Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Improving Conformer based End-To-End Manipuri Automatic Speech Recognition using Wav2vec2 model

Thangjam Clarinda Devi[1*], Kabita Thaoroijam[2], Kishorjit Nongmeikapam[1]
[1]*Department of Computer Science and Engineering, Indian Institute of Information Technology Manipur, India*
[2]*School of Computer Science and Artificial Intelligence, SR University Warangal, India*
Corresponding author: thangc@iiitmanipur.ac.in

ARTICLE INFO

Received: 26 Dec 2024
Revised: 14 Feb 2025
Accepted: 22 Feb 2025

ABSTRACT

In order for end-to-end speech recognition systems to function successfully, a lot of labeled speech data is required for training. Due to the availability of huge labeled voice corpora for high-resource languages like English, this condition tips the scales in favor of those languages. On the other hand, transcriptions of speech for most languages spoken around the world are scarce. This work builds a Conformer based end-to-end automatic speech recognition system for Manipuri, one of the Indian and low resource languages. ULCA (Unified Language Contribution APIs) Manipuri speech corpus of around 10 hours is used. The proposed method uses two approaches for extracting features- Log Mel and wav2vec2 speech features. Log Mel features are extracted from the input speech signals and speed perturbation technique is used to effectively increase the amount of training data and wav2vec2 model are used as a pre-encoder for speech features. Word Error Rate (WER) and Character Error Rate (CER) are used to gauge how well the trained conformer-based model performs. The best performance achieved by the proposed system is 29.7% WER and 8.3% CER. The results are compared with the baseline LSTM based ASR system and it was found that the proposed system gave an absolute improvement of 34% in WER and 23.5% in CER.

Keywords: *Speech recognition, wav2vec2, conformer, machine learning, manipuri speech, manipuri text*

INTRODUCTION

The use of hybrid modeling based on deep neural networks (DNN) configured approximately ten years ago has led to significant improvements in the accuracy of automated speech recognition (ASR) systems [1]. By using DNNs for the acoustic probability evaluation instead of the traditional Gaussian mixture model, this invention preserved the hybrid ASR system's design, which included lexicon, language, and acoustic models. A significant milestone in the speech community's shift from hybrid modeling to end-to-end (E2E) modeling was recently reached [2–6]. A single network converts the audio sequence into the token sequence output in E2E modeling. Even more amazing is the fact that E2E modeling does away with every element of conventional ASR system modeling, which has been around for many years. E2E models have many more advantages than their hybrid equivalents. The optimization process under a single objective function that coincides with the goal of the ASR system is the main emphasis of E2E models. Traditional hybrid models, on the other hand, optimize each component separately, making it impossible for them to reach a global solution. As a result, E2E models have been shown to outperform traditional hybrid models in both scholarly research [7] and practical applications [8]. Second, by directly generating letters or even phrases, E2E models greatly simplify the ASR pipeline. On the other hand, the intricate design of conventional hybrid models necessitates years of ASR expertise and a great deal of specialized knowledge. Third, compared to conventional hybrid models, E2E models are considerably more compact because the entire model is based on a single network. Such characteristics make E2E models easier to deploy to end devices unlike hybrid models which require multiple integrations. By performing a benchmark evaluation on the Universal Language Contribution APIs (ULCA) Manipuri speech dataset hosted on their website, this work aims to improve ASR for a low resource language. The following contributions are made by this work:

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[1] The proposed work combines the conformer based and wav2vec2 model for Manipuri ASR. This approach has not been investigated in the earlier Manipuri ASR.

[2] Dataset was manually annotated with speaker information in order to make sure no same speaker utterances overlap in train set and test set.

[3] To the best of our knowledge, no prior ASR work on this dataset exists yet, thus the proposed Manipuri ASR model is developed as the benchmark evaluation.

The structure of this paper is as follows.The relevant work is reviewed in Section 2. Section 3 discusses methodology and how it is applied. Section 4 presents the findings and discussion. The outcomes are the main topic of discussion in this section. Section 5 presents our work's conclusions.

