-vector+PLDA+AHC+VB-HMM	38.74
VAD variation		
S3	Internal VAD+X-vector+PLDA+AHC	49.11
Replacing AHC with SC		
S4	BL SAD+X-vector+SC	51.49
S5	BL SAD+X-vector+PLDA+SC	37.67
S6	BL SAD+X-vector+PLDA+SC+VB-HMM	35.99
Finetuned Model		
S7	BL SAD+X-vector+PLDA+SC	31.08
Replacing X-vector with Voxceleb SID		
S8	BL SAD+VoxCeleb SID (0.25)	45.29
S9	BL SAD+VoxCeleb SID (0.25)+VB-HMM	38.23
Previous Work		
S10	Silero VAD+ECAPA-TDNN-1+SC [6]	39.29
S11	Silero VAD+ECAPA-TDNN-1+SC+VB-HMM	39.02
Towards Best System		
S12	BL SAD+ECAPA-TDNN-1+SC	36.93
S13	BL SAD+ECAPA-TDNN-1+SC+VB-HMM	36.30
S14	BL SAD+ECAPA-TDNN-2+SC+VB-HMM	35.64

4.1 Enhancements Using Clustering Techniques

We replaced the PLDA and AHC modules with spectral clustering. However, this gave poor performance compared to the baseline system. We observed that the input to the spectral clustering algorithm needs to consist of well-separated “connected components” [18] for robust clustering. As PLDA is expected to perform the required vector discrimination, we included the PLDA block before spectral clustering. The DER in Table 1 shows indeed a large improvement is observed when applied spectral clustering on X-vector with PLDA vectors (S5) than only X-vectors alone (S4).

4.2 Investigating the Separability of Speaker Embeddings

We observed that the spectral clustering works well if the speaker embedding vectors are well-separated, as shown in Sect. 4.1. We further explored different versions of speaker embeddings for their noticeable level of separability across speakers. As part of the analysis, we plot X-vector, X-vector with PLDA, Vox-celeb SID and ECAPA-TDNN embeddings for audio in Fig. 2. The scatter plot in Fig. 2a shows that speaker discrimination is not sufficient enough with X-vectors. The PLDA scoring helps improve the speaker separation capability of the X-vectors resulting in better discrimination among the multi-lingual speakers as shown in Fig. 2c. The speaker distinction is the best using ECAPA-TDNN without the need for PLDA as seen from Fig. 2c.

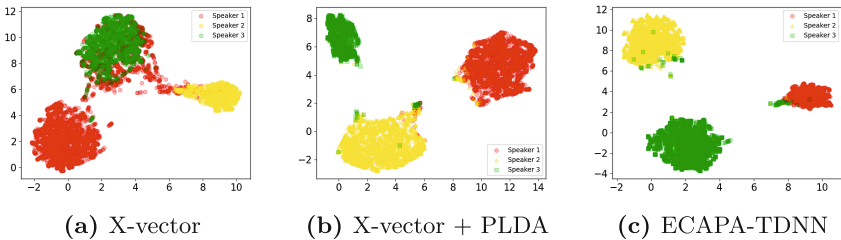


Fig. 2. Scatter plots of (a) X-vector, (b) X-vector+PLDA, and (c) ECAPA-TDNN based enhanced embeddings after the U-map based dimensionality reduction. Red, blue and green color indicate three different speakers in an utterance. (Color figure online)

We were able to achieve 36.93% DER on Dev sets - a marginal improvement compared to Baseline numbers, using ECAPA-TDNN with a TDNN-based SAD system. We observe a further reduction in DER (S13 and S14) compared to baseline (S1) along with VB-HMM rescoring. As seen from Table 1, all the ECAPA-TDNN-based model performances are almost comparable with the X-vector+PLDA+SC approach. This is in line with the observations from Fig. 2. Furthermore, we observe from Table 1 that for the sliding window hop period, $p=0.25$ gives better results than when $p=0.75$. This is because the short speaker utterances like ‘yes’, ‘no’, ‘oh’, etc. can be easily accounted for with small hops.

4.3 Voice Activity Detection

As the DISPLACE challenge data was recorded in far-field conditions, we tried to remove noise or reverberation from the utterances using a DNN pre-processing model. This was followed by utterance segmentation into speech-only regions using LSTM VAD trained on noise-augmented Librispeech data discussed in Sect. 3.2.

