Evaluating the Efficacy of Large Acoustic Model for Documenting Non-Orthographic Tribal Languages in India

Tonmoy Rajkhowa, Amartya Roy Chowdhury, Hrishikesh Ravindra Karande, S.R. Mahadeva Prasanna
Indian Institute of Technology, Dharwad
{212022001, amartya.chowdhury, 210010020, prasanna}@iitdh.ac.in

Abstract

Pre-trained Large Acoustic Models, when fine-tuned, have largely shown to improve the performances in various tasks related to spoken language technologies. However, their evaluation has been mostly on datasets that contain English or other widely spoken languages, and their potential for novel under-resourced languages is not fully known. In this work, four novel under-resourced tribal languages that do not have a standard writing system were introduced and the application of such large pre-trained models was assessed to document such languages using Automatic Speech Recognition and Direct Speech-to-Text Translation systems. The transcriptions for these tribal languages were generated by adapting scripts from those languages that held a prominent presence in the geographical regions where these tribal languages are spoken. The results from this study suggest a viable direction to document these languages in the electronic domain by using Spoken Language Technologies that incorporate LAMs. Additionally, this study helped in understanding the varying performances exhibited by the Large Acoustic Model between these four languages. This study not only informs the adoption of appropriate scripts for transliterating spoken-only languages based on the language family but also aids in making informed decisions in analyzing the behavior of particular Large Acoustic Model in linguistic contexts.

Keywords: Large Acoustic Models, Automatic Speech Recognition, Direct Speech-to-Text Translation

1. Introduction

Many of the languages spoken around the world lack a standardized writing system making them under-represented in the digital domain. To preserve such non-orthographic languages, one can record the conversations and store them in the form of audio. This approach may help in documenting the cultural nuances and transmission of wisdom and knowledge only to a limited audience who are well-versed with that language. However, to make the language more accessible to a wider audience requires documenting in a written form. With the absence of a written script, transliterating into neighboring languages that has a standard writing system with similar acoustical characteristics and is geographically closer may be a viable alternative solution for documentation. Subsequently, these transliterated texts can be translated into various languages broadening their appeal to a global audience for research purposes. To achieve this, leveraging Spoken Language Technologies (SLT), such as Automatic Speech Recognition (ASR) and Direct Speech-to-Text Translation (DS2TT) may prove instrumental. With this context, this work will attempt to investigate the effectiveness of the proposed approach by building corpora for four such low-resourced under-represented tribal languages of India and developing ASR and DS2TT for their documentation. The four languages were Soliga (Spandana et al., 2023) and Lambani (Naik and Naik, 2012; Boopathy, 1972; Chowdhury et al., 2022) spoken in the state of Karnataka and Kui (Winfield, 1929) and Mundari (Osada, 2008) spoken in the state of Odisha in India. Since these languages lack their distinct writing systems, the recorded audio content will be transliterated into Kannada for Soliga and Lambani and Odia for Kui and Mundari.

Automatic Speech Recognition (ASR) involves transcribing speech spoken in a language into its textual form (Baevski et al., 2020; Bérard et al., 2018) whereas Speech-to-Text Translation (S2TT) translates the speech spoken in one language into the text of another (Berard et al., 2016; Bansal et al., 2017). Traditionally, ASR systems relied on statistical methods for transcription (Anantaram et al., 2016; Bassil and Alwani, 2012; Bohac et al., 2012; Cucu et al., 2013; Shugrina, 2010). Similarly, for S2TT tasks, the ASR-generated text underwent further translation into desired languages using Statistical Machine Translation (SMT) systems. Consequently, the S2ST task involved two distinct systems, *viz.* ASR and SMT, in a cascaded pipeline. However, with the advent of Deep Learning Encoder-Decoder architectures, traditional ASR and MT systems transitioned into a neural architecture and has outperformed traditional statistical approaches. Additionally, because of Deep Learning approaches, the neural-based ASR and MT systems were merged into a unified architecture called Direct Speech-to-Text

