Heterographic Supervision for Direct Speech-to-Text Translation in Scriptless Tribal Languages

Prashant Bannulmath, Swapnil Sontakke, Deepak K T
Indian Institute of Information Technology Dharwad
prashantb@iiitdwd.ac.in, swapnil.sontakke@iiitdwd.ac.in, deepak@iiitdwd.ac.in

Abstract

Direct Speech-to-Text Translation (DS2TT) generates target language text directly from source speech without requiring intermediate source language transcriptions. However, developing DS2TT for low resource, scriptless tribal languages is challenging due to limited parallel data. We investigate heterographic transcriptions as auxiliary supervision in DS2TT for four Indian tribal languages: Lambani, Soliga, Kui, and Mundari. We evaluate three transcription strategies: contact language scripts, contact language influenced IPA, and HuBERT based pseudo text. Using parallel corpora of speech, heterographic transcriptions, and English translations, we train DS2TT models with cross modal learning. Results show that heterographic transcriptions significantly improve translation quality over speech only baselines, while inference remains transcription free, requiring only speech input.

1 Introduction

Speech-to-text translation converts spoken input in one language into written text in another language (Bérard et al., 2016; Duong et al., 2016; Anastasopoulos et al., 2016; Weiss et al., 2017; Bérard et al., 2018). Traditional speech translation systems employ a cascaded architecture consisting of Automatic Speech Recognition (ASR) followed by Machine Translation (MT) (Waibel and Fugen, 2008). Although effective, cascaded systems suffer from error propagation and increased inference latency (Jia et al., 2019; Yang et al., 2022; Indurthi et al., 2019; Bérard et al., 2016; Duong et al., 2016). Direct Speech-to-Text Translation (DS2TT) addresses these limitations by translating speech directly into target language text without intermediate transcriptions. However, training DS2TT models remains challenging for low-resource languages due to the scarcity of parallel speech translation data (Gaido et al., 2021; Zhang and Feng, 2023).

Most DS2TT datasets assume the availability of standardized writing systems and contain speech, source transcriptions, and target translations (Di Gangi et al., 2019; Salesky et al., 2021; Iranzo-Sánchez et al., 2020). This assumption does not hold for many Indian tribal languages such as *Lambani*, *Soliga*, *Kui*, and *Mundari*, which are predominantly oral and lack standardized orthographies (Besacier et al., 2006; Martin et al., 2015; Adams et al., 2016a,b; Anastasopoulos and Chiang, 2017). Previous studies have improved low-resource DS2TT using pretraining, curriculum learning, mutual learning, and contrastive learning (Stoian et al., 2020; Masumura et al., 2020; Cheng et al., 2021; Ye et al., 2022). Recently, discretized speech representations, often referred to as speech units, have gained attention for their role in processing spoken conversations (Lakhotia et al., 2021). These units are typically generated by clustering representations extracted from self-supervised speech models, capturing acoustic features that are comparable to traditional linguistic units like phonemes. A key innovation lies in the treatment of these speech units as a form of pseudo-text (Kim et al., 2024), opening new avenues for conducting textless natural language processing without textual transcriptions, particularly beneficial for low resource and unwritten languages. Inspired by these developments, this study explores leveraging cross modal heterographic transcriptions and self-supervised speech units for training DS2TT models in low resource, scriptless language scenarios.

The main contributions of this work are:

1. We develop a multilingual parallel corpus comprising speech, heterographic transcriptions, and English translations for four low resource Indian tribal languages.

2. We compare three heterographic transcription strategies: contact language scripts, contact

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 59–64
October 17-21, 2026 ©2026 Association for Computational Linguistics

language influenced IPA, and HuBERT based pseudo text.

3. We demonstrate that heterographic transcriptions substantially improve DS2TT performance in low resource settings while preserving end-to-end, transcription free inference.

2 Corpus Creation

Lambani is an Indo-Aryan language influenced by Kannada (Madar, 2016), Soliga is a Dravidian language spoken in Karnataka (Morlote et al., 2011; Spandana et al., 2023), Kui is a South-Eastern Dravidian language spoken in Odisha (Friend-Pereira, 1909), and Mundari is an Austroasiatic language spoken in Odisha and neighboring regions (Anderson, 2015). None of these languages has a standardized writing system, motivating the use of heterographic transcriptions.

English sentences were collected from textbooks (using OCR), news articles, and the Swadesh list (Alexei et al., 2010), followed by manual cleaning and linguistic verification. The curated corpus was translated into Kannada and Odia by bilingual translators and subsequently translated into Lambani, Soliga, Kui, and Mundari by native speakers. All translations were manually validated. Speech recordings were collected from native speakers. The final corpus consists of speech, heterographic transcriptions, and English translations for all four languages. Corpus statistics are summarized in Table 1.

