Significance of Indic Self-supervised Speech Representations for Indic Under-Resourced ASR

Sougata Mukherjee(✉), Jagabandhu Mishra, and S. R. Mahadeva Prasanna

Indian Institute of Technology Dharwad, Dharwad, India
{211022004,183081002,prasanna}@iitdh.ac.in

Abstract. Automatic speech recognition is a mature speech technology, almost able to attend human label recognition performance conditioned on the availability of sufficient labeled training data. However, the performance of the system struggles to achieve deployable performance in the under-resourced scenario. In such a scenario, most of the work suggests traditional frameworks are preferable over state-of-the-art deep learning frameworks. This work creates a dataset for the Lambani language of 6 hours duration, and attempts to develop an ASR system. The system provides a character error rate (CER) of 39.1% and 24.1% using the GMM-HMM framework and TDNN framework, respectively for Lambani dataset. The language doesn't have enough publicly available speech and corresponding text transcription resources of its own. Motivating by the same, this work uses the publicly available wav2vec2.0 (W2V) pre-trained model (trained on 23 Indian languages' unlabeled speech data) and fine-tuned it with the labeled data of the Lambani language. After that using the fine-tuned framework as a non-linear feature extractor, the ASR task is performed with GMM-HMM and TDNN framework. The proposed approach provides a relative improvement of 53.4% and 32.1% for the GMM-HMM and TDNN frameworks, respectively.

Keywords: Lambani · Wav2vec2.0 · GMM-HMM · TDNN · W2V Features · MFCC

1 Introduction

Automatic speech recognition (ASR) system converts the spoken utterances to the corresponding textual form. Generally building the state-of-art ASR systems requires a large amount of transcribed speech data having speaker variability, a pronunciation dictionary, and a large amount of text data for building the language model. However, it is difficult to get such repositories for an under-resourced language. On the other hand, to avoid the digital divide and encourage people to use speech-based applications in their own language, it is essential to develop ASR in such languages. Further, such technological intervention in their own language encourages people to use their own language, instead of

A. Karpov et al. (Eds.): SPECOM 2023, LNAI 14339, pp. 100–113, 2023.
https://doi.org/10.1007/978-3-031-48312-7_8

adopting to a resource-rich language. This may help in minimizing the conversion of the resource scare language to the dead language. With this motivation, this work initially attempts to create a dataset suitable for ASR development for the Lambani language. Lambani is a spoken language, doesn't have a written script, and is spoken by the tribal community of Western and Southern parts of India.

Gaussian Mixture Model (GMM)-Hidden Markov Model(HMM) and Time Delay Neural Network (TDNN) [13] are known to be the state-of-art classical approaches for building a speech recognition system. As these approaches do not require a high amount of data, these approaches may be suitable for under-resourced settings [13]. To further improve the performance, the improvement can be done in either of the three levels of the ASR framework, i.e. feature level, modeling level, and decision/hypothesis level. Out of these, this work focuses on the feature level to improve the ASR performance in under-resourced settings.

In wav2vec2.0 in the pretraining stage, the network is trained to predict the masked sub-word units in order to learn about the contextual information [9–11]. In [5], they have shown that when such a learned wav2vec2.0 pre-trained model is used for fine tuning on low-resourced settings it gives a decent performance. Taking motivation from there, in our work, we are using a fine tuned W2V model that has been trained on 23 Indian languages and 10,000 hours of data as a feature extractor. These extracted features are used for GMM-HMM and TDNN training for low-resourced settings in order to get improved performance

The rest of the paper is organized as follows:- Section 2 gives a description of how the lambani corpora was built. Section 3 shows the proposed framework for this paper. Section 4 gives a brief description about the available resources. Section 5 shows the results obtained using MFCC and speech representation extracted from self-supervised wav2vec2.0 approach. Finally, Sect. 6 concludes the report.