LREC-COLING 2024, pages 6475–6483
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Translation (DS2TT) system (Waibel and Fugen, 2008; Duong et al., 2016). This DS2TT system became more computationally efficient compared to the cascaded approach due to the involvement of a single system. Nevertheless, these Deep Learning systems require a substantial amount of data as these systems are known to be data-intensive. To meet this data requirement, various techniques, such as Data Augmentation, ASR pre-trained encoders, and recently Large Acoustic Models (LAMs), trained using thousands of hours of speech have been employed. These LAMs achieved state-of-the-art performances on many well-established corpora. However, their performances were demonstrated on those datasets whose languages, either in spoken or textual form, were part of the training data used for developing the LAMs. For instance, LAMs excelled when evaluated on corpora containing English or other well-known languages but for less popular languages, i.e. languages that have not been included during its building and under-resourced scenarios, the performance remains uncertain. This motivated us to take the work in this direction to investigate the performance of LAMs fine-tuned on the four novel under-resourced corpora for both ASR and DS2TT tasks.

The advent of the Transformer (Vaswani et al., 2017b) shifted the dynamics of Deep Learning technologies leading to a transition from Recurrent Neural Network (RNN) based Encoder-Decoder architectures. This led to the development of state-of-the-art ASR as well as DS2TT systems with improved performances over its recurrent counterparts. The efficacy of ASR systems received an additional boost with the development of pre-trained Large Acoustic Models(LAMs), such as Wav2Vec2 developed by Huggingface (Baevski et al., 2020), Conformer-CTC by Nvidia-NeMo (Peng et al., 2021), Whisper by Open-AI (Radford et al., 2023). When fine-tuned on standard corpora, these models produced benchmark results. However, it is worth noting that these models performed very well on standard datasets containing well-known languages, but their evaluation of novel languages remains an area of exploration. Hence, this work will attempt to assess the performance of one of these LAMs, i.e. Whisper Large V2 pre-trained model, on four previously unencountered tribal languages, *viz.* Soliga, Lambani, Kui, and Mundari, shedding light on its adaptability to linguistically diverse and under-resourced contexts.

The primary objective of this study is to introduce a novel approach for documenting the four non-orthographic tribal languages that lack a standardized writing system in the digital platform by incorporating speech technologies. Given the low-resourced nature of these languages, *i.e.* duration of audios for training is less than 12 hours, this work also aims to explore the utilization of Large Acoustic Models to enhance the performances, particularly in the domains of ASR and DS2TT systems. This work endeavors to provide a method for documenting and preserving such non-orthographic low-resourced languages in the digital domain. This paper is organized as follows: Section 2 describes the Experimental Methodology, including the description of the languages and corpora creation methodology, the Large Acoustic Model used for fine-tuning, and tools to develop the ASR and DS2TT systems along with their evaluation methodologies. Section 3 delves into the Results and Discussions. Conclusions and Future Works were presented in Section 4.

2. Experimental Methodology

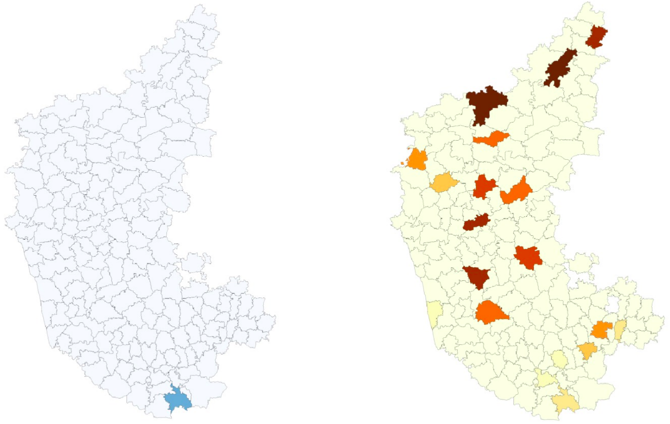
This section describes the languages and their data collection and corpus creation strategies along with the details of the Large Acoustic Model that is used for fine-tuning for implementing the Automatic Speech Recognition and Direct Speech-to-Text Translation systems.

