Exploring different techniques in Spoken Languages Diarization in Indian Languages code mixed with English Language

B.Tech Major Technical Project 2 Report

to be submitted by

Yash Sharma

B20241

for the partial fulfillment of the degree

of

BACHELOR OF TECHNOLOGY IN

Computer Science Engineering

SCHOOL OF COMPUTER AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

KAMAND-175075, INDIA

May, 2024

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the ever-evolving landscape of voice-operated technology, the current paradigm often requires users to pre-select a single language for interactions with smart devices. However, this approach doesn't align with the dynamic reality of human communication, where individuals seamlessly blend languages in their daily discourse. This incongruity underscores the crucial role of Spoken Language Diarization (LD), a transformative technology designed to overcome the limitations of traditional language presets and dynamically adapt to the inherently multilingual nature of human expression.

1.1 Motivation for the Problem

The motivation for this project stems from the growing disparity between the fixed language configurations of contemporary devices and the fluid, code-mixing practices inherent in real-world conversations. As individuals increasingly engage in nuanced multilingual interactions—such as the pervasive use of **Hinglish**, a blend of Hindi and English—LD emerges as an indispensable solution. By precisely identifying and distinguishing between languages within a single utterance, LD not only elevates the user experience in multilingual environments but also addresses the evolving expectations for technology to seamlessly align with the diverse linguistic expressions of its users. In essence, LD serves as a bridge, connecting the limitations of existing language configurations with the dynamic and expressive reality of human language.

1.2 Dependence on Language Identification (LID)

Spoken Language Diarization (LD) intricately relies on the foundational capabilities of Language Identification (LID) systems. LID, tasked with automatically discerning the language within speech utterances, serves as the cornerstone for the multifaceted approach of LD. Building upon the precision of LID, LD goes beyond mere language identification, encompassing tasks such as speaker diarization, temporal segmentation, and contextual preservation within the intricate tapestry of multilingual conversations. The effectiveness and robustness of LD hinge on the accuracy of LID, underscoring the need for advanced language identification technologies to provide a solid foundation for the nuanced tasks performed by LD systems.

1.3 Challenges in Language Identification (LID)

The landscape of Language Identification (LID) encounters formidable challenges, particularly in the dynamic context of multilingual and code-mixing-rich environments. High interclass similarities between languages present a persistent hurdle, especially in regions like India, where linguistic diversity is abundant. Intraclass variations within languages, coupled with the widespread practice of code-mixing, further compound the complexity. In real-world scenarios, such as those in India, where code-mixing, exemplified by expressions like Hinglish, is prevalent, LID systems face the intricate task of accurately identifying languages within seamlessly blended utterances. Moreover, the dynamic nature of real-world settings introduces variations in acoustic conditions, emotional states, and recording devices, posing additional challenges to the precision and adaptability of LID systems.

The challenges extend beyond the linguistic realm, encompassing domain-mismatch issues between training and testing samples. In the Indian linguistic landscape, characterized by diverse phonetic structures, dialectical nuances, and intricate linguistic variations, LID encounters hurdles that demand a nuanced and culturally sensitive approach. Bridging the gap between idealized training conditions and the complexities of real-world scenarios remains a formidable task, necessitating innovative solutions

for accurate and contextually aware language identification.

1.4 Research Gap in Indian Native Languages

In the context of Indian native languages, a conspicuous research gap exists, particularly in the development and adaptation of LID and LD systems. The linguistic diversity in India, boasting a multitude of languages and dialects, introduces layers of complexity to language technology applications. While strides have been made in addressing challenges in English-centric contexts, the scarcity of research on LID and LD for Indian languages, especially those marked by code-mixing like Hinglish, represents a significant void. The unique phonetic structures, linguistic variations, and intricate dialectical nuances pose specific challenges that demand targeted research efforts.

To address these challenges effectively, research initiatives must focus on developing culturally sensitive and linguistically robust LID and LD systems. Bridging this research gap is essential for ensuring that language technologies are tailored to the linguistic intricacies of the diverse Indian populace. Through comprehensive and focused research endeavors, the aim is to elevate the adaptability and accuracy of LID and LD systems, enabling their seamless integration into the rich tapestry of Indian language scenarios.

1.5 State of the art in the field

While Language Identification (LID) systems play a pivotal role in the nuanced realm of Spoken Language Diarization (LD), the state-of-the-art still grapples with persistent challenges in classifying and diarizing multilingual scenarios. Despite notable advancements, issues such as high interclass similarities, intraclass dissimilarities, and domain mismatches persist, hampering the effectiveness of these systems.

In the complex landscape of multilingual interactions and code-mixing-rich environments, distinguishing between languages and accurately attributing speech to specific speakers remains a formidable task for existing LD frameworks. Interclass similari-

ties, wherein languages from the same family exhibit significant overlap in phonetic structures, create confusion, especially in short speech segments. Additionally, intra-class dissimilarities stemming from dialects, accents, and variations within a language further challenge the robustness of LD systems. The persistent problem of domain mismatch, where training conditions differ from real-world usage, adds another layer of complexity, impacting the adaptability of these systems.

Existing literature highlights the critical role of Language Identification (LID) systems in the intricate dance of Spoken Language Diarization (LD). Notable papers and works, such as [1] , underscore the foundational capabilities of LID as a cornerstone for the multifaceted approach of LD. Moreover, techniques for LD in dynamic multilingual and code-mixing-rich environments, have been documented in [2].

Acknowledging these challenges, our approach aims to leverage state-of-the-art technologies, such as Wave2Vec2, to address the shortcomings in current LD systems. Wave2Vec2, a powerful neural network architecture, excels in capturing intricate acoustic features and is poised to enhance the discriminative capabilities of LD models. By harnessing the capabilities of Wave2Vec2, we aspire to mitigate the impact of inter-class similarities, handle intraclass dissimilarities more effectively, and bridge the gap caused by domain mismatches. This strategic integration of cutting-edge technology represents a step towards eradicating the complexities posed by multilingual scenarios, offering a more accurate and adaptable solution for Spoken Language Diarization.

1.6 Paving the Way for Multilingual Technological Harmony

In conclusion, as we navigate the complex terrain of Spoken Language Diarization (LD) and its reliance on Language Identification (LID), we uncover the intricate dance between cutting-edge technology and the diverse linguistic expressions of human communication. The symbiotic relationship between LD and LID underscores the need for advancements in language technology that mirror the fluidity and dynamism of multilingual interactions.

While challenges persist in the realms of LID, particularly in the face of code-mixing and real-world variability, our journey is guided by the pursuit of solutions. The unique linguistic landscape of India, exemplified by the prevalent Hinglish code-mixing phenomenon, represents an untapped reservoir of linguistic richness awaiting exploration. Bridging the existing research gap in Indian native languages is not merely a scholarly endeavor; it is a commitment to enhancing the inclusivity and effectiveness of language technologies.

1.7 Expected Outcome

The anticipated outcome of this research project is twofold: first, the refinement of Language Identification (LID) and Spoken Language Diarization (LD) systems tailored to the intricacies of Indian languages, particularly in the context of Hinglish code-mixing; and second, the generation of valuable insights that transcend regional boundaries, contributing to a global understanding of multilingual technology. Through innovative solutions and a culturally sensitive approach, we aim to enable seamless language identification and diarization within Hinglish code-mixed utterances, ensuring a smooth transition between Hindi and English. Ultimately, this project aspires to serve as a catalyst in advancing the harmonious coexistence of cutting-edge tools with the rich tapestry of human communication, where Spoken Language Diarization becomes a vital linchpin bridging technology and the intricate melodies of multilingual conversations.