2 Building Lambani Corpora

2.1 Text Data Collection

Lambani is a language which has a spoken form, but written form of Lambani is not available. So, the written form of Lambani had to be prepared with a lot of manual effort and time. We had to make sure that the native Lambani speakers could articulate the words easily. So, initially 1000 English sentences containing a swadesh list [16] of words were prepared taking help from a Linguist. These are 4 to 10 word sentences. Examples of short and long sentences from the swadesh list include "All kids want sweet" and "Before I went to her house I changed my clothes". An ASR system requires several hours of speech data to train. The duration of the recording of the sentences containing Swadesh list of words vary from 2 to 4 words. So, the number of sentences had to be increased following the same procedure as discussed above. During the increment in the number of sentences text had to be extracted from several sources. Major sources of English sentences were NCERT and Wikipedia. Among the books published by NCERT, we focused on English language textbooks meant for the students of

the lower, middle, and higher secondary schools. For retrieving text, we used text extractors written by us using the Python language. We used optical character recognition to extract sentences from publicly available scanned versions of the books on Lambani and languages using Adobe Reader's API.

2.2 Text Processing

The text data extracted from various sources like NCERT, wikipedia quite often contain incomplete sentences, semantically incorrect sentences, and long sentences which are difficult to speak. Hence The following preprocessing steps were applied to the raw text to improve the readability of the sentences.

- The passages of extracted text were processed to derive a set of sentences. The sentences containing fewer than 3 words were eliminated. Sentences longer than 10 words were removed as they will be difficult to utter for illiterate or older tribal people.
- Incomplete sentences, syntactically or semantically incorrect sentences, and sentences containing symbols and characters not present in the Roman script were removed.
- Sentences containing words that may be too complex for a tribal person to speak were discarded. Text containing controversial statements including political statements was removed from the set of sentences.

The English sentences that successfully passed through the above-mentioned preprocessing steps qualify to be a part of the sentence corpus. The selected English sentences were converted to the Kannada language (contact language) using the Kannada script as it was the formal language in the area. Then, those Kannada sentences were translated to the Lambadi language using the Kannada script by the Lambani native speakers. The Lambani text data in Kannada script was preserved in digital format by writing them in a spreadsheet.

2.3 Speech Files Recording

These sentences collected as text data were spoken by multiple native Lambani speakers which was recorded using Laptop. Graphical User Interface (GUI) was designed to collect data through Laptop. The Lambani sentence is displayed on the GUI. The recorded voice is replayed to assess its quality. If necessary, the GUI offers a feature to re-record the current sentence. Every speaker will record about one hour of data in seven sessions, which means almost 100 recordings per session. The GUI which is used for ASR recording has been shown in Fig. 1

The entire process of data collection strategy can be summarized in the flowchart 2

3 Related Work and Motivation

Efforts have been made to build speech recognition systems for under-resourced languages. But, due to the lack of resources, it becomes very difficult to

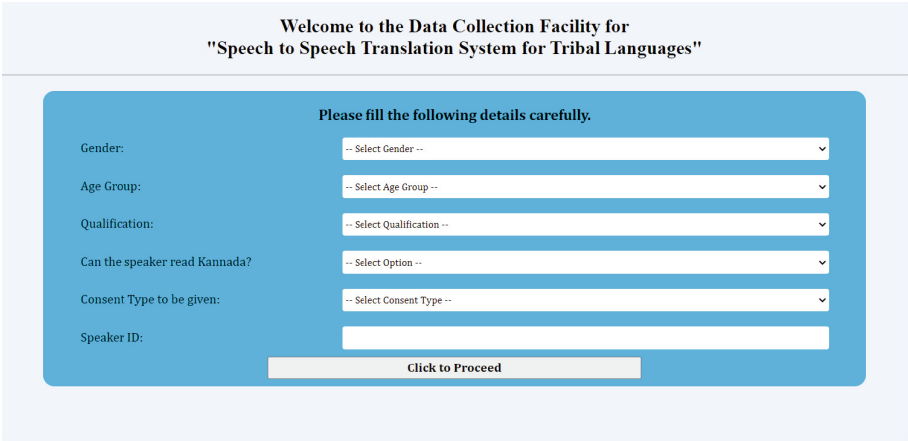

Fig. 1. GUI used for ASR recording

build such systems. Still, people have undergone research in this field applying various methodologies to overcome the challenges. Initially, ASR building for under-resourced language started with cross-lingual adaptation [8] of the Vietnamese language. They also tried to show the potential of cross-lingual context-dependent and independent modeling in this task. The same paper shows grapheme-based acoustic modeling when there is an absence of a pronunciation dictionary. With grapheme based acoustic modeling the performance that they have achieved using data-driven approach is an SLER of 43.9% and a WER of 50.6%.