2.1. Language Descriptions

The section describes the four non-orthographic low-resourced tribal languages that were chosen for this work.

Lambani Lambani (K.Ramalingareddy, 2023) belongs to the Indo-Aryan family of languages spoken by nomadic tribes across various regions of Western and Southern India. With a population of approximately 68.9 million speakers, Lambani exhibits a variety of dialects and accents influenced by the major language of the geographical areas where it is spoken. This language lacks a standardized writing system, thereby relying on oral medium for the transmission of linguistic and cultural nuances. For this study, the dialect of Lambani spoken in the northern region of Karnataka is chosen. This particular dialect of Lambani is influenced by the Kannada language, which is the predominant language of Karnataka. As a result, it has evolved as a blend of Indo-Aryan and Dravidian languages.

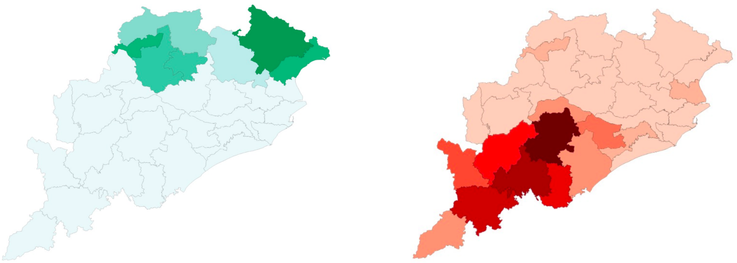
Soliga Soliga (Morlote et al., 2011) is a language spoken by the Soliga tribe, residing in the landscapes of Biligiri Rangaswamy Hills (B. R. Hills) of Southern Karnataka. This language is a part of the Dravidian language family spoken by a small population of 40 thousand individuals. Dedicated linguists and researchers in collaboration with the



(a) Regions of Karnataka where Soliga is spoken

(b) Regions of Karnataka where Lambani is spoken

Figure 1: Maps of Karnataka illustrating the locations where Soliga and Lambani are spoken. Darker shades represent the area that contains the majority of speakers.



(a) Regions of Odisha where Mundari is spoken

(b) Regions of Odisha where Kui is spoken

Figure 2: Maps of Odisha representing locations where Mundari and Kui are spoken. Darker shades represent the area that contains the majority of speakers.

Soliga community have taken numerous initiatives to document and revitalize the Soliga language to ensure continuous usage and transmission to future generations thereby preserving its cultural and linguistic heritage.

Kui Kui (Winfield, 1929), also known as Kandh, Khondi, Kanda, Kodu, or Kuinga, is an oral South-Eastern Dravidian language spoken in the hilly and tribal regions of Odisha in Eastern India. This language, native to a tribal community called Kandhas or Kondhs, uses this language as a means of communication for various cultural practices and rituals. Despite its cultural significance, Kui faces the threat of extinction due to the growing influence of the dominant Odia language. Addressing this concern, researchers and linguists are dedicatedly working to preserve this language.

Mundari Mundari (Wolf-Sonkin, 2021), an Austro-asiatic language, spoken by the Munda tribes in the state of Odisha, India. It does not possess a script but is typically transcribed using the Devanagari or Odia alphabets. This language serves as a means of communication for at least 2.23 million people. One distinct feature of this language is the absence of word classes due to which the nouns, verbs, and adjectives can be distinguished by relying on contextual cues only.

Maps depicting the geographical regions where these four tribal languages were spoken can be visualized in Figures 1 & 2. Figure 1 illustrates the regions where Soliga and Lambani are spoken in the maps of Karnataka whereas Figure 2 illustrates the maps of Odisha portraying the regions in which Kui and Mundari is spoken. The density of speakers is denoted by varying shades of color, with darker shades indicating a higher density of speakers whereas lower shades indicate lower density.