Chapter 2

System Model

2.1 Description of Scenario

Our system operates in the dynamic landscape of multilingual conversations, focusing on the intricacies of spoken language diarization. Within this context, the unique challenge arises in scenarios where code-mixing is prevalent, and individuals seamlessly switch between languages. Leveraging the Wave2Vec2 transformer model, our approach involves a two-phase system to address this complexity. In the initial phase, hidden features are extracted from the audio data. Secondly, depending on the approach, Time-Delay Neural Network (TDNN) is trained to learn a specialized Language Identification (LID) x-vector representation while Attention based approach is followed to learn u-vector representation. Subsequently, in the diarization phase, a signal processing approach is applied to the outputs obtained from the LID phase, aiming to complete the diarization process and provide precise language differentiation.

2.2 Problem Statement

The central problem addressed by our system is spoken language diarization in multilingual conversations, especially within the context of code-mixing. The first phase focuses on Language Identification (LID), utilizing hidden features obtained from the Wave2Vec2 transformer model to train a TDNN for learning a specialized represen-

tation (x vector) and Attention based method to learn u-vector representation. This representation is crucial for accurate LID. In the second phase, the challenge lies in refining and completing the diarization process using a signal processing approach applied to the LID-specific representation, ultimately providing accurate differentiation between languages within a given utterance.

2.3 Speech corpora in the study

We have used the subset of data from EkStep datasets which comprises the 12 languages (11 Indian languages and English) for fine-tuning the wav2vec. The actual durations and number of utterances of each language are shown in Table 2.3.

Before fine-tuning wav2vec, the EkStep data was converted into 2sec of chunks. Further, these chunks were used to fine-tune the wav2vec model.

Language	**Duration (in Hours)**	**Number of Utterances**
Assamese	17.11	7363
Bengali	17.16	6411
English	15.50	5820
Gujarati	14.68	6087
Hindi	18.32	6300
Kannada	15.75	6014
Malayalam	15.63	6597
Marathi	17.03	6266
Odia	15.23	6428
Punjabi	17.71	5245
Tamil	16.31	5082
Telugu	16.61	6512

Table 2.1: EkStep dataset duration and number of utterances for 12 languages used for fine-tuning the wav2vec.

2.4 Proposed Methodology

The proposed methodology consists of a two-phase approach aimed at solving the spoken language diarization problem in multilingual conversations. The pipeline, as

illustrated in Figure 2.1, details the key steps involved in both the Language Identification (LID) and Spoken Language Diarization (LD) phases.

The primary algorithm encompasses two broad approaches, denoted as S1 and S2, utilizing different embeddings: x-vector and u-vector. Both approaches share a common feature extraction step at the frame level, employing the wav2vec model fine-tuned with 2-second utterances. The resultant hidden features are processed to derive either x-vector or u-vector based on the selected approach.

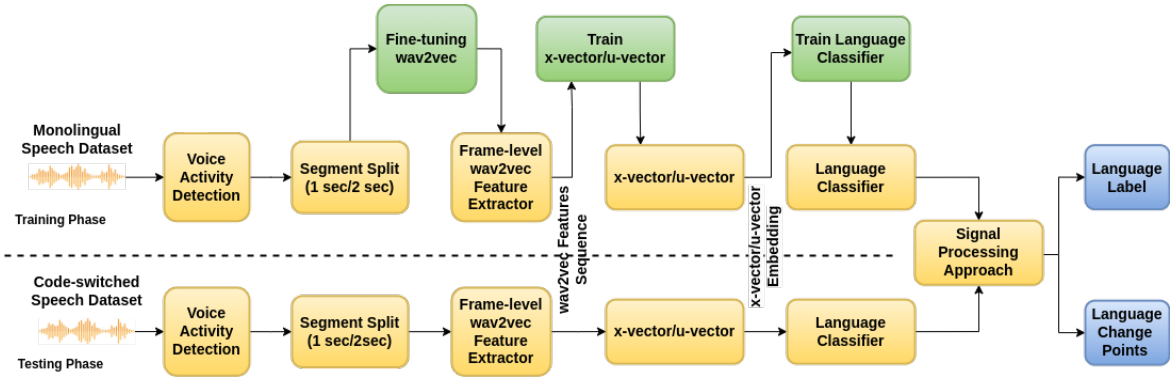

Fig. 2.1: Schematic diagram of speech production.

2.5 Proposed framework for LID using LID-senones based x-vector

2.5.1 Frame-level Feature extraction using Wave2Vec2

In the initial stage of our process, the primary task is to extract frame-level features from audio data. To achieve this, we employ the Wave2Vec2.0-base model, a robust self-supervised speech representation learning architecture. This model, pre-trained on unlabeled data following the framework presented in the "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations" paper [3], is fine-tuned for speech recognition tasks using a Connectionist Temporal Classification (CTC) loss on the LibriSpeech dataset, encompassing 960 hours of audio.

The Wave2Vec2.0-base model incorporates a multi-layer convolutional feature en-

Parameter	Details
Model Architecture	Wave2Vec2.0-base
Pre-training Task	Self-Supervised Learning of Speech Representations
Fine-Tuning Task	Speech Recognition with CTC loss on LibriSpeech dataset
Pre-training Data	Unlabeled data
Fine-Tuning Data	LibriSpeech dataset (960 hours of audio)
Feature Encoder	Multi-layer Convolutional Feature Encoder
Transformer Layers	12
Feature Dimension	768
Expected Sample Rate	16 kHz
Hidden Features Size	(#frames, 1024)

Table 2.2: Wave2Vec2.0-base Model Configuration

coder that processes raw audio inputs to generate latent speech representations. These representations are then fed into a Transformer, allowing the model to capture information across the entire audio sequence. The Transformer in the base model comprises 12 layers, and the resulting feature dimension is 768.

It's important to note that the Wave2Vec2.0-base model is optimized for audio samples recorded at a 16 kHz sample rate, ensuring compatibility with the common sampling rate of audio data.

For more comprehensive details, we recommend referring to the associated repository and model card for Wave2Vec2.0-base.

We for our purpose have finetuned the wave2vec2-base over the available audio speech corpus. We have used the following pipeline while finetuning as shown in Figure 2.1.

2.5.2 Hidden Features Processing with TDNN

In this section, we present the Time-Delay Neural Network (TDNN) architecture used for Language Identification (LID), along with its key specifications.