In [18] they have shown that the performance obtained (36.5% accuracy) from a 30 dimensional posterior features multi-layer perceptron is trained on 15 hrs of German and 16 hrs of Spanish which is adapted to 1 hr of English so that several phonetic attributes of speech get covered from the out-of-language data. Almost in a similar approach, people showed the importance of bottleneck features and tandem features extracted from multi layer perceptron trained 15 hrs of German,16 hrs of Spanish trained on 1 hr of English (low-resource setting) for low-resource large vocabulary continuous speech recognition task [19]. In [6] they have explored multilingual information with KL-HMM when the available data is less than 75 mins and showed how the accuracy of KL-HMM varies with respect to other systems with increase in training data from 5 mins to 808 min. [17] they have explored the performance using both longer acoustic units which are syllables and shorter lexical and language modeling units i.e. morphemes. They have decomposed the rare syllables (if the number is less than 17.9%) into phones and trained the hybrid system. Importance of out-of-domain language data to improve the performance of under-resourced speech recognisers is shown in [7]. Here, they have shown that they are improving the performance using out-of-domain data that using 81 hrs of Dutch data along with 3 hrs of Afrikaans data they are achieving a improved phone accuracy of 68.8% with respect to

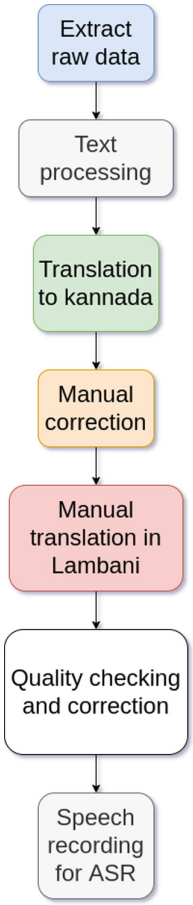

Fig. 2. ASR corpus creation flowchart

the monolingual system which gives a phone accuracy of 60.6%. They have also shown that KL-HMM is giving the best performance after acoustic model adaptation as compared to MLP, MLLR and MAP techniques. In [15] they have discovered a new speech feature named Intrinsic Spectral Analysis (ISA) which is performing better than FBANK,MFCC and PLP features. It gives a phone error rate of 10.42% on a training and testing set of 10.7 hrs and 2.2 hrs of Afrikaans language data. [2] shows that they have achieved best performance of 5.6%for Afrikaans language using multitask learning where they are learning the triphone senones and trigrapheme senones of multiple phonemes and a universal phone set and grapheme set which contains all the phones and graphemes of all the under-resourced languages Afrikaans, Siswati and Sesotho. The entire literature review is summarized in Table 1. [4,20] showed that training pre-trained weights helps to regularize and converge better rather than random initialization. Wav2vec2.0 is known to learn speech representations. So,in our work we

Table 1. Literature review of speech recognition systems for under-resurced languages

Authors	Language	Dataset	Techniques	Performance
Viet-Bac Le et al.	Vietnamese	Training-14 hrs Testing-408 sentences by 3 spkrs	Grapheme based AM	SLER-43.9% WER-50.6%
			Context dependent cross-lingual model adaptation	SLER-36.6% WER-42.7%
Thomas S et al.	English	Training-15 hrs German, 16 hrs Spanish, 1 hr English Testing-1.8 hrs	30D Multi-stream cross-lingual posterior features	Accuracy- 36.5%
Thomas S et al.	English	Training-15 hrs German, 16 hrs Spanish, 1 hr English Testing-1.8 hrs	DNN features	WA-41%
Imseng et al.	Afrikaans	Training-81 hrs Dutch, 3hrs Afrikaans Testing-50 mins	KL-HMM	PA-68.8%
Tachbelie et al.	Amharic	Training-20 hrs Testing-5K set from ATC_120K Corpus	Hybrid acoustic units (phones and syllables)	WER-17.9%
Sahraeian et al.	Afrikaans	Training-10.7 hrs Testing-2.2 hrs	Intrinsic spectral analysis for SGMM	PER-10.2%
Dongpeng et al.	Afrikaans	Train-3.37 hrs	Multitask learning of trigpaheme senones and triphone senones	WER-5.6%

are trying to learn speech representation from multiple Indic languages and then converge the weights accordingly to a particular language of an under-resourced setting.