2.2. Corpora Creation Methodology

In this section, the proposed approaches adopted to develop both the speech and text corpora for the tribal languages were given. The overview of the corpora creation methodology is illustrated in Figure 3. The overall process consists of the following stages: (1) Relevant data collection; (2)

Data preprocessing; (3) Translation to contact languages (Kannada and Odia); (4) Audio recordings; and (5) Manual quality checking

2.2.1. Data Collection

To create high-quality data, the following steps adopted are summarized below:

1. **Text compilation from various sources:** Optical Character Recognition (OCR) feature of Adobe Reader has been employed to extract sentences from textbooks. Additionally, English texts from English courses of the National Council of Educational Research (NCERT) textbooks (of Educational Research and Training) were extracted. The focus has been centered on textbooks intended for lower and middle schools. Furthermore, with the involvement of a linguist, a list of 1000 swadesh sentences (Hymes, 1970) was compiled which contained sets of basic English words that cover the fundamental concepts of English grammar.

2. **Data Preprocessing**: The extracted texts frequently include a substantial amount of noise, which poses a challenge for the native speaker in providing accurate translations. To address this issue, the extracted texts were subjected further to the following data preprocessing techniques.

   - It has been observed that the tribal speakers typically communicate using concise and simple sentences. Therefore, sentences containing less than three and more than ten words were discarded.
   - Incomplete sentences with unclear meanings were removed.
   - Manual quality checking is done by a linguist and any sentences found to be syntactically or semantically incorrect were removed.
   - Sentences containing URLs, and unknown characters were removed.

3. **Relevancy pruning** Every sentence has been assessed for its relevance and assigned a ranking, where a score of 1 indicates relevance, and 0 signifies irrelevance based on the subject matter related to daily conversations by tribals. For instance, sentences containing controversial or political statements, which were not typically used in daily conversational activities, were marked as irrelevant and subsequently removed. Following this refinement, approximately 80% of the sentences were retained. This led to the creation of 10,000 textual sentences in English.

4. **Translation to Contact language**: The Lambani and Soliga speakers primarily residing in Karnataka were fluent in Kannada where Kannada serves as the dominant language. Similarly, Mundari and Kui speakers were fluent in Odia, given that Odia is the principal language of Odisha. Therefore, Kannada and Odia were selected as contact languages that have their orthography. The compiled set of 10,000 English sentences were then subsequently translated into Kannada by employing bilingual experts, proficient in both English and Kannada and into Odia, by experts proficient in English and Odia.

5. **Contact language to Tribal language translation**: The translated 10,000 Kannada sentences were then again translated to Soliga and Lambani by the native speakers using Kannada script whereas Odia is translated again to Kui and Mundari using Odia script. The translated texts were subjected to further validation by various bilingual experts employed to ensure the translation quality.

6. **ASR and TTS Recordings** The translated sentences in all the four tribal languages were subsequently recorded by native speakers and the quality check of the recorded audios were performed by language experts to ensure high quality. The recordings were carried out in a studio environment for TTS recordings and in normal environments too, such as open fields etc, for ASR recordings. This created the corpora for both ASR and TTS task. However, the corpora created for ASR task was used for this study.

Table 1 shows some examples of English sentences with their translations that were transliterated in Lambani and Soliga using Kannada script. Corresponding examples from English translated to Kui and Mundari were shown in Table 2.