Table 1: Corpus statistics across four tribal languages

Language	Utterances	Duration (hrs)
Lambani	29,661	16.15
Soliga	9,977	7.80
Kui	29,336	36.56
Mundari	19,923	33.91

3 Methodology

Figure 1 shows the proposed DS2TT framework. The model consists of a pre-trained Speech encoder, an Embedding layer, and a Transformer encoder-decoder. Raw speech is first processed by a pre-trained speech encoder to extract contextualized speech representations. A 1-D convolution layer with stride 2 further downsamples the speech features before they are passed to the Transformer

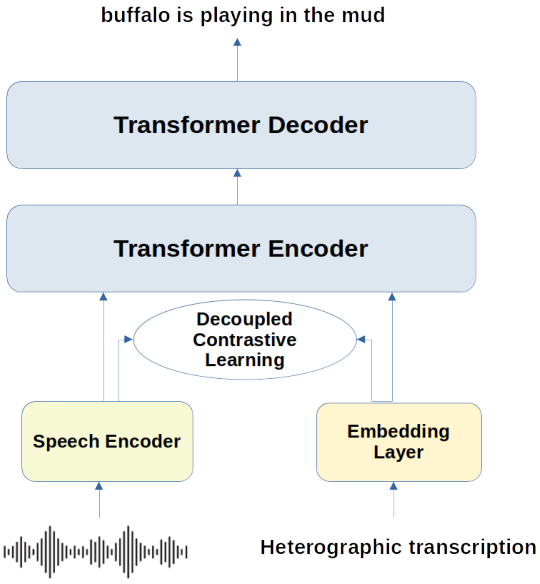


Figure 1: Architecture of the decoupled contrastive learning based DS2TT system leveraging heterographic transcription.

encoder. The Transformer decoder then autoregressively generates the target language text.

To reduce the modality gap between speech and heterographic transcription, we employ Decoupled Contrastive Learning (DCL). Sentence level speech and heterographic transcription representations are obtained using mean pooling:

$$a = \text{MeanPool}(s), \tag{1a}$$
$$b = \text{MeanPool}(x), \tag{1b}$$

where s and x denote the speech and corresponding heterographic transcription, respectively. The DCL objective (eq.2) encourages matching speech-heterographic transcription pairs to have similar representations while pushing apart non-matching pairs within a mini-batch:

$$L_{DC} = -\frac{\text{sim}(a,b)}{\tau} + \log \sum_{j \neq i} \exp\left(\frac{\text{sim}(a_i, b_j)}{\tau}\right), \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature parameter.

Unlike conventional DS2TT systems that rely on native orthographic transcripts, we investigate

three heterographic transcription strategies for contrastive alignment:

- **Contact-language scripts:** Tribal speech is transcribed using the script of the dominant contact language (Kannada for Lambani and Soliga, and Odia for Kui and Mundari).
- **Contact language-influenced IPA:** The source speech is represented using International Phonetic Alphabet (IPA) transcriptions that capture pronunciations influenced by the contact language phonology.
- **HuBERT-based pseudo-text:** In the absence of human transcriptions, discrete speech units extracted using HuBERT (Hsu et al., 2021) and clustered into 100 units using K-means are treated as pseudo-text for contrastive alignment, enabling learning in transcription free settings.

During training, the contrastive objective aligns speech with these heterographic transcriptions, while the translation objective optimizes the Transformer encoder-decoder. During inference, the model requires only speech input to generate the target language text.

4 Experiments & Results

4.1 Training Configuration

For each of the four language pairs, approximately 90% of the data was used for training, with the remaining 10% equally divided between the development and test sets. The speech encoder used in our experiments is wav2vec 2.0 pre-trained on Librispeech to generate speech representations, which are passed through a CNN layer with hidden size 512, kernel size 5, and stride 2, followed by a base Transformer consisting of six encoder layers and decoder layers with eight attention heads and 2048-dimensional feed-forward hidden states. The input speech data is mono-channel sampled at 16 kHz. Optimization is performed using the Adam optimizer with a learning rate of 1×10^{-4} and a warmup of $25k$ steps, while the decoupled contrastive loss uses a temperature parameter $t = 0.02$ and contrastive weight $\lambda = 1.5$. Training is carried out on two NVIDIA RTX A5000 GPUs within the fairseq environment.