4 Proposed Framework

As discussed in the previous section, self-supervised learning representations give better results as compared to hand-crafted features . So, MFCC is being replaced by self-supervised speech representations in order to get better performance for under-resource settings. For the build up of these framework we are adopting the following strategies:-

4.1 Character Level Speech Recognition

In this work, kaldi recipes have been used for building frameworks for GMM-HMM and TDNN. Generally, the GMM-HMM and TDNN recipes in kaldi [14] take either word level transcription or phone level transcription as input for training the model and give word level or phone level transcription as decoded output. But in our case due to the absence of a lexicon for under-resourced language Lambani we first built systems which take character level transcription as input for building the trained model and gives character level transcription as decoded output.

For, GMM-HMM framework we have used the TIMIT recipe. There we have replaced the phone level transcription training and testing text files with character level transcription in the data preparation stage. Initially, for speaker independent GMM-HMM training MFCC features, Δ and $\Delta\Delta$ were used. Here, the

MFCCs were subject to cepstral mean variance normalization. But, for speaker dependent case Feature space Maximum Likelihood Linear Regression (FMLLR) features were used [3]. The frame shift and frame width are 10 ms and 25 ms respectively. GMM-HMM acoustic model(AM) was trained using maximum likelihood(ML) condition. Along with these a bi-gram statistical language model (LM) was used while decoding. The training and decoding strategies are shown in the Fig. 3(a)

TDNN is good for modeling long-range temporal dependencies. In the case of TDNN framework, Mini-Libispeech chain recipe in kaldi was followed which uses a factorised TDNN network. In this case, 40 dimensional MFCC and 100 dimensional i-vector was used for every time step. A 13 A Lattice-Free (LF) variant of the Maximum Mutual Information (MMI) criterion is used for chain model training without frame-level cross-entropy pre-training. The training and decoding strategy are shown in the Fig. 3(b)

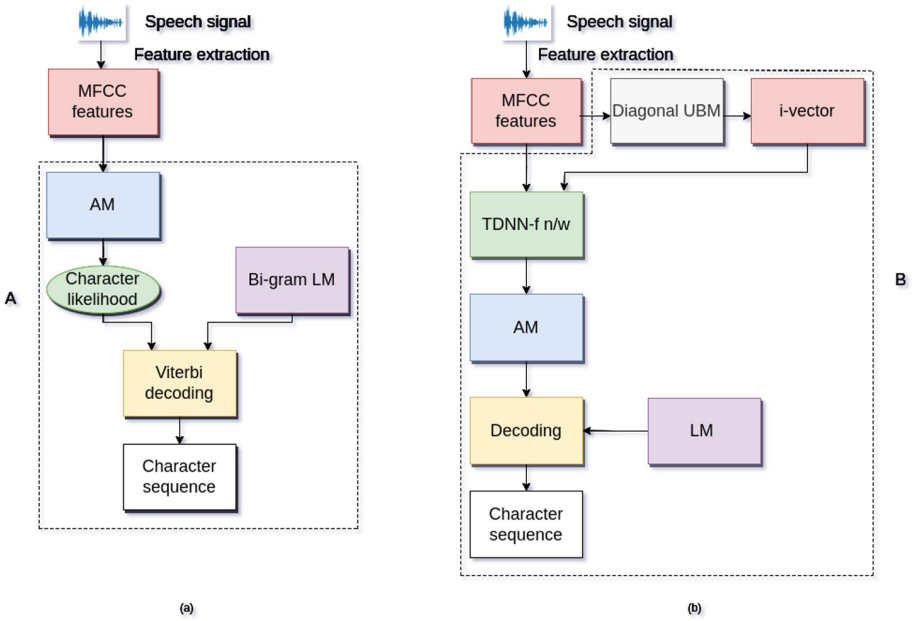

Fig. 3. (a) GMM-HMM framework using MFCC (b) TDNN framework using MFCC

4.2 Wav2vec2.0 Feature Extraction

Before feature extraction wav2vec2.0 involves 2 steps namely pre training and finetuning which are described as follows:-

The pretraining network is shown in the Fig. 4. The latent representations (Z) from raw speech are extracted using convolutional neural networks. The latent representations are quantised to discrete units(Q) which act as targets during the contrastive task. The latent representations are masked randomly and fed to the transformer allowing the network to predict the context representation of the masked regions. Q is compared with C using contrastive and diversity loss. Multilingual pretrained model(CLSRIL-23) [5] which has been trained on 23 indian languages to learn the contextual speech representations has been used here.