2.3. Data Preparation and Preprocessing

A comprehensive set of 10,000 parallel sentences were created for all the tribal languages along with the contact languages. These sentences were then divided into train, validation, and test sets while maintaining the parallelism among the sets. A total of 9,000 sentences were allocated for training, 500 sentences for validation, and 480 sentences for testing thereby forming a parallel corpora containing all the tribal languages. The corresponding audio recordings were similarly divided based on the text divisions. The duration of audios for each respective language amounted to

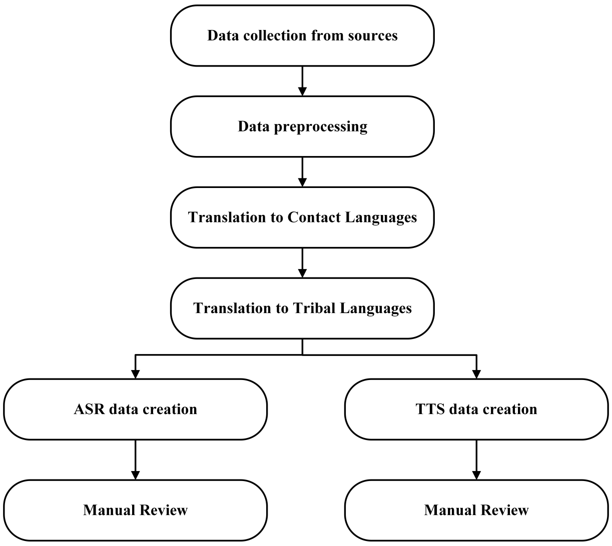

Data collection from sources
Data preprocessing
Translation to Contact Languages
Translation to Tribal Languages
ASR data creation
TTS data creation
Manual Review
Manual Review

Figure 3: Block diagram representing the flow of corpora creation

Sl. No.	English	Lambani	Soliga
1	Buffalo playing in the mud	ಅಂಗೋಳೋ ಆರೋಚ್	ಎಮ್ಮೆ ಬದೀಲಿ ಆಟ ಆಡಿದ್ದೆ
2	The bull hit the cart	ಬಳದ ಆನ ಗಾಡಿನ ಮಾರ್ದಿನೋಚ	ಎತ್ತು ಬಂದು ಗಾಡಿಕ ಗುದ್ದಿಕಿತು
3	A lot of food was wasted	ಘಣೋ ಖಾಣೋ ಹಾಳ ವೇಗೋಚ	ಸೇನೆ ತೀನಿಯ ಹಾಳಮಾಡಿಯವರೆ
4	Dust storm is coming	ಅಂಗೋಳೋ ಆರೋಚ್	ದೂಳಿನ ಬಿರುಗಾಳಿ ಬೀಸಿದ್ದೆ

Table 1: Samples of Lambani and Soliga examples transliterated using Kannada alphabets along with their translated meaning in English.

Sl. No.	English	Kui	Mundari
1	Buffalo playing in the mud	ରଣ୍ଡେ କୋରୁ ଗଦେତୋନି କାହାଇମାନେ	ମିଆଁ କଡ଼ୋ ଲସକ୍ ରେ ଏନକେତାନେ
2	The bull hit the cart	ରଣ୍ଡେ ଷଣ୍ଢକୋଡି ଶଗଡ଼ି ଗାଡ଼ିତିନି ଡୁଞ୍ଜାଅତେ	ମିଆଁ ସଁଣ ଉରିକ୍ଗାଡିକେ ଧକ୍କାଲେଃ
3	A lot of food was wasted	ଆଣ୍ଟ ତିନ୍ବା ନଷ୍ଟି ଆ>ତ	ବାହୁତ୍ ଯମ୍ ଜିନିଷ ଖାରାପ ନାନା
4	Dust storm is coming	ରଣ୍ଡେ ଦୁଲିବାରୁ ବାଇନେ	ଧ୍ୱଡ଼ି ହୟ ହୀଜୁଃତାନା

Table 2: Samples of Kui and Mundari examples transliterated using Odia alphabets along with their translated meaning in English.

approximately 11 hours for training and 30 minutes each for both validation and testing. For ASR experiments, the target text is the transliterated texts of tribal languages whereas for DS2TT experiments, the target text is English. Once the data preparation stage was completed, the audios were transformed to 16-bit 16kHz mono-channel which served as raw inputs. 80-dimensional Mel filterbanks were then extracted from these raw inputs serving as audio features. The transliterations and translations were tokenized using SentencePiece and these tokens served as features for the text. Thus, a set of comprehensive features was extracted from both audio and text.