In order to incorporate more data, we further tried replacing the TDNN based VAD with Silero VAD as discussed in Sect. 3.2. We see that the use of Silero VAD

reduces the DER compared to when baseline TDNN SAD is used. However, a close analysis of failed cases indicates that the Silero VAD improves the performance for the utterances with high noise and reverberation, while TDNN SAD works best in the case of clean utterances. The development set did not contain many noisy utterances which led to aggregate performance deterioration for S13.

4.4 Results on Dev and Eval Datasets

Results on the development set using various experiments are shown in Table 1. Here, S7 is the finetuned version of the model S5. In general, overfitting happens when the model performs better with very low error rates on test data set [16]. On the contrary, the DERs are higher on the development data set due to the presence of largely divergent data conditions in this task. So, we finetuned the weights of model S5 using the development set and created model S7. In principle, it is not meaningful to measure the DER on the development set itself using system S7. However, as the DER is still on the higher scale even after the finetuning, we consider system S7 as one of the competing systems in our experiments.

Performance Evaluation on Phase-1. We selected four models based on development set results and observations, that is, S13, S14, S11 and S7. The corresponding evaluation set results are shown in Table 2. We observed that the Silero VAD proves to be more robust in the presence of noise variability and works well with short window and hop sizes. Corresponding system S11, therefore, outperforms very short and noisy utterances. That is, if a speaker speaks for a small time in a conversation, used ECAPA-TDNN embedding computed with a small window hop size provides an advantage in helping more accurate speaker change detection for system S14. We combine our four system’s outputs for final submission, considering that each system has its characteristics which help in specific aspects of the task. We performed the fusion based on the maximum voting criteria. The speaker label which appears more times for a given frame is voted as the final speaker label. If none of the four systems claimed the same speaker label, we retained the labels from S7 - the system performing best on the development set. After fusion, the results improved further to achieve a lower DER of 27.70% as shown in Table 2.

Table 2. Results of different combination SAD, speaker vectors and clustering methods in DER (%) on eval phase 1 and phase 2 set.

System No	Systems	Dev	Phase-1	Phase-2
S1	BL X-vector + PLDA + AHC [1]	40.24	39.60	32.50
S13	BL SAD+ECAPA-TDNN (p 0.75)+SC+VB-HMM	36.30	28.11	29.45
S14	BL SAD+ECAPA-TDNN (p 0.25)+SC+VB-HMM	35.64	29.67	28.85
S11	Silero SAD +ECAPA-TDNN (p 0.75)+SC+VB-HMM	39.02	32.41	28.86
S7	BL SAD+X-vector+PLDA+SC	31.08	31.84	29.87
S15	Fusion of S13, S14, S11 and S7	NA	27.67	28.7
S16	Fusion of S14, S11 and S7	NA	27.89	28.59

Table 3. Individual DER results on the eval set for the individual audio files.

Audios →	B015	B016	B020	B023	B027	B029	B035	B039	B051	B053	B054	M007	M008	M010	M017	M044	M047	M048	M050	M053
Sys ↓																				
S13	25.67	41.52	34.55	22.01	7.19	15.81	6.52	33.16	30.04	32.35	22.64	21.89	37.45	30.34	30.81	33.06	15.02	35.47	56.75	13.40
S14	25.60	41.94	34.75	23.94	7.32	15.76	6.45	33.25	29.96	32.33	23.20	30.30	37.37	30.18	30.69	34.70	30.55	35.44	62.32	16.54
S11	27.99	42.03	32.43	31.03	9.69	20.13	7.62	36.08	39.64	41.72	25.86	30.80	37.17	34.46	32.40	27.02	18.97	46.50	77.24	17.23
S7	41.75	48.65	38.53	22.01	7.19	27.48	19.82	33.33	34.08	34.21	22.64	23.57	38.97	50.44	32.61	33.06	15.02	37.90	56.75	13.40

Performance Evaluation on Phase-2. We decided to improve the fusion technique further based on the DOVER approach [8] for the Phase-2 part of the challenge. DOVER-LAP is a method to combine multiple diarization system hypotheses while handling the overlap between multiple speakers. The DOVER-LAP S15 and S16 systems were used to combine individual systems based on empirically selected custom weights. These weights were calculated based on the leave-one-out cross-validation performance on the development set.

4.5 Analysis After Phase-1 Evaluation

We performed a detailed analysis of the audio after the completion of the Phase-1 evaluation. The individual file-wise results for Phase 1 Eval-set are shown in Table 3. We observe that for a few files, the DER is relatively very less (e.g., file B027), while for some others the DER is very high (e.g., file M050).