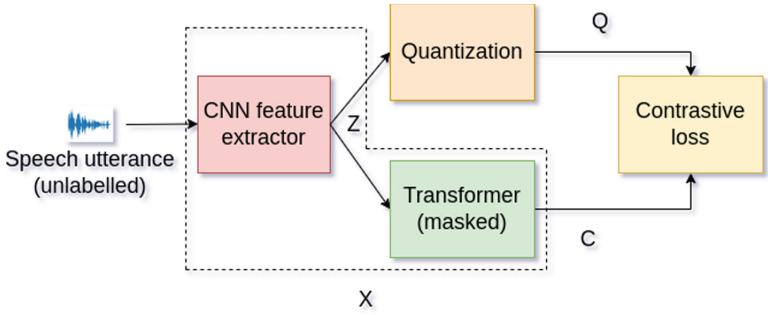

Fig. 4. Wav2vec2.0 pretrainig architechture

The CLSRIL-23 model has been fine tuned on both Mini-Librispeech and Lambani dataset. The fine tuning framework is shown in Fig. 5. During fine tuning as shown in the figure the network X is borrowed from pretraining architecture and a randomly initialized softmax linear layer is added on top of it which is optimized using connectionist temporal classification (CTC). The size of the linear layer is equal to the vocabulary size (V) of the language.

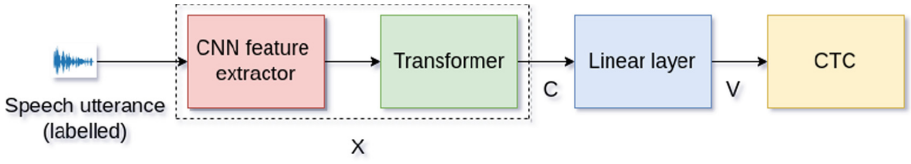

Fig. 5. Wav2vec2.0 finetuning framework

4.3 Modification of Features

CLSRIL-23 pre-trained model is finetuned and has been used as a feature extractor in our proposed framework. In CLSRIL-23 base architecture of the wav2vec2.0 framework is used which contains 12 transformer blocks with a model

dimension of 768. Wav2vec2.0(W2V) features which are of 768 dimension is extracted from the last layer of transformer encoder of the fine tuned model with a frame shift of 20 ms and a frame size of 25 ms of the speech signal [10]. These features are used to train the GMM-HMM and TDNN framework in place of MFCC and the rest of the procedure is kept intact. The blocks A and B are the same as shown in Fig. 3(a) and 3(b), only the feature is replaced with W2V features.

5 Database Settings

Two types of language datasets have been used for performing the experiment, those are English and Lambani and the name of the datasets are Mini-Librispeech and Lambani respectively.

The Mini-Librispeech dataset is a subset of the Librispeech data [12]. The Mini-Librispeech dataset comes with transcribed training and testing sets named as train-clean-5 and dev-clean-2 respectively. Train-clean-5 and dev-clean-2 contain 5 hrs and 2 hrs of speech data. Speakers from the dev-clean-2 are chosen randomly so that it sums up to 1hr of testing data. The training set of Mini-Librispeech contains 12 males and 16 females speaker data i.e. 28 speakers data in total, whereas the testing set contains 9 males and 3 females i.e. 12 speakers in total. The speech file has 16 kHz sampling frequency and bit rate of 256 kbits/sec. The number of channels for each speech file is 1. There are 1519 utterances and audio files in the training set and 534 utterances and audio files in the testing set. That means there is one audio file corresponding to each utterance.