2.4. Implementation of ASR and DS2TT systems

To implement the baseline ASR and DS2TT systems, the Transformer-based Encoder-Decoder architecture (Vaswani et al., 2017a) is utilized using the Fairseq toolkit (Wang et al., 2020). The Transformer model consisted of two layers of convolutional subsampler. The convolutional layer is responsible for downsampling the input mel filterbank features by a factor of k. In these cases, k is taken as 4. The sampler is followed by a 12-encoder layer. Each encoder layer consisted of eight attention heads, with an embedding dimension of 512, an attention dropout of 0.1, and a dropout of 0.1 elsewhere. Before feeding the target/source text into the decoder they are tokenized using the SentencePiece tokenizer (Kudo and Richardson, 2018). The tokenized texts were then passed through a 6-layer decoder, each containing eight attention heads, an embedding dimension of 512, and a dropout of 0.1. The default model hyper-parameters and learning rate scheduler without any model-specific fine-tuning were used. After training the weights of ten checkpoints were averaged with the lowest validation loss to obtain the final model. The detailed implementation is illustrated in the Figure 4.

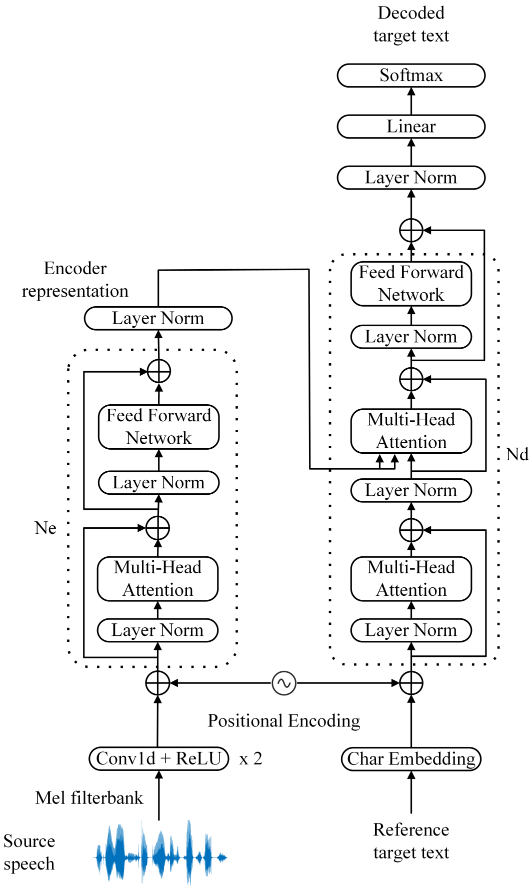


Figure 4: Block diagram of Transformer based ASR/DS2TT system where N_e refers to encoder block and N_d refers to decoder block

2.5. Large Acoustic Model

This study made use of the pre-trained Whisper large-V2 model developed by OpenAI. Whisper is developed using 680,000 hours of multi-lingual supervised data demonstrating multitasking capabilities. Notable, this model demonstrated improved performance over other existing models for the dataset involving English language (Jain et al., 2023). Hence, this motivated to utilize this pre-trained model and assess its capability for these low-resourced tribal languages for both ASR and DS2TT tasks. The architectural framework is similar to that of Figure 4 in which the encoder is initialized using the parameters from this pre-trained model.

2.6. Metrics for Evaluation

For inferencing, a beam size of 5 is used for decoding for all the systems. The ASR systems were evaluated using Word Error Rate (WER). For DS2TT, de-tokenized BLEU scores were computed using SacreBLEU (Post, 2018) for evaluating DS2TT systems. BLEU is the most widely used metric to evaluate translation quality.