RELATED WORK

The authors in reference [9] have described a variety of methods for developing ASR systems and the requirements for ASR-based technologies. They performed speech recognition based on E2E for two different languages. According to their findings, E2E ASR could enhance speech recognition across a range of speech patterns, including accents and languages [10]. The E2E-based speech recognition system performs better than the current techniques in a variety of difficult situations, including noisy and clear speech [3]. Nonetheless, the DNN model can greatly enhance the creation of a reliable ASR system. Using E2E speech recognition models with various data bases, a significant decrease in word error rate (WER) was seen [11]. An attention-based model outperforms another E2E strategy in terms of performance. On noisy data, though, it produces subpar results [12]. An E2E speech recognition system can be trained by using the CTC approach to label unsegmented sequences [13]. According to [14], self-attention and multihead attention layers can be used for transformer model encoding and decoding. The core of an end-to-end ASR system, which streams ASR and requires that an output be produced as soon as a word is spoken, is formed by Transformers. Time-restricted self-attention is applied to the encoder and to the encoder-decoder attention mechanism prompt attention is applied. The new fusion attention technique on the Wall Street Journal task yields a WER decrease of 16.7% against the non-fusion standard transformer and 12.1% against the transformer-based benchmarks of other authors. Advancements in ASR systems and the development of large speech corpora have resulted from speech recognition research in the English language. CommonVoice (1400 hours), LibriSpeech (960 hours), TedLium-3 (450 hours), Switchboard (300 hours), and SPGISpeech (5000 hours) are a few of the English language common speech corpora [15]. All specified English corpora, with the exception of Switchboard, are freely downloadable for academic and non-commercial use. As for the WER (Word Error Rate), English speech recognition accuracy has greatly improved in recent years. Having decreased from 5.33% [16] to 1.4% [17], the WER on the LibriSpeech test-clean [18] dataset has improved. Low-resource Indian languages do not, however, possess a comparable sizable dataset for Automatic Speech Recognition (ASR) investigation. Additionally, there have been several attempts to use augmentation to create large speech corpuses for languages with inadequate resources [19]. Very few works in Manipuri ASR have been done. In [20], the telephonic read speech data is collected and used for GMM-HMM and DNN-HMM based ASR with 13.57% WER. Also, similar work on GMM-HMM acoustic model has been done by [21]. Time Delay Neural Network (TDNN) model outperformed with 2.53% WER. Recently, the authors in [22] implemented an ASR Model based on E2E approach with both Connectionist Temporal Classification (CTC) and Attention type model. While E2E systems are simple to understand and train, they still require a large amount of an aligned and annotated speech-text corpus. Additionally, alignment of labels to the speech data is still a problem for E2E systems. Although E2E model has improved ASR, most of the work is done on well-studied languages whereas work on low resource languages like Manipuri receives little attention. Manipuri (also known as Meeteilon) is the Official Language of the northeastern state of India, Manipur. ASR tasks in Manipuri are significantly under-resourced and lacking in data, which is one of the primary reasons why the development of ASR systems is stalled for this language.

METHODS

ULCA Manipuri dataset [23] is used for the experiment. The corpus was collected from All India Radio Manipur led by EkStep Foundation under National Language Translation Mission (NLTM). The dataset consists of recordings of sentences from News domain recorded by male and female News readers. Dataset is about 10 hours of speech

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

utterances. It has 5449 number of utterances in total. While splitting the data into Train, Valid and Test set, it was ensured that no speakers utterances are overlapped in training and testing set. Table 1. shows the dataset details.

Table 1: Details of ULCA dataset used.

Dataset	Train		Valid		Test		Total	
	Utterance	Hours	Utterance	Hours	Utterance	Hours	Utterance	Hours
ULCA Manipuri data	3649	-	855	-	945	-	5449	10

E2E systems are smaller and easier to install on servers than traditional systems because they only require one network to map speech frames from input to characters. With all of the benefits mentioned above, E2E systems are now the new SOTA for all ASR assignments. However, since ASR accuracy is not the only metric for these services, most commercially deployed systems continue to use traditional models. Several functional aspects, including streaming, latency, and domain-specific adaptation, are also crucial and can be supplied by well-designed module blocks that make up the conventional ASR pipeline. These conventional models are still in use in many commercial devices because they are production-optimized. With the development of large datasets for speech recognition, research started on building deep learning frameworks for Automatic Speech Recognition. In an effort to map the input audio sequence to the corresponding output text sequence, these deep learning models are E2E, meaning they simultaneously learn the alignment and pronunciation modules.