The Lambani dataset comes with raw audio files and its corresponding text transcriptions along with the utterance ids. For each utterance there is only one audio file corresponding to that particular utterance. Initially the sampling rate of the speech files were 44.1 kHz which has been changed to 16 kHz which is compatible for all the frameworks. The Lambani training dataset has 7 males and 8 females speaker data which is 15 speakers data in total and the testing data contains 2 females and 1 male speaker data which is 3 speakers data in total. The speech file has a bit rate of 256k and the number of channels of each audio file is 1. The total number of utterances and audio files in the testing set are 770. Here, also there is one audio file corresponding to each utterance.

A summary of the entire dataset is shown in Table 2.

5.1 Experimental Results and Discussion

For GMM-HMM and TDNN based frameworks we have carried out all the experiments using kaldi and for Wav2Vec2.0 we have carried out all the experiments using vakyansh [1] toolkit. We have carried out the experiments with MFCC features and the speech representations from Wav2Vec2.0 which we are calling here as W2V features.The best experimental results are shown in the Table 3.

Table 2. Specification of the dataset after preprocessing and organisation

Parameters	Mini-Librispeech	Lambani
Amount of training data	5 hrs	5 hrs
Amount of testing data	1hrs	1 hrs
No. of spkrs	28(12 males, 16 females) in taining set	12(9 males, 3 females)
	15(7 males, 8 females)	3(2 males, 1 female)
Sampling rate	16 kHz	16 kHz
Bit rate	256k	256k
Channels	1	1
Utterances	1519 in training set	3390 in training set
	534 in testing set	770 in testing set
No. of Audio files	1519 in training set	3390 in training set
	534 in testing set	770 in testing set

Table 3. CER(%) for different frameworks with MFCC and W2V features

FRAMEWORKS	FEATURES	DATASET	
		Mini-Librispeech	Lambani
GMM-HMM	MFCC	38.5	39.1
	W2V	18.2	18.2
TDNN	MFCC	24.1	30.8
	W2V	15.2	20.9

While using MFCC features for building acoustic models for character level ASR systems using GMM-HMM and TDNN frameworks TIMIT recipe and Mini-Librispeech chain recipe from kaldi were used respectively. i-vectors were used in addition to MFCC in the TDNN framework. A bi-gram language model was used while decoding which was built using the text of the entire training data. IRSTLM toolkit was used to build the language model. 5 hours of transcribed Mini-Libispeech and Lambani were used for training and building the acoustic model for the GMM-HMM and TDNN framework.

So, here we can see that W2V features are performing better than the MFCC features for both the frameworks. We have considered speaker adaptation while carrying out the results of Mini-Librispeech but in the case of Lambani it wasn't considered.

Sample for a particular utterance of the decoded transcript for both Lambani and Mini-Librispeech dataset is shown below .
TV stands for truth value
GHM stands predicted sequence for GMM-HMM using MFCC
TM stands predicted sequence for TDNN using MFCC
GHW stands predicted sequence for GMM-HMM using W2V features
TW stands predicted sequence for TDNN using W2V features.

D stand for deletion
I stands for insertion
S stands for sustitution

Mini-Librispeech

TV:	* s h e w a s i n d e e d a c l e v e r b i r d
GHM:	T s h e w a s i n d e e d a c l A v e r P E r d
Eval:	I S S S
GHW:	W s h e w a s i n d e e d a c l e v e r b i r d
Eval:	I
TM:	* s h e w a s E n d e e d a c l I v e r b U r d
Eval:	S S S
TW:	* s h e w a s i n d e e d a c l e v e * b i r d
Eval:	D

Lambani

Fig. 6. Lambani predicted text showing its alignment with ground truth

So, we can see that in case of Mini-Librispeech dataset we are getting the least error in the decoded transcription for TDNN framework for W2V features i.e. TW. GHW is also performing better than TM as it has 3 character errors whereas TM has one character error. Among all these GHM is giving the worst result with 4 character error as its performance is the poorest. Here, among the predicted characters the characters which are inserted and substituted are marked in capital letters.Hence the result is justified

As shown in Fig. 6 in the case of the Lambani dataset, GHW is giving the best performance with only 6 character errors. GHM is the worst one with 11 errors. TM has 9 errors and TW has 8 errors so TM has better performance than TW. Hence the result is verified.