2.5.3 TDNN Architecture Details

The TDNN architecture is a crucial component for learning language-specific representations, including the utterance-level embedding (x-vector). Below is a summary

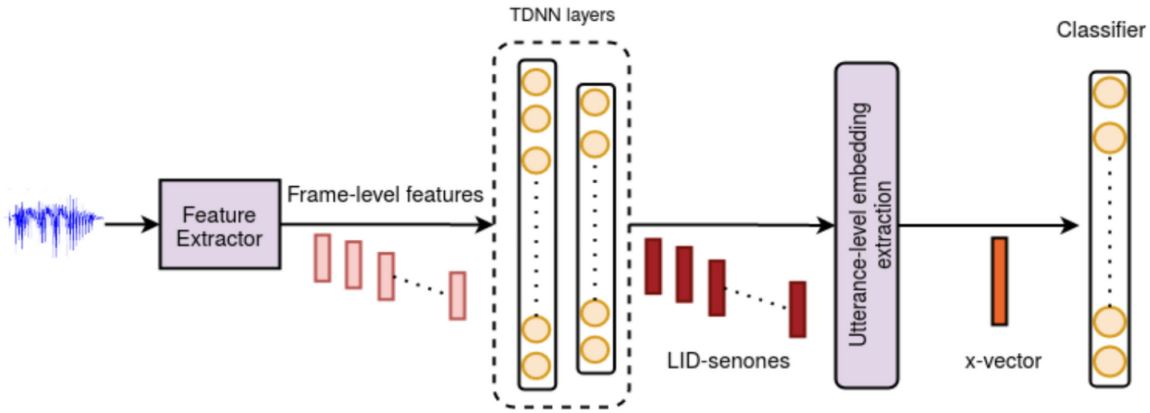

Fig. 2.2: Proposed TDNN Unit [1]

Layer	Input Dimension	Output Dimension	Context Size	Dilation
TDNN1	39	512	5	1
TDNN2	512	512	3	1
TDNN3	512	512	2	2
TDNN4	512	512	1	1
TDNN5	512	512	1	3
Segment6	1024	512	-	-
Segment7	512	512	-	-
Output	512	*11*	-	-

Table 2.3: TDNN Architecture Configuration

of the TDNN layers and their configurations:

In the Language Identification (LID) phase, the output of the Time-Delay Neural Network (TDNN) layers, known as LID-senones [1], obtained over the entire speech sample, undergoes further processing by an utterance-level embedding extractor. In the standard x-vector architecture, a statistics pooling layer is employed. This layer computes the mean and standard deviation of the output of TDNN layers, followed by a dense layer to obtain the utterance-level embedding, commonly referred to as the x-vector [1]. The x-vector is then fed into the output layer with softmax activation to predict the language label.

In the statistics pooling-based approach, an additional layer called the statistics pooling layer is introduced. Let $H = (h_1, h_2, \ldots, h_t, \ldots, h_T)$ represent the sequence of LID-senones obtained by passing the input frame level features vectors through the TDNN layers of the proposed LID network (denoted as x-vec-Net). Here, T represents

the number of speech chunks, and $h_t \in \mathbb{R}^D$ represents the t-th LID-senone. The statistics pooling layer computes the mean (μ) and standard deviation (σ) of all LID-senones as follows:

$$\mu = \frac{1}{T}\sum_{t=1}^{T} h_t$$

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(h_t - \mu)^2}$$

The utterance-level embedding (x-vector) is then obtained by concatenating the mean vector with the standard deviation:

$$x = [\mu, \sigma]$$

2.6 Proposed framework for LID using robust u-vector representations

The framework for Language Identification (LID) utilises the hidden features based on u-vector aimed at extracting an utterance-level embedding (u-vector), and a language classifier block. The BLSTM layers generate LID-seq-senones. These senones are then utilized by an utterance-level embedding extractor to produce the u-vector, subsequently fed into the classifier for language prediction.

The network is trained end-to-end using categorical cross-entropy loss, enabling the BLSTM layers to discern between languages using temporal variations in the input sequence of BNF vectors. Consequently, the resulting LID-seq-senones contain richer language-discriminative information compared to simple frame-level features.

To address the challenge of varying-length representations of speech due to fixed-size chunks, two approaches are proposed for combining LID-seq-senones into an utterance-level representation (u-vector). The approach employs a self-attention-based fusion technique.

2.6.1 Self-attention based fusion

To overcome the limitation in statistics pooling based fusion, we propose to use a self-attention based strategy. The self-attention mechanism dynamically assigns the weights to LID-seq-senones depending on their relevance in determining the language label.

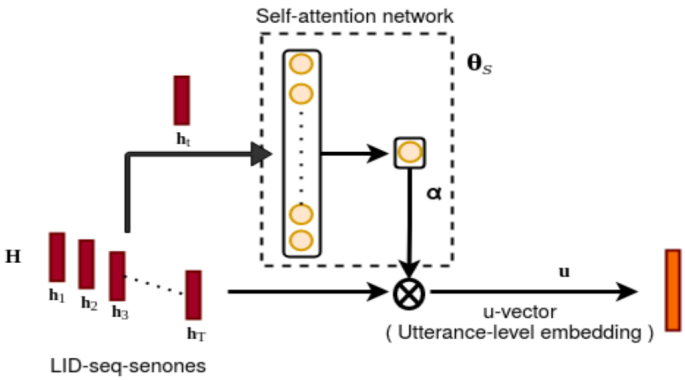

Fig. 2.3: Self-attention based fusion of LID-seq-senones to get u-vector.

Let $H = (h_1, h_2, \ldots, h_t, \ldots, h_T)$, be the sequence of LID-seq-senones obtained by passing the input BNF vectors through the BLSTM layers of the proposed LID network (denoted as u-vec-Net). Where, T represents the number of chunks of speech, and $h_t \in \mathbb{R}^D$ represents the t^{th} LID-seq-senone. As shown in Figure 2.3, this sequence of LID-seq-senones is then fed to the self-attention network, which contains a dense (fully connected) layer with N_a number of nodes followed by a layer with a single node. Both layers have tanh activation function. Note that, this self-attention network can also be viewed as a fully connected neural network with 2 layers. Given H, the self-attention network first computes an intermediate representation at the output of the dense layer with a single node, denoted as $\gamma = [\gamma_1, \ldots, \gamma_t, \ldots, \gamma_T]^T$, as follows:

$$z_t = \tanh(W_a^T h_t + b_a); \text{ for } t = 1, 2, \ldots, t, \ldots, T \tag{2.1}$$

$$\gamma_t = \tanh(w_\gamma^T z_t + b_\gamma) \tag{2.2}$$

Where, $W_a \in \mathbb{R}^{D \times N_a}$ and $b_a \in \mathbb{R}^{N_a}$ are respectively the weights and biases of the dense

layer with N_a nodes, $w_\gamma \in \mathbb{R}^{N_a}$ and b_γ are respectively the weights and bias of the layer with a single node. Note that, all these parameters of the self-attention network (W_a, b_a, w_γ, and b_γ) are collectively denoted as Θ_S. Using the intermediate representation γ, the attention vector $\alpha = [\alpha_1, \ldots, \alpha_t, \ldots, \alpha_T]^T$, is then computed as:

$$\alpha = \text{softmax}(\gamma) \tag{2.3}$$

Each value in this attention vector (α_t) indicates the weight associated with the corresponding LID-seq-senone h_t. Using H and α, a fixed-length representation of the speech utterance (u-vector) is then computed as:

$$\mathbf{u} = H\alpha \tag{2.4}$$

This u-vector (obtained using either statistics pooling or self-attention based pooling), is then fed to the classifier network for identifying the language.