3. Results and Discussions

The results of the ASR experiments conducted on the four tribal languages are presented in Table 3. The performances are evaluated using the Word Error Rate (WER) metric. It can be observed that the WERs for the baseline systems on the four tribal languages were above 100%, whereas, with the use of pre-trained Whisper LAM, the performances were improved for these languages. Significant improvement can be seen for Soliga at 22.09%. It is also worth noting that the performance gains were more pronounced for Lambani and Soliga compared to Kui and Mundari. This variance in performance can be attributed to the familiarity of the LAM with the Kannada language, which is a Dravidian language that includes Soliga and Lambani in its family. Hence, the Whisper LAM can capture the linguistic nuances resulting in better recognition of speech by the ASR system. Among Kui and Mundari, Kui had an edge in performance over Mundari as it had Dravidian elements but it has been transcribed using Odia script which is not a part of the Dravidian language family. The performance of Kui using a Dravidian script would be an area of exploration in the future. Furthermore, the better performance for Soliga can be attributed to it being a pure Dravidian language unlike Lambani, which is an Indo-Aryan language influenced by Kannada.

The results of the DS2TT experiments, where the source audios were the tribal languages and the target texts were English, are presented in Table 4. A similar performance trend can be observed in which Soliga had the biggest gain when fine-tuned with the pre-trained Whisper-Large V2 LAM.

The results from both the ASR and DS2TT experiments indicate a notable improvement in the performance attributed to the fine-tuning using a pre-trained Whisper Large V2 Acoustic Model when compared with the results of the baseline systems. The results from the baseline systems fell significantly short, offering less or no practical utility. This justifies the adoption of LAMs to make the ASR and DS2TT systems effective for documenting similar spoken-only languages. Furthermore, the results suggest that borrowing scripts

Sl. No.	Language	WER (baseline)	WER (Fine-tuned)
1	Kui	106.53	64.57
2	Mundari	101.30	70.00
3	Lambani	100.37	49.61
4	Soliga	103.24	22.09

Table 3: Baseline and Whisper fine-tuned evaluations for ASR task on the four tribal languages using Word Error Rate (WER) metric (in percentages).

Sl. No.	Source-Target Language Pair	BLEU (baseline)	BLEU (Fine-tuned)
1	Kui-English	2.44	5.58
2	Mundari-English	1.67	5.09
3	Lambani-English	1.47	7.62
4	Soliga-English	1.44	10.07

Table 4: Baseline and Whisper fine-tuned evaluations using BLEU metric (in percentages) for DS2TT task with the tribal languages as source speech translated to English text.

within the same language family impacts the performance of speech recognition tasks. This is evident in the difference in performances between Kui and Mundari and between Soliga and Lambani for the speech recognition task. Additionally, it is evident that those tribal languages, which originated from the Dravidian language family, performed better as compared to their Odia counterpart. This may be due to the reason that even though this particular LAM was trained using the Odia language, however, it did not support Odia for inferencing (C., 2023). Hence, these insights collectively may help in building corpora for such spoken-only languages and document them through the implementation of more robust Spoken Language Technologies(SLTs) incorporating LAMs.

4. Conclusions and Future Works

The primary objective of this work is to assess a method for documenting four novel under-resourced tribal languages of India that lacked a standardized writing system using Speech Recognition and Translation systems. To enhance the practicality of these systems, this work justified the use of pre-trained Large Acoustic Models. Additionally, this work also sheds light on understanding the performance characteristics exhibited by LAMs in the four novel languages. Notable, it is also observed that better performance is achieved for those under-resourced languages that were transcribed using scripts from the same language family as that of the audio. This study may serve in setting guidelines for creating corpora for similar spoken-only languages and document it with better-spoken language technologies by taking advantage of LAMs. This study was limited in performing evaluations using only Whisper LAM which is trained using a fully-supervised approach. In the future, LAMs that were built using a self-supervised approach will be explored, and their performance for these languages. Also, the performances of the relatively under-performing tribal languages will be evaluated by introducing appropriate script.