We observe the scatter plots for the files M007 and M050 as shown in Fig. 3. For the M007 file the speakers are clearly distributed from each other which results in less DER (as reported in Table 3). In particular, with X-vector embeddings the DER obtained is relatively less compared to ECAPA-TDNN embeddings that are reflected from Fig. 3a and Fig. 3b. In addition, we also observe that the M050 recording is highly noisy. Figure 3c and 3d show the X-vector and ECAPA-TDNN embeddings respectively for this recording. In both cases, the speaker embeddings show a large overlap. Due to the noisy nature of speaker embeddings and distance metrics, the quality of the affinity matrix degrades affecting the spectral clustering, thereby leading to poor performance.

On the other hand, the audios B027 and B035, are near-field audios with relatively less noise and reverberation. Further, the gender-related speaker transitions in these audios are generally well-defined. That is, even within the same gender, different speakers have clearly distinct voices due to noticeable variations in pitch and timber. However, these audios do not have overlapping speakers. Each speaker speaks for a longer time as opposed to a short 2 to 3-sec duration. All these are helpful in getting better discrimination across embeddings of different speakers and hence better speaker diarization.

As shown in Table 2, stand-alone system S7 performed better on the dev set and did not perform well on the Eval set compared to other systems. Further investigation shows that most of the audios from the Dev set are clean, while the audios from the Eval set contain comparatively more noise under reverberation conditions. This indicates the importance of domain mismatch not only in terms of language and accent but also in terms of noise and reverberation as well.

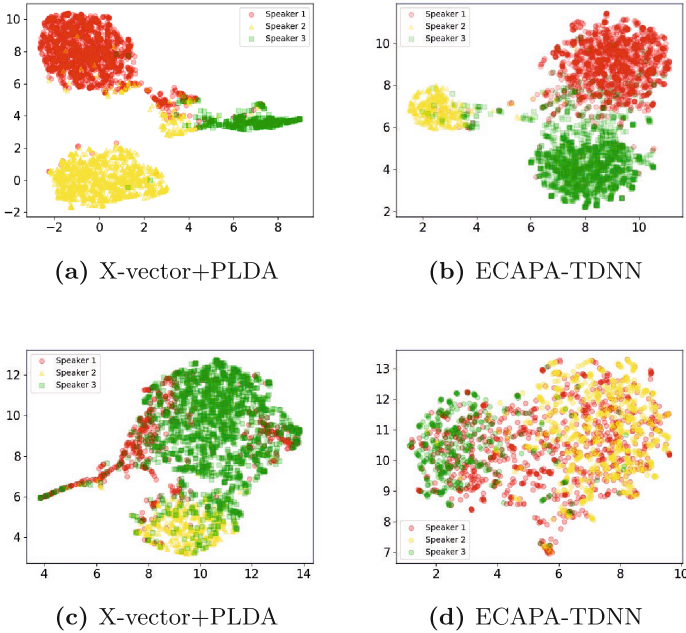


Fig. 3. Scatter plots of (a) X-vector+PLDA, and (b) ECAPA-TDNN embeddings for M007 recording. Scatter plots of (c) X-vector+PLDA, and (d) ECAPA-TDNN embeddings for M050 recording. The U-map based dimensionality reduction is performed before plotting scatter plots.

As shown in Table 3, we observe that all baseline VAD systems perform poorly for the M044 audio, while system S11 using Silero VAD performs much better. M044 audio is extremely noisy with very high pitch and loudness variations due to frequent switching between near-field and far-field conditions. This condition seems to have been captured well by Silero VAD compared to the baseline VAD.

5 Conclusion

In this paper, we built a system for Track-1 of the DISPLACE Challenge that aims at speaker diarization in multilingual scenarios. The system is implemented using different combinations of core sub-modules such as VAD, robust speaker embeddings, and clustering methods. The best system is the combination of different systems obtained using the DOVER-LAP fusion techniques. This represents a significant relative improvement over the baseline by 30.05% which led us to the **second** position for both phase-1 and phase-2 of the DISPLACE challenge. Although there is still room for improvement, we do believe that these are promising results. Our systems perform noticeably well with gender-specific transitions compared to same-gender conversations under multilingual and far-field conditions. The mismatch in the Dev and Eval sets due to huge reverberation

and noisy data made the task more challenging. The use of models trained on large datasets helped with reducing the data mismatch under challenging acoustic conditions.