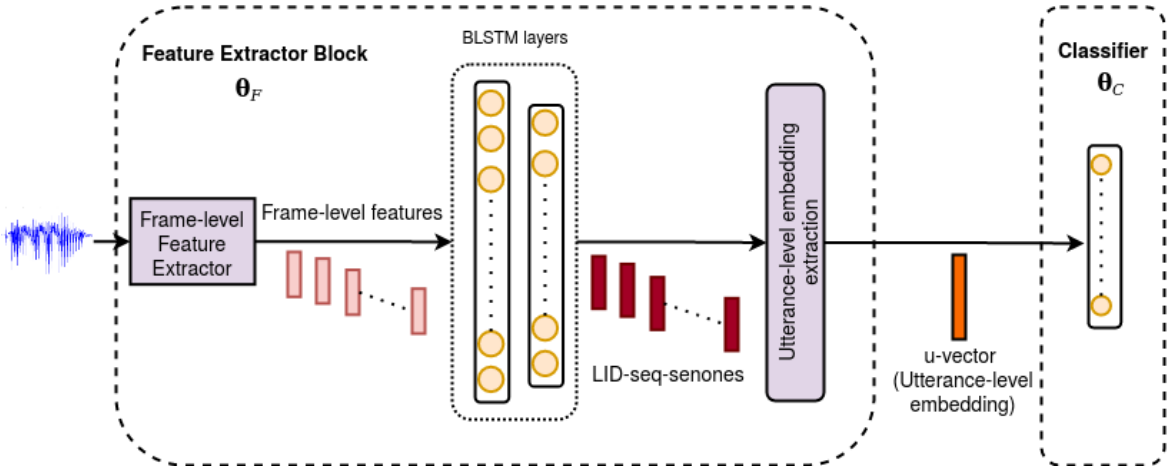

Fig. 2.4: Block diagram of the framework for u-vector.

2.6.2 Comparison with x-vector

The proposed u-vector approach shares similarities with the state-of-the-art x-vector network. Both networks involve a frame-level feature extractor followed by processing layers to obtain embeddings. However, key differences lie in the utilization of BLSTM

units instead of TDNN units and the use of self-attention-based fusion instead of statistical pooling. These distinctions contribute to the u-vector representation being considered more discriminative than the x-vector.

2.7 Spoken Language Diarization (LD) Phase: Signal Processing for Change Point Detection

The output of the last classifier layer from each architecture provides probabilities for each segment belonging to languages L1 or L2, stored in S0 and S1, respectively. The approach draws inspiration from the paper [2], and the key details are outlined below. Subsequently, the following algorithm is employed:

1. **Step 1:** Suppose the output of the network corresponding to the overlapping speech segments belongs to language 1 is $S1(n)$ and language 2 is $S2(n)$. The Gaussian smoothed output $Sg1(n)$ and $Sg2(n)$ can be computed as:

$$Sgi(n) = \frac{1}{M} \sum_{k=-(M-1)/2}^{(M-1)/2} Si(n + (M-1)/2 + k) \cdot Wg(k)$$

   where $1 \leq i \leq 2$, $0 \leq n \leq L - (M-1)/2$, M is the size of the Gaussian window, L is the number of 200 msec segments present in the code-switched utterance, and $Wg(k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{k^2}{2\sigma^2}}$.

2. **Step 2:** After smoothing, the signed difference of $Sg1(n)$ and $Sg2(n)$ is computed, i.e.,

$$dSg = \text{sign}(Sg1(n) - Sg2(n))$$

3. **Step 3:** Then, the first-order difference of dSg is computed as:

$$fdSg(n) = dSg(n+1) - dSg(n)$$

4. **Step 4:** The language change points (CP) and the segment labels (SL) can be

computed as follows:

$$CP = \arg(\{j | 0.5 \times |fdSg(n)| == 1\})$$

$$SL(n) = \arg\max(Sg1(n), Sg2(n))$$

where 'sign' represents the signum function.

This signal-processing approach effectively identifies language transitions within the utterance, contributing to the Spoken Language Diarization process.

2.8 Configurations

The key configurations set while training the models and the machines used are mentioned in 2.4.

Configuration	**Details**
Supercomputer	Param Siddhi HPC, Pune
Finetuning Approach	Semi-batch (256) for Wave2Vec2
Training Algorithm	Stochastic Gradient Descent (SGD) for TDNN
GPU Configuration	Used distributed finetuning with 8 GPUs (A100-SXM4-40GB)
CPU Configuration	Employed 256 CPUs for computational power
Training Precision	Mixed-precision training using f16

Table 2.4: Machine Configuration

The training and testing phases are configured to address the complexities of spoken language diarization, considering the nuances of monolingual and code-mixed scenarios.

2.8.1 Training Phase Configuration

During the training phase, the system is exposed to monolingual audio samples. The audio undergoes Voice Activity Detection (VAD) to remove silence, ensuring a focus on speech segments. Subsequently, the audio is segmented into non-overlapping 2-second chunks, providing a disjoint representation for each segment. This setup aims

to train the system on clear and distinct language instances, facilitating robust learning of language-specific features. Later depending on the methodology used, x-vector/u-vector embeddings are learnt. A simple classifier is used to classify language based on the embeddings. This completes are LID (Language Identification). The LD system take these classified probabilities to further find the language labels and change points.

2.8.2 Testing Phase Configuration

In contrast, the testing phase introduces code-mixed samples, where speakers seamlessly switch between languages. To adapt to this real-world complexity, the testing audio is **not subjected to VAD** and segmented into overlapping 4-second chunks. The overlapping configuration ensures that the system can effectively handle the fluid transitions between languages within an utterance. This configuration aligns with the dynamic nature of multilingual conversations, providing a realistic evaluation scenario for the diarization system.

Chapter 3

Results

Our study focuses on language diarization (LD) within Indian dialects, encompassing both English and non-English variants, as part of the **DISPLACE-2024** challenge conducted at Interspeech-2024 by IISc Bangalore. Employing advanced wav2vec2 embedding features alongside x-vector and u-vector models, we have achieved precise language identification (LID). Our methodology is fortified by employing signal processing techniques, ensuring accurate diarization and bolstering system robustness.

The fine-tuning of the wav2vec2 model on EkStep data has enabled nuanced linguistic capture, further optimized on the Displace challenge dataset. The results of our approach demonstrate promising outcomes. We have achieved a Diarization Error Rate (DER) of 12.57% and 35.22% on the development and evaluation datasets, respectively. Moreover, we have observed an 11.59% enhancement over the baseline result of 40.58% on the development dataset. However, there is a 2.54% reduction compared to the baseline result of 32.68% on the evaluation dataset.

These results highlight significant progress in LD methodologies tailored for diverse linguistic contexts, particularly within the Indian dialect landscape.

3.0.1 Wave2Vec2 Fine-tuning

We performed fine-tuning on the `facebook/wav2vec2-large-xlsr-53` model, which was pre-trained on a subset of the dataset. The fine-tuning was done in a distributed manner, leveraging the "Param Siddhi Supercomputer" in Pune. The model was

Language	Precision	Recall	F1-Score	Support
asm	0.97	0.98	0.98	3566
ben	0.98	0.97	0.98	3987
eng	0.98	0.92	0.95	1104
guj	0.97	0.99	0.98	3408
hin	0.98	0.96	0.97	3958
kan	0.96	0.98	0.97	1858
mal	0.94	0.95	0.95	1081
mar	0.91	0.96	0.93	1030
odi	0.97	0.98	0.97	2695
pun	0.97	0.98	0.97	3618
tam	0.98	0.98	0.98	3862
tel	0.95	0.89	0.92	1193
Accuracy	0.97			
Macro Avg	0.96	0.96	0.96	31360
Weighted Avg	0.97	0.97	0.97	31360

Table 3.1: Wave2Vec2 Classification Report

trained using a semi-batch size of 256. The training details include a final training loss of 0.3367, a validation loss of 0.3434, and an accuracy of 97.47%. A detailed accuracy report for the final evaluation of best of finetuned wave2vec2 is also given in 3.1

3.0.2 TDNN Training Using Hidden Features from finetuned Wave2Vec2

In this section, we discuss the training of the Time-Delay Neural Network (TDNN) from scratch using the hidden features obtained from the finetuned Wave2Vec2. The hidden features are of size (1, 99, 1024) for a single 2-second chunk.