6 Conclusion and Future Work

In this work, data collection strategy for an under-resourced language Lambani has been shown. Character level speech recognition for under-resourced settings

using GMM-HMM and TDNN has also been explored in this paper. The focus was to show the significance of self-supervised speech representations extracted from wav2vec2.0 for under-resourced settings. So, using the wav2vec2.0 approach as a non-linear feature extractor we are getting a relative improvement of 53.4% and 32.1% for GMM-HMM and TDNN frameworks respectively for under-resourced language Lambani. In the Mini-Librispeech dataset, the relative improvement in the performances are 35.5% and 36.92% for GMM-HMM and TDNN frameworks respectively. Similarly, as a part of future work, this approach can be explored in case of other pretrained models and other deep self-supervised learning methodologies can be explored as a feature extractor in place of wav2vec2.0.

Acknowledgements. The Lambani data collection is a part of the"Speech to Speech translation project". The authors would like to acknowledge the Ministry of Electronics and Information Technology (MeitY), Govt. of India, for funding us in this project. The authors would also like to thank the data associates who have helped in collecting Lambani data. The authors are grateful to Mr.Swapnil Sontakke for building the GUI which played a crucial role in collection of data.

References

1. Chadha, H.S., et al.: Vakyansh: ASR toolkit for low resource Indic languages. arXiv preprint arXiv:2203.16512 (2022)
2. Chen, D., Mak, B.K.W.: Multitask learning of deep neural networks for low-resource speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(7), 1172–1183 (2015)
3. Gales, M.J.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
4. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
5. Gupta, A., et al.: CLSRIL-23: cross lingual speech representations for Indic languages. arXiv preprint arXiv:2107.07402 (2021)
6. Imseng, D., Bourlard, H., Garner, P.N.: Using kl-divergence and multilingual information to improve ASR for under-resourced languages. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4869–4872. IEEE (2012)
7. Imseng, D., Motlicek, P., Bourlard, H., Garner, P.N.: Using out-of-language data to improve an under-resourced speech recognizer. Speech Commun. **56**, 142–151 (2014)
8. Le, V.B., Besacier, L.: Automatic speech recognition for under-resourced languages: application to Vietnamese language. IEEE Trans. Audio Speech Lang. Process. **17**(8), 1471–1482 (2009)
9. Mishra, J., Gandra, J., Patil, V., Prasanna, S.R.M.: Issues in sub-utterance level language identification in a code switched bilingual scenario. In: 2022 IEEE International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2022)

10. Mishra, J., Patil, J.N., Chowdhury, A., Prasanna, S.M.: End to end spoken language diarization with wav2vec embeddings
11. Mishra, J., Prasanna, S.R.M.: Importance of supra-segmental information and self-supervised framework for spoken language Diarization task. In: Prasanna, S.R.M., Karpov, A., Samudravijaya, K., Agrawal, S.S. (eds.) International Conference on Speech and Computer, vol. 13721, pp. 494–507. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20980-2_42
12. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. IEEE (2015)
13. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Sixteenth Annual Conference of The International Speech Communication Association (2015)
14. Povey, D., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. No. CONF, IEEE Signal Processing Society (2011)
15. Sahraeian, R., Compernolle, D.V., Wet, F.d.: Under-resourced speech recognition based on the speech manifold. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
16. Swadesh, M.: Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north American Indians and Eskimos. Proc. Am. Philos. Soc. **96**(4), 452–463 (1952)
17. Tachbelie, M.Y., Abate, S.T., Besacier, L.: Using different acoustic, lexical and language modeling units for ASR of an under-resourced language-Amharic. Speech Commun. **56**, 181–194 (2014)
18. Thomas, S., Ganapathy, S., Hermansky, H.: Cross-lingual and multi-stream posterior features for low resource LVCSR systems. In: Eleventh Annual Conference of the International Speech Communication Association (2010)

19. Thomas, S., Ganapathy, S., Hermansky, H.: Multilingual MLP features for low-resource LVCSR systems. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4269–4272. IEEE (2012)
20. Yu, D., Deng, L., Dahl, G.: Roles of pre-training and fine-tuning in context-dependent DBN-HMMS for real-world speech recognition. In: Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning. sn (2010)