5. Acknowledgements

This work is funded by the Ministry of Electronics & Information Technology, Government of India for the ongoing project titled ”Speech-to-Speech Translation for Tribal Languages”. The authors highly acknowledge the experts and associates from different organizations that were involved in the data collection and corpora building process for the four tribal languages. The authors also would like to thank the people from the tribal communities who agreed to donate their voice (with proper documented consent) for the preparation of the corpus.

References

C. Anantaram, Sunil Kumar Kopparapu, Chirag Patel, and Aditya Mittal. 2016. Repairing general-purpose asr output to improve accuracy of spoken sentences in specific domains using artificial development approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4234–4235. AAAI Press.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of* *the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 474–479, Valencia, Spain. Association for Computational Linguistics.

Youssef Bassil and Mohammad Alwani. 2012. Post-editing error correction algorithm for speech recognition using bing spelling suggestion. *arXiv preprint arXiv:1203.5255*.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6224–6228. IEEE Press.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation.

Marek Bohac, Karel Blavka, Michaela Kucharová, and Svatava Skodová. 2012. Post-processing of the recognized speech for web presentation of large audio archive. *2012 35th International Conference on Telecommunications and Signal Processing (TSP)*, pages 441–445.

S. Boopathy. 1972. Languages of Tamil Nadu: Lambadi, an Indo-Aryan dialect. *Census of India 1961, Tamil Nadu*, ix, part XII.

Johanna C. 2023. Whisper API FAQ. `https://help.openai.com/en/articles/7031512-whisper-api-faq/`.

Amartya Chowdhury, Deepak K. T., Samudra Vijaya K, and S. R. Mahadeva Prasanna. 2022. Machine translation for a very low-resource language - layer freezing approach on transfer learning. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 48–55, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Horia Cucu, Andi Buzo, Laurent Besacier, and Corneliu Burileanu. 2013. Statistical error correction methods for domain-specific asr systems. In *Statistical Language and Speech Processing*, pages 83–92, Berlin, Heidelberg. Springer Berlin Heidelberg.

Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 949–959, San Diego, California. Association for Computational Linguistics.

George Abraham Grierson. 1928. *Linguistic Survey of India*, volume 9. Office of the Superintendent of Government printing, India.

Dell Hymes. 1970. Morris swadesh.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of whisper models to child speech recognition. *ArXiv*, abs/2307.13008.

K.Ramalingareddy. 2023. Lambanis: Origin and socio cultural profile. *ArXiv*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Diana M Morlote, Tenzin Gayden, Prathima Arvind, Arvind Babu, and Rene J Herrera. 2011. The soliga, an isolated tribe from southern india: genetic diversity and phylogenetic affinities. *Journal of human genetics*, 56(4):258–269.

Chandrashekar Naik and D. Paramesha Naik. 2012. Banjara stastical report karnatka state, india.

National Council of Educational Research and Training. Ncert.

Toshiki Osada. 2008. Mundari. *The Munda languages*, 99.

Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. 2021. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 367–376.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Maria Shugrina. 2010. Formatting time-aligned ASR transcripts for readability. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 198–206, Los Angeles, California. Association for Computational Linguistics.

TV Spandana, Anusuya Kamath, and BK Veena. 2023. Soliga tribes' culture and identity. *International Research Journal of Modernization in Engineering Technology and Science*, 5(2):916–922.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Waibel and Christian Fugen. 2008. Spoken language translation. *IEEE Signal Processing Magazine*, 25(3):70–79.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Walter Warren Winfield. 1929. A vocabulary of the kui language: Kui-english. Asiatic society of Bengal.

Lawrence Wolf-Sonkin. 2021. Proposal to encode the mundari bani script in the universal character set.