3.0.3 Evaluation Metric: Diarization Error Rate (DER)

The evaluation of language diarization systems commonly employs the Diarization Error Rate (DER) as a key metric. DER quantifies the dissimilarity between a system-generated diarization and a reference diarization, typically represented in Reference Text Time Markup (RTTM) 3.1 files. RTTM files play a pivotal role in the field of language diarization, offering structured annotations for segmenting and labeling

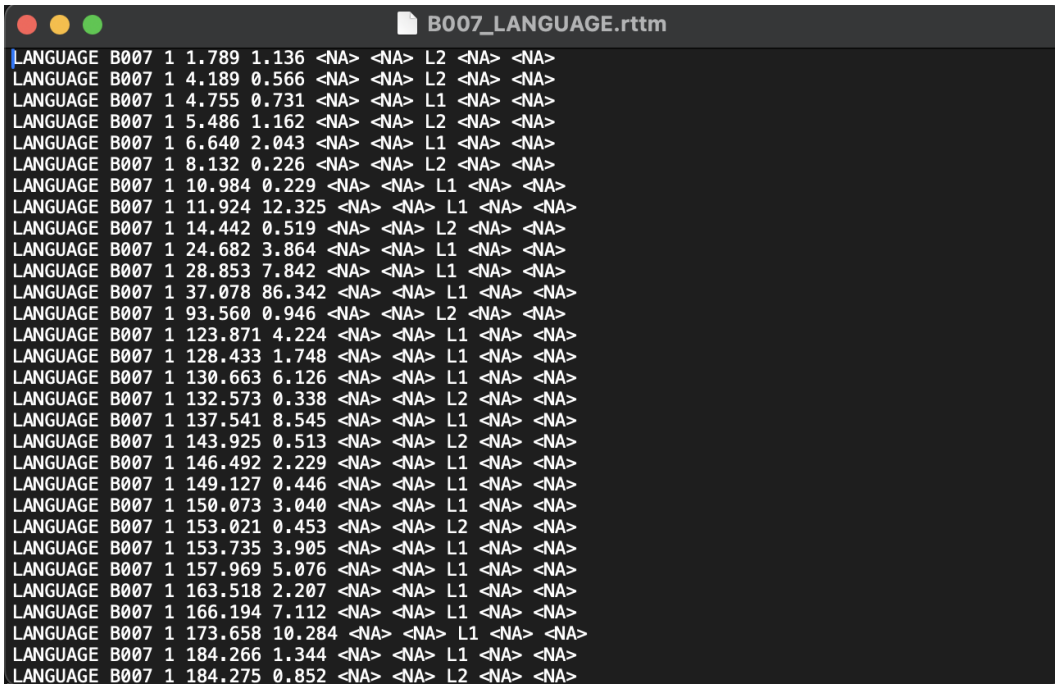

Fig. 3.1: A typical RTTM file

speakers within audio recordings. These files provide essential tools for annotating language turns, facilitating the analysis and comprehension of spoken language data.

The formula for calculating DER is as follows:

$$DER = \frac{FalseAlarm + Miss + Overlap + Confusion}{ReferenceLength}$$

Where:

- **False Alarm**: False positive speaker segments detected by the system.
- **Miss**: Speaker segments present in the reference but not detected by the system.
- **Overlap**: Overlapping speaker segments detected by the system.
- **Confusion**: Speaker segments incorrectly labeled by the system.
- **Reference Length**: Total length of the reference audio in seconds.

DER provides a comprehensive measure of the performance of diarization systems, capturing various aspects of errors including false positives, misses, overlaps, and con-

fusions. Lower DER values indicate better performance, reflecting higher accuracy in speaker segmentation and labeling.

3.0.3.1 Training Loss, Accuracy, and Validation Metrics

The training process involved training the TDNN with a focus on language discrimination using the extracted hidden features. The following figures illustrate the trends in training loss, accuracy, and validation metrics during the training phase:

Metric	**Value**
Train Loss	0.144
Train Accuracy	0.9689
Validation Loss	0.0964
Validation Accuracy	0.9545

Table 3.2: TDNN Training Metrics

3.1 LD Results on 12 Indian Languages

In this section, we present the results of the LD applied to the outputs of the x-vector/u-vector model for Hindi codemixed with English language. We needed to mask the other languages probabilities as this systems is only designed to handle only 2 languages codemixed.

The first evaluation was done on a synthetically stitched togerther single-speaker audio to make a codemixed one. Refer to 3.1 for the final DER on the synthetic data for variable window size.

System	**Window Size**	**DER**	**JER**	**B3-P**	**B3-F1**
Baseline	-	-	-	-	-
S-1 (x-vector)	16000	30.39	5233	0.61	0.72
	32000	**18.86**	18.63	0.94	0.84
	48000	23.29	39.24	0.66	0.76

Table 3.3: Language Diarization results on synthetically-generated Codemixed data

The RTTM for a particular synthetic Audio can be visualized. Refer to 3.2 for the same. Using 12 languages on the DISPLACE-2024 didn't yield impressive DER