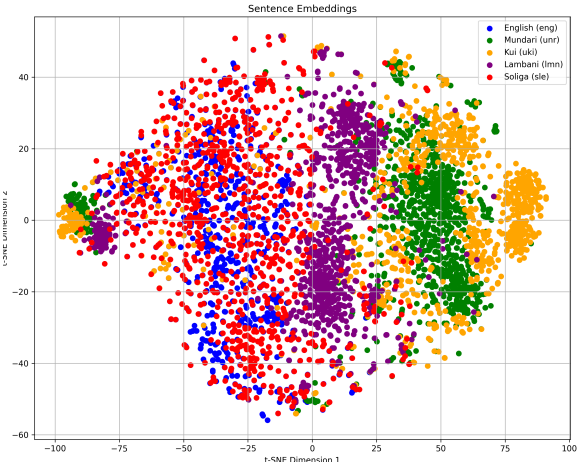


Figure 2: t-SNE visualization of the sentence embeddings for English and the four tribal languages.

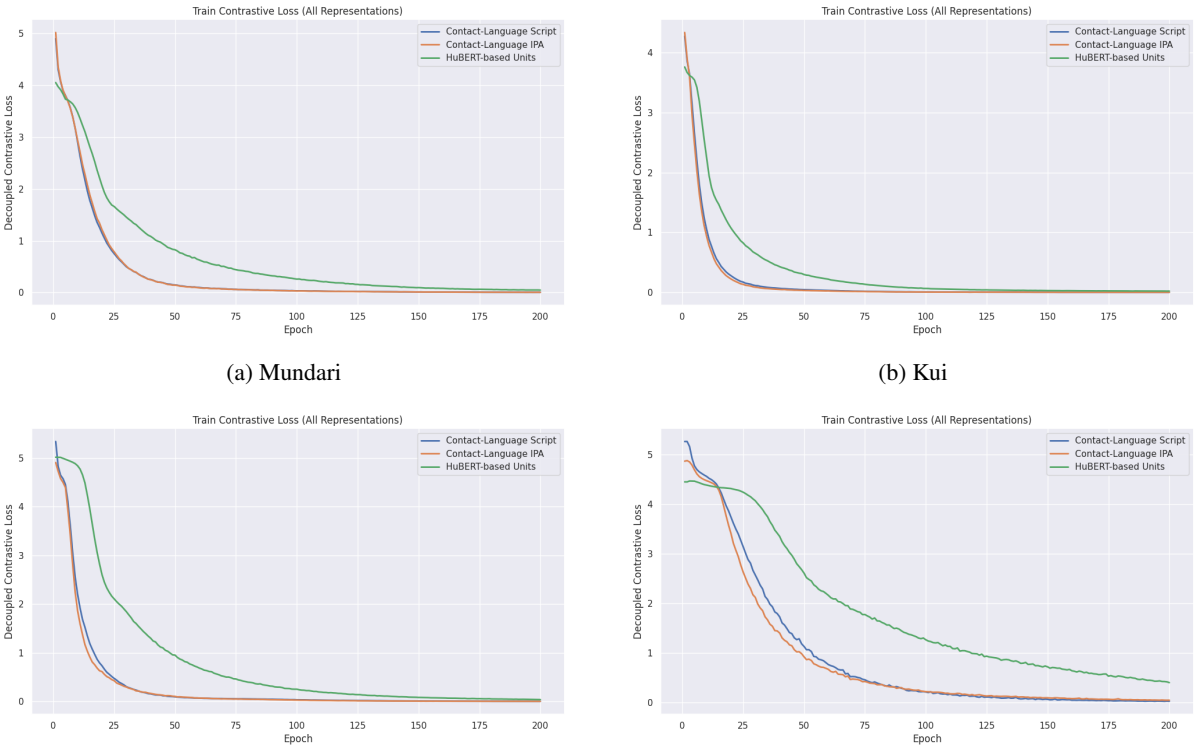
4.2 Results & Discussion

Table 2 reports the test SacreBLEU scores of the proposed DS2TT system using three auxiliary supervision strategies across four low resource tribal languages. For comparison, the table also reports the SacreBLEU scores obtained using a machine translation model based on IndicTrans2 (Gala et al., 2023), which serves as a text-to-text translation baseline for comparison with the proposed speech-to-text translation approaches. The no-transcription baseline performs poorly for all languages, with SacreBLEU scores below 0.5, highlighting the difficulty of learning speech-to-text translation without auxiliary supervision. Introducing heterographic transcriptions substantially improves translation performance. Contact language script achieves the best performance for Mundari and Kui , while contact language influenced IPA yields the highest SacreBLEU score for Soliga. These results suggest that both orthographic and phonetic representations provide effective supervision for learning robust alignments. HuBERT based pseudo text outperforms the no-transcription baseline and achieves the best performance for Lambani, demonstrating that discrete speech units are a possible alternative when manual transcriptions are unavailable.

Figure 3 shows the decoupled contrastive loss curves for the four languages. Across all settings, the contrastive loss decreases steadily and converges, indicating stable optimization and effective alignment between speech and heterographic transcriptions. Models trained with contact language script and IPA exhibit faster convergence than HuBERT based pseudo text, reflecting the

Table 2: Test SacreBLEU scores of the proposed DS2TT system using different heterographic transcription strategies across four low-resource tribal languages.