References

1. Displace challenge. <https://codalab.lisn.upsaclay.fr/competitions/10588>
2. Displace challenge evaluation plan. https://displace2023.github.io/docs/DISPLACE_Evaluation_Plan_v1.pdf
3. openslr.org. <https://www.openslr.org/28/>
4. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: a review of recent research. *IEEE Trans. Audio Speech Lang. Process.* **20**(2), 356–370 (2012)
5. Chung, J.S., Nagrani, A., Zisserman, A.: VoxCeleb2: deep speaker recognition. In: *Proceedings of the Interspeech 2018*, pp. 1086–1090 (2018). <https://doi.org/10.21437/Interspeech.2018-1929>
6. Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., Na, H.: ECAPA-TDNN embeddings for speaker diarization. In: *Proceedings of the Interspeech 2021*, pp. 3560–3564 (2021). <https://doi.org/10.21437/Interspeech.2021-941>
7. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: *Proceedings of the Interspeech 2020*, pp. 3830–3834 (2020). <https://doi.org/10.21437/Interspeech.2020-2650>
8. Raj, D., et al.: DOVER-Lap: a method for combining overlap-aware diarization outputs. In: *2021 IEEE Spoken Language Technology Workshop (SLT)* (2021)
9. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: a new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(2), 652–662 (2019)
10. Graves, A., Graves, A.: Long short-term memory. In: *Supervised Sequence Labelling With Recurrent Neural Networks*, pp. 37–45 (2012)
11. Gudepu, P., Koroth, M.J., Sabu, K., Shaik, M.A.B.: Dynamic encoder RNN for online voice activity detection in adverse noise conditions. In: *Interspeech (Accepted for Publication)*. Dublin, Ireland (2023)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018)
15. Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S.: A study on data augmentation of reverberant speech for robust speech recognition. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224. IEEE (2017)
16. Li, T.W., Lee, G.C.: Performance analysis of fine-tune transferred deep learning. In: *IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pp. 315–319 (2021)

17. Lin, Q., Yin, R., Li, M., Bredin, H., Barras, C.: LSTM based similarity measurement with spectral clustering for speaker diarization. In: Proceedings Interspeech 2019, pp. 366–370 (2019). <https://doi.org/10.21437/Interspeech.2019-1388>
18. von Luxburg, U.: A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007)
19. Lyu, D.C., Chng, E.S., Li, H.: Language diarization for code-switch conversational speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7314–7318. IEEE (2013)
20. Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. In: Proceedings of the Interspeech 2017, pp. 2616–2620 (2017). <https://doi.org/10.21437/Interspeech.2017-950>
21. Park, T.J., Kanda, N., Dimitriadis, D., Han, K.J., Watanabe, S., Narayanan, S.: A review of speaker diarization: recent advances with deep learning. *Comput. Speech Lang.* **72**, 101317 (2022)
22. Prabhavalkar, R., Hori, T., Sainath, T.N., Schliüter, R., Watanabe, S.: End-to-end speech recognition: a survey. *arXiv* (2023). [arXiv:2303.03329](https://arxiv.org/abs/2303.03329)
23. Ryant, N., et al.: The third DIHARD diarization challenge. In: Proceedings of the Interspeech 2021, pp. 3570–3574 (2021). <https://doi.org/10.21437/Interspeech.2021-1208>
24. Sarikaya, R.: The technology behind personal digital assistants: an overview of the system architecture and key components. *IEEE Signal Process. Mag.* **34**(1), 67–81 (2017). <https://doi.org/10.1109/MSP.2016.2617341>
25. Sertsi, P., Boonkla, S., Chunwijitra, V., Kurpukdee, N., Wutiwiwatchai, C.: Robust voice activity detection based on LSTM recurrent neural networks and modulation spectrum. In: Proceedings of the APSIPA ASC, pp. 342–346 (2017). <https://doi.org/10.21437/Interspeech.2021-941>
26. Shankarappa, R., Tiwari, S.: A faster approach for direct speech to speech translation. In: IEEE WINTTECHCON. Bangalore (2022)
27. Snyder, D., Chen, G., Povey, D.: MUSAN: a music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484* (2015)
28. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: robust DNN embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5329–5333. IEEE (2018)
29. Team, S.: Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. <https://github.com/snakers4/silero-vad> (2021)
30. Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4052–4056 (2014)
31. Zeinali, H., Wang, S., Silnova, A., Matějka, P., Plchot, O.: BUT system description to VoxCeleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592* (2019)
32. Zhou, Y., Tian, X., Li, H.: Language agnostic speaker embedding for cross-lingual personalized speech generation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 3427–3439 (2021). <https://doi.org/10.1109/TASLP.2021.3125142>