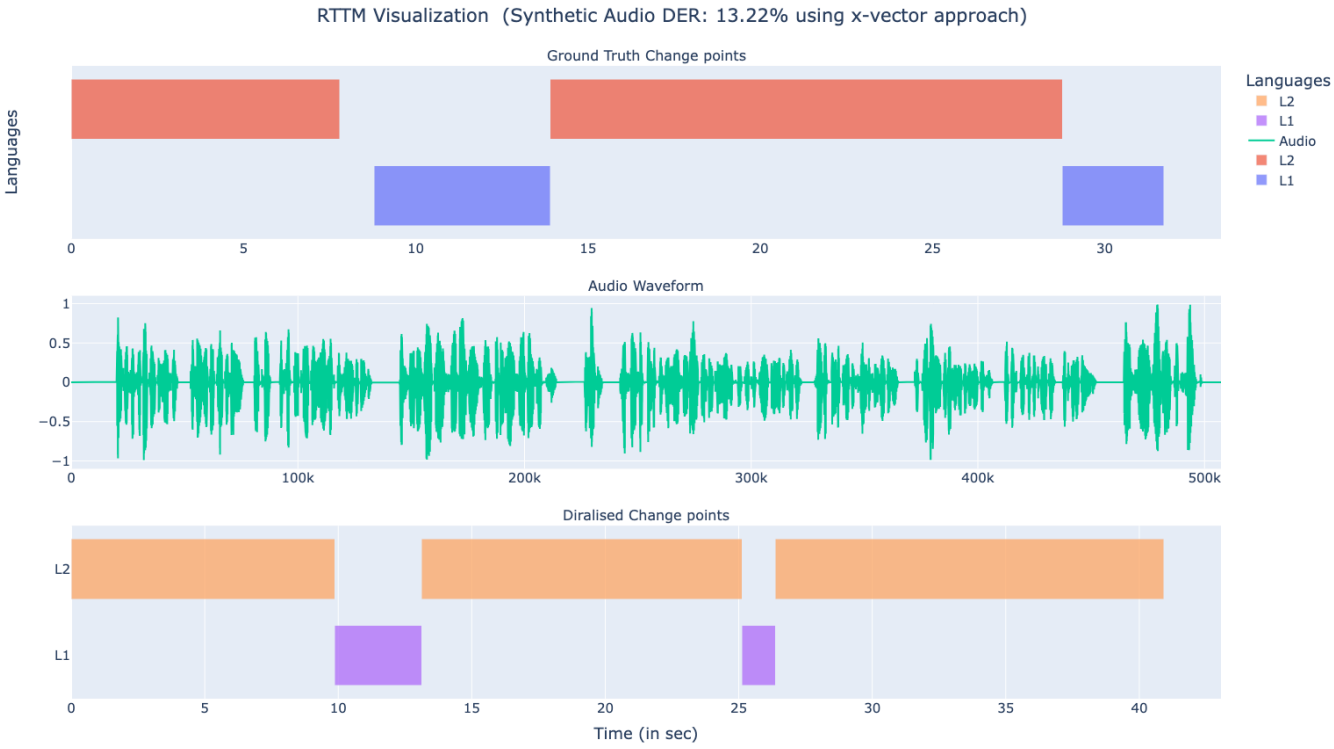

Fig. 3.2: RTTM Visualization for synthetic code-mixed data

(45.33 %). Thus, we had to resort to 2 Languages system in LID which is outlined in the next section.

3.2 Results on DISPLACE-2024 challenge

In this section, we will present the result obtained on the TEST Set of the DISPLACE-2024 challenge. We reduced our dataset from 12 languages to just 2 languages mainly eng and not-eng. This way our system (LID) just had to classify for a smaller number of classes making it easy for our LD to diralize. This yielded some impressive results on both Development 3.2 and Test Dataset 3.2. The RTTM visualization for a given audio in DISPLACE-2024 can be found in 3.3

System	Window Size	DER	JER	B3-P	B3-F1
Baseline	-	40.58	-	-	-
S-1	16000	28.99	47.48	0.56	0.63
	32000	29.61	48.40	0.55	0.62
	48000	28.99	47.48	0.56	0.63
S-2	16000	33.62	46.79	0.73	0.62
	32000	31.62	43.01	0.71	0.72
	48000	**12.57**	27.02	0.73	0.81

Table 3.4: Language Diarization results on Development set

System	Window Size	DER	JER	B3-P	B3-F1
Baseline	-	32.68	-	-	-
S-1	16000	38.55	66.74	0.54	0.60
	32000	37.20	65.67	0.55	0.61
	48000	**35.22**	64.64	0.55	0.62
S-2	16000	42.71	69.76	0.54	0.61
	32000	37.47	66.24	0.55	0.63
	48000	35.45	64.60	0.56	0.63
	64000	36.18	65.21	0.55	0.62

Table 3.5: Language Diarization results on Evaluation set

3.3 TSNE Plot

In this section we will show the tsne plots of the extracted hidden-features from wave2vec2 in 3.4 and the x-vector embeddings learnt for the 2 class language classification in 3.5.

3.4 Ablation Study

In order to comprehensively evaluate the effectiveness of our proposed language diarization pipeline, we conducted an ablation study. This systematic investigation involved selectively modifying or removing various components within our pipeline to assess their individual impact on performance. Through this study, we aimed to identify the key components that significantly contribute to the overall effectiveness of our language