Language	No Auxiliary Supervision	Contact Language Script	Contact Language IPA	HuBERT Pseudo Text	Machine Translation
Mundari	0.24	**18.64**	18.39	8.41	17.50
Kui	0.39	**32.84**	31.62	17.54	30.70
Lambani	0.24	2.69	2.48	3.34	**26.80**
Soliga	0.24	18.36	**20.19**	10.31	20.10



(a) Mundari

(b) Kui

(c) Lambani

(d) Soliga

Figure 3: Decoupled contrastive loss curves for the four tribal languages.

benefit of linguistically informed supervision. Figure 2 visualizes the sentence representations using t-SNE. The embeddings form language specific clusters while maintaining partial overlap with English representations. Overall, the experimental results demonstrate that heterographic supervision effectively improves DS2TT performance for scriptless, low-resource languages.

5 Conclusion

This study investigates the effectiveness of heterographic transcriptions in improving DS2TT system for four low-resource, scriptless Indian tribal languages (Lambani, Soliga, Kui, and Mundari). Given the absence of standardized orthographies for these languages, introducing any form of transcription whether linguistic or data-driven significantly improves the average SacreBLEU score across the four languages by approximately 10–18 points compared with training without auxiliary supervision. Moreover, the proposed DS2TT framework achieves translation performance comparable to text-to-text translation baseline for most of the evaluated languages while eliminating the need for source-language transcriptions during inference. These findings underscore the potential of heterographic transcriptions as both a practical documentation strategy and an effective computational bridge for modeling languages without standardized orthographies. Future work will aim to improve generalization to unknown speakers by expanding the dataset with more speaker diversity and by applying advanced cross lingual adaptation techniques.

Acknowledgments

The authors would like to thank the Science and Engineering Research Board (SERB), Government of India, for supporting this research.

References

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. 2016a. Learning a translation model from word lattices. In *INTERSPEECH*, pages 2518–2522.

Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Truong Do, and Satoshi Nakamura. 2016b. Learning a lexicon and translation model from phoneme lattices. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2377–2382.

Kassian Alexei, Starostin George, Dybo Anna, and Chernov Vasiliy. 2010. The swadesh wordlist. an attempt at semantic specification. *Voprosy yazykovogo rodstva*, (16 (59)):46–89.

Antonios Anastasopoulos and David Chiang. 2017. A case study on using speech-to-translation alignments for language documentation. *arXiv preprint arXiv:1702.04372*.

Antonios Anastasopoulos, David Chiang, and Long Duong. 2016. An unsupervised probability model for speech-to-translation alignment of low-resource languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1255–1263, Austin, Texas. Association for Computational Linguistics.

Gregory DS Anderson. 2015. *The Munda Languages*. Routledge.

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)*, pages 6224–6228. IEEE.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Laurent Besacier, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *2006 IEEE Spoken Language Technology Workshop*, pages 222–225. IEEE.

Yao-Fei Cheng, Hung-Shin Lee, and Hsin-Min Wang. 2021. Allost: low-resource speech translation without source transcription. *arXiv preprint arXiv:2105.00171*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics.

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.

JE Friend-Pereira. 1909. A grammar of the kui language. *Bengal Secretariat Book Depot, Calcutta*.

Marco Gaido, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 55–62, Trento, Italy. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, Raghavan Ak, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2019. Data efficient direct speech-to-text translation with modality agnostic meta-learning. *arXiv preprint arXiv:1911.04283*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184.

Minsu Kim, Jeongsoo Choi, Dahun Kim, and Yong Man Ro. 2024. Textless unit-to-unit training for many-to-many multilingual speech-to-speech translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Shiddappa Madar. 2016. Out-migration of lambani community: an observation. *Editorial Board*, 5(4):175.

Lara J Martin, Andrew Wilkinson, Sai Sumanth Miryala, Vivian Robison, and Alan W Black. 2015. Utterance classification in speech-to-speech translation for zero-resource languages in the hospital administration domain. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 303–309. IEEE.

Ryo Masumura, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, and Takanori Ashihara. 2020. End-to-end automatic speech recognition with deep mutual learning. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 632–637. IEEE.

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.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation. *arXiv preprint arXiv:2102.01757*.

TV Spandana, Anusuya Kamath, and BK Veena. 2023. Soliga tribes'culture and identity. *International Research Journal of Modernization in Engineering Technology and Science*.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

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

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur. 2022. JHU IWSLT 2022 dialect speech translation system description. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 319–326, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. *arXiv preprint arXiv:2205.02444*.

Shaolei Zhang and Yang Feng. 2023. End-to-end simultaneous speech translation with differentiable segmentation. *Preprint*, arXiv:2305.16093.