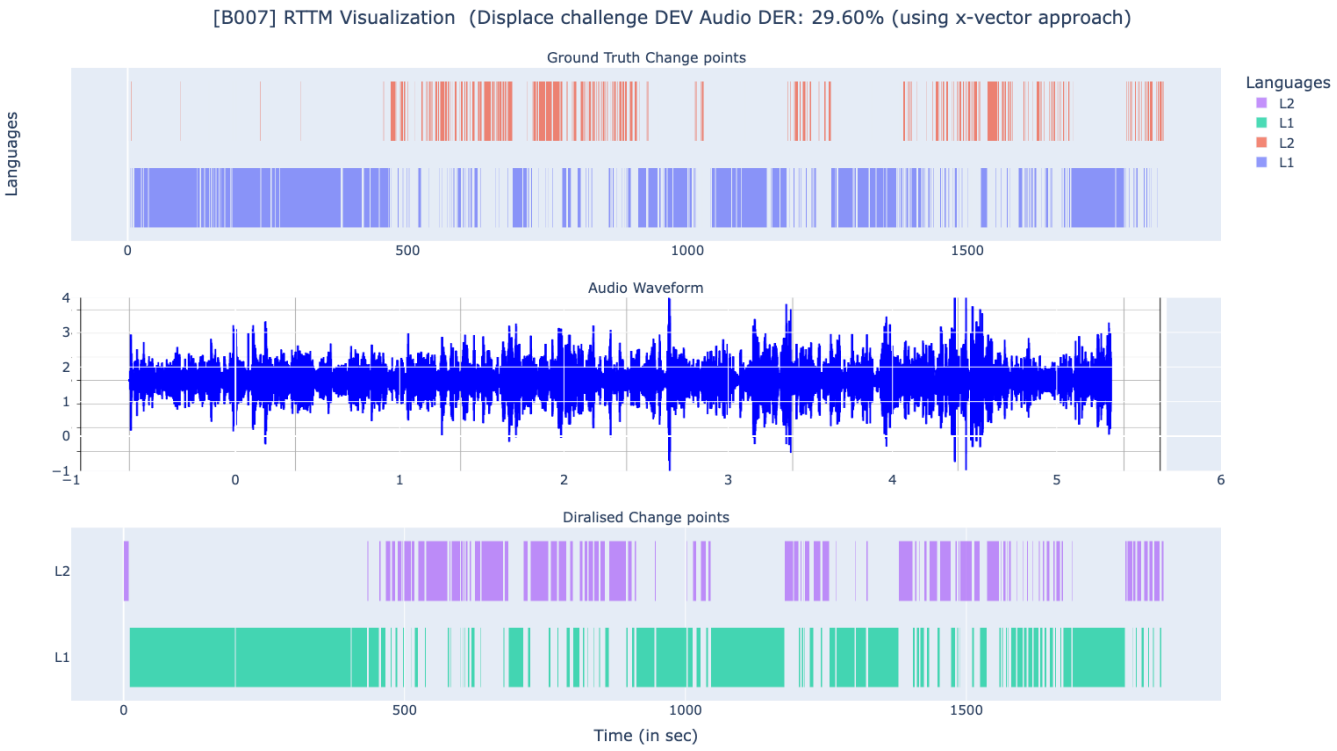

Fig. 3.3: RTTM Visualization for an audio from Displace2024 DEV Dataset

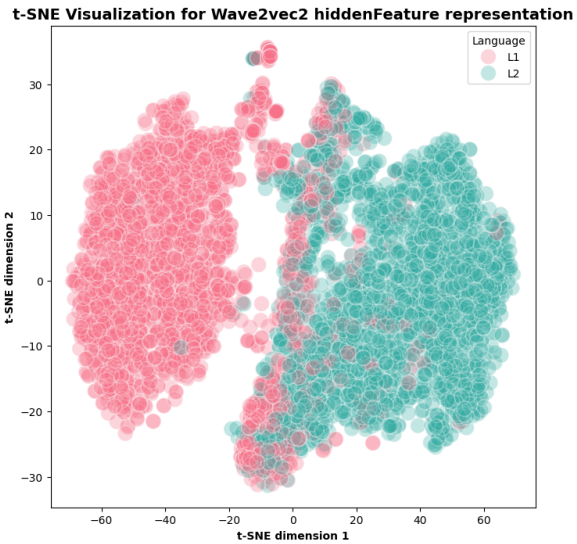

Fig. 3.4: TSNE Plot for Hidden Features

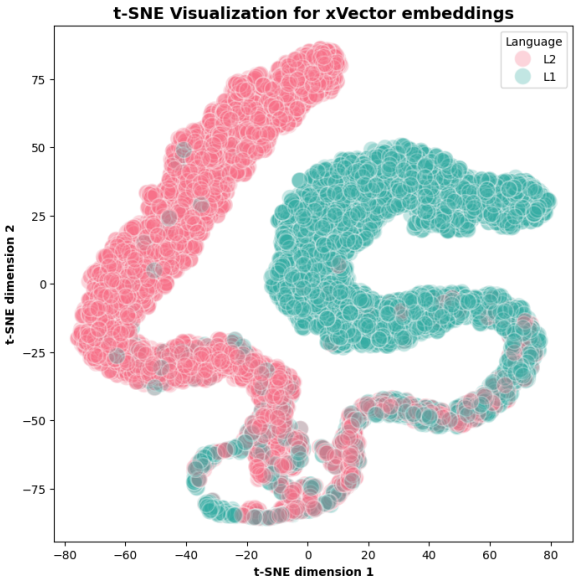

Fig. 3.5: TSNE Plot for x-vector embeddings

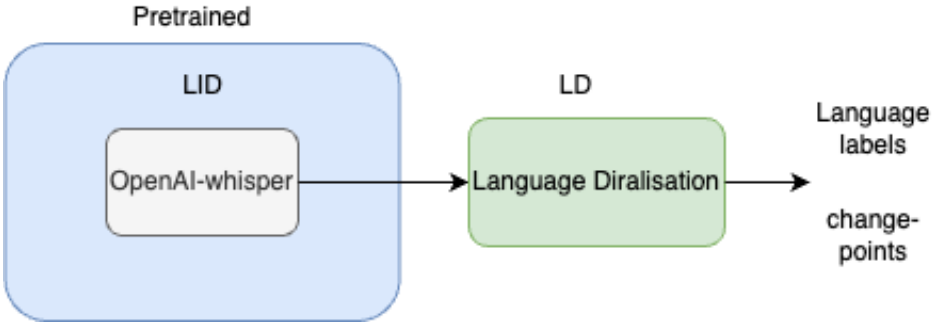

Fig. 3.6: OpenAI Whisker as LID

diarization approach. The results of the ablation study provide valuable insights into the importance of different aspects of our methodology and guide further refinements for optimal performance.

3.4.1 OpenAI Whisper as LID

In this we replaced the complete LID of our pipeline with one of the most popular state-of-the-art audio/speech processing model, OpenAI whisper model [4]. In contrast to wave2vec2 which is an encoder-only, this is an encoder-decoder architecture. We used the pipeline similar to in 3.6. A pretrained base-version of OpenAI whisper was used which is trained on more that 40 languages.

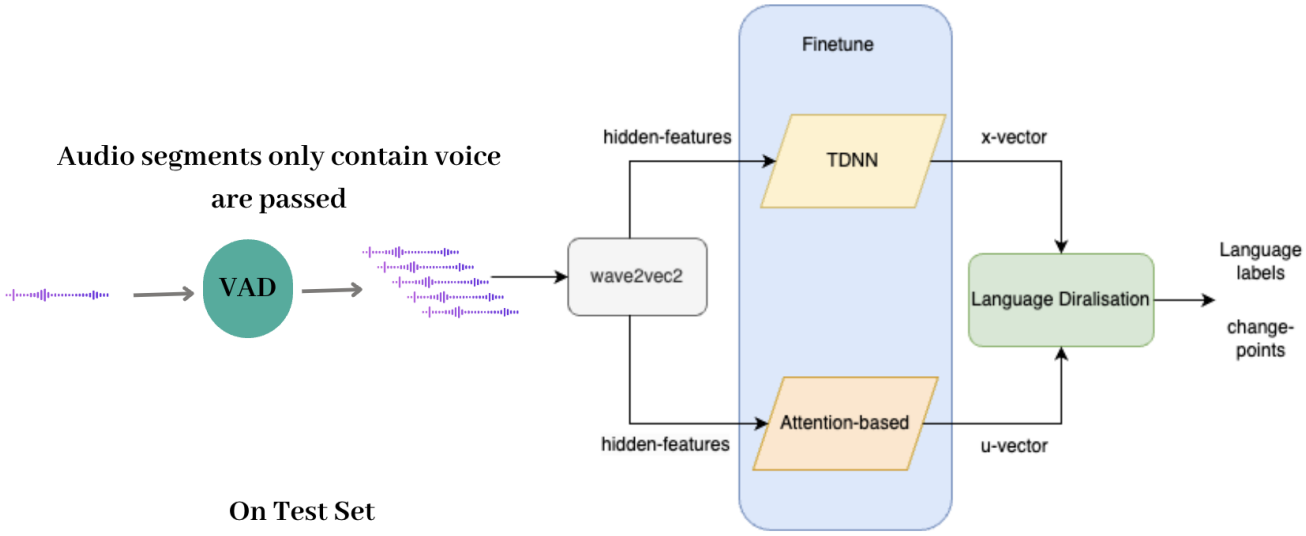

Fig. 3.7: VAD at inference

We achieved a **42.70 %** DER on test displace dataset.

3.4.2 Using VAD (Voice Activation Detection) during Inference

Previously, in our pipeline we are not using VAD during the inference time. We added this component during the inference time. Using this approach, the test audio was broken into various chunks that contains any amount of voice and all the silence will be omitted. The new pipeline can be found in 3.7.

Using this appraoch we achieved a increase in DER (**63.94 %**).Supposedly, VAD is not a good appraoch that works in harmony with our LD approach. The breaking up of main audio into multiple chunks make the transition detection algorithm (Signal processing approach) weak.

3.4.3 Using spring-labs pretrained Wave2vec2

In this method, we made use of pretrained wave2vec2 from spring-labs as a feature extraction block instead of manually finetuning it. The models can be accessed at Models. This model was trained on 12 Indian native languages. We used this model for feature extraction rest of the pipeline remains the same. A DER of **49.70 %** was

achieved using this approach.

Chapter 4

Inferences

4.1 Key Inferences and Explanations

- Almost comparable performance in **Displace Challenge 2024** as compared to the baseline simply represents the power of Language Identification (LID) combined with Language Diarization (LD). These robust x-vector/u-vector embeddings can replace the traditional clustering-based diarization.

- Increasing the window size used in training/validation helped a lot in decreasing Diarization Error Rate (DER).

- **Considering the complexity of Indian languages**, surpassing the linguistic diversity covered by the current model trained on 53 languages, there's potential for improved performance with models trained on more intricate language structures.

- Using OpenAI Whisker did not yield any improvement in the present pipeline. Additionally, adding Voice Activity Detection (VAD) at inference highly degraded the performance of the LD approach. The present approach requires continuous audio without breaks to detect smooth transitions.

- Present approach seems to work fine with the 2 language diralization while an approach for expanding this approach to multilingual codemixed scenarios still

needs to be devised.

Chapter 5

Future Works

5.1 Future Works and Ongoing Explorations

- **Testing Multiple Wave2Vec2 Variants:** Future exploration involves testing at least three variants of Wave2Vec2 with increasing numbers of parameters. This iterative testing approach aims to identify the optimal model configuration for enhanced feature extraction and improved language discrimination.
- **Extension to Other Native Languages:** Successful experiments in Hindi-English code-mixing suggest potential applicability to other native languages mixed with English. This success motivates further investigations into adapting the approach for languages beyond Hindi and English.
- **Continuous Exploration of Innovative Methods:** Continuous exploration of innovative methods and models, alongside staying updated on field advancements, is crucial. This ongoing exploration ensures the refinement and advancement of the proposed approach for multilingual language diarization.
- **Future Plan:**
  - Planning to engage in a significant technical project at IIT Madras Spring Labs during the upcoming semester as part of the NLTM project, of which Dileep Sir is a part.

– Focused on creating **"Indian large foundation models,"** a semi-supervised initiative with the goal of extracting language-specific frame-level features unique to Indian languages, akin to Wave2Vec.

– The project, named "Indic-Wave2Vec," aims to advance the representation and comprehension of Indian languages, drawing inspiration from the success of Wave2Vec.

– This endeavor aligns with our commitment to developing robust and tailored models for Indian linguistic diversity.

5.2 Project Timeline

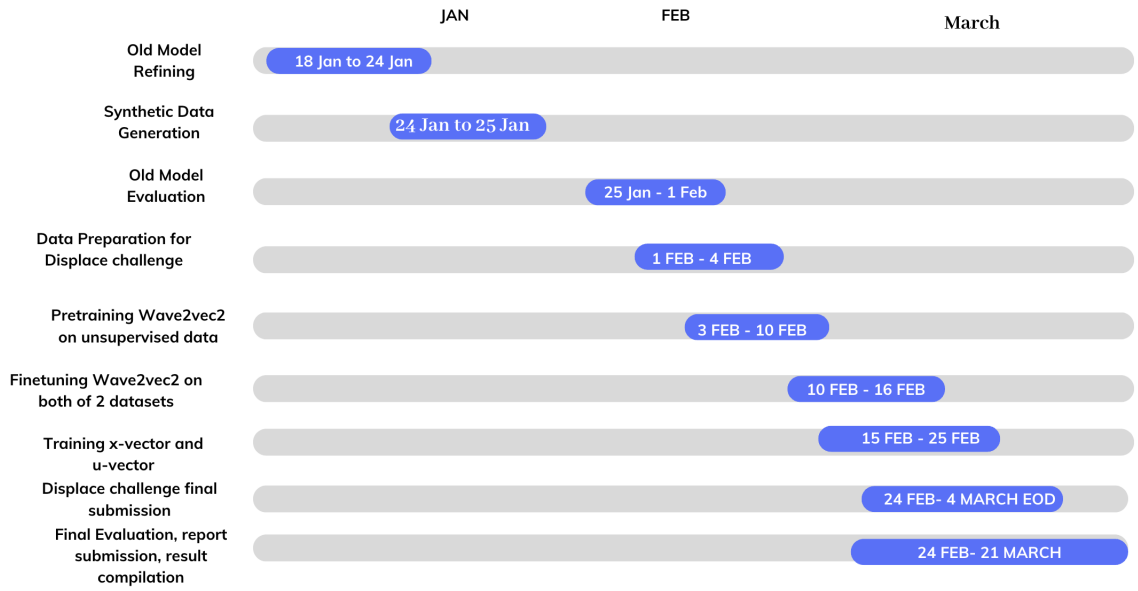

Fig. 5.1: Gantt Chart -Upto Midsem

In this section, we present the project timeline, showcasing the key milestones achieved so far and outlining the planned future works. The Gantt charts below visually represent the timeline of tasks and their durations.

Figure 5.1 displays the Gantt chart representing the tasks completed during the project upto Mid-Sem. The chart illustrates the duration and completion status of each milestone, providing an overview of the progress made.

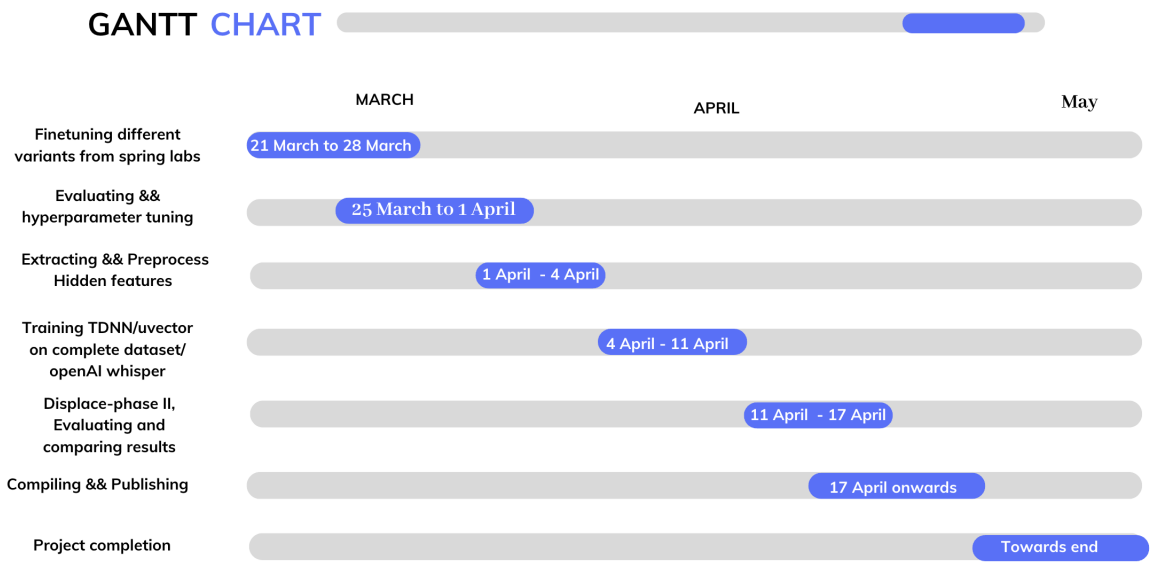

Fig. 5.2: Gantt Chart - After Midsem

The Gantt chart in Figure 5.2 outlines the task done after Mid-Sem Project evaluation which also marks the completion of the present project.

Chapter 6

Conclusion

This project delves into the critical intersection of Language Identification (LID) and Spoken Language Diarization (LD), with a specific focus on Indian native languages. In the era of voice-operated smart devices dominating digital communication, the need for nuanced language understanding is apparent, especially in India's linguistically diverse landscape dominated by code-mixed languages like Hinglish (Hindi mixed with English).

Language Diarization, an extension of LID, offers a solution to the challenges posed by code-mixing and multilingual environments. Despite advancements, persistent challenges include accurately discerning languages and speakers within speech samples, particularly in the context of Indian languages. This thesis contributes to enhancing the robustness of Language Diarization systems tailored for Indian languages, employing innovative approaches, particularly leveraging Wave2Vec models.

Our approach aims to tackle these challenges and, hopefully, yield new results in this domain using Wave2Vec. We eagerly anticipate obtaining valuable insights and results as we continue our work into this, further refining and advancing our understanding of Language Diarization in the context of Indian languages. The current results on the Evaluation Set of the challenge represent the potential in the current pipeline and thus pave the way for further investigation and research in the same domain. All the developments and code can be found at Github

Bibliography

[1] H. Muralikrishna, P. Sapra, A. Jain, and D. A. Dinesh, "Spoken language identification using bidirectional lstm based lid sequential senones," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 320–326.

[2] J. Mishra, A. Agarwal, and S. Prasanna, "Spoken language diarization using an attention based neural network," 06 2021.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 12 2022.