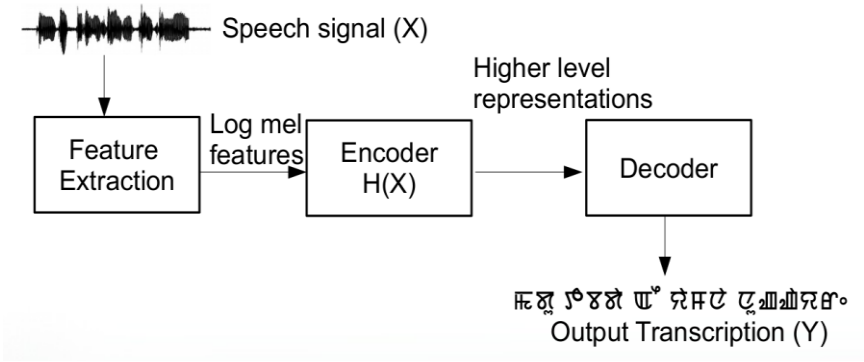

Figure 1. Structure of an End-to-End speech recognition system

Figure 1. depicts the structure of an End-to-End speech recognition system. The mapping of the input audio sequence to a higher dimensional feature sequence is done by the encoder, frequently a fixed context vector. In order to produce the character sequence, the aligner and the decoder work together to learn the phoneme alignments and pronunciations. Despite the figure's seemingly pipeline-like appearance, the training and decoding process is completed in an end-to-end fashion. The model outputs the hypothesis sentence after receiving raw audio files as input. The E2E framework also resolves the optimization issue by attempting to reduce the loss function in order to achieve global optimization of the entire pipeline.

The encoder that converts speech sequences into higher level feature representations is the most important component of any E2E ASR model. In earlier developments of E2E ASR, the most common unit was a Long Short Term Memory (LSTM) network [24]. The encoder could be constructed as either multilayer uni-directional LSTM-RNN or a multilayer bidirectional LSTM (BLSTM) RNN. While LSTMs can model some short-term dependencies, models based on Transformers [25,26] are much more efficient at capturing long-range interdependencies within audio sequences due to the self-attention mechanism incorporated into the transformer architecture. Large-scale datasets comprising audio recordings and transcriptions, like LibriSpeech or Common Voice, are frequently used to pretrain transformer-based models. The pretraining stage enables the models to pick up linguistic and acoustic representations, thereby capturing complex associations between textual data and auditory signals. In this work,

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Conformer [27] encoders were used as it combines Transformer and CNNs efficiently and extracts local patterns for enhancement.

The mathematical expression for Conformer encoders can be expressed as follows:

$$\text{Conformer}(x) = x + 1/2.\text{FFN}_1(x) + \text{MHA}(x) + \text{Conv}(x) + 1/2.\text{FFN}_2(x) \quad (1)$$

where, x is the input signal, FFN stands for Feed-forward network, that provides non-linear transformation to the inputs, MHA stands for Multi-head attention layer for global sequence modeling, Conv module does the local pattern extraction.

In addition to training from scratch, this work attempts to utilize well trained wav2vec2 [28] speech model. A popular SSL model for speech recognition is Wav2vec2, which uses contrastive learning to pretrain on a large number of unlabeled speech samples. By learning to predict future speech segments from past ones, the model produces high-quality representations that can be enhanced on tiny labeled datasets for ASR tasks. This model has shown significant improvements in ASR performance for low-resource languages, even with a limited amount of labeled data. The wav2vec2 model is positioned in front of the encoder so that it can accept its output as acoustic characteristics extracted from the input audio file. In this study, the wav2vec2 model are used as feature extractors for the encoder design.

Word Error Rate (WER): It is the most often used metric for evaluating the efficacy of ASR systems, especially in general-purpose applications. In comparison to the ground truth, it determines how many words were misclassified.

$$\text{WER} = (S + D + I) / N \quad (3)$$

where, S is the number of substitutions (wrong words), D is the number of deletions (missing words), I is the number of insertions (additional words), N is the total number of words in the reference transcript (ground truth). Better ASR performance is indicated by a lower WER.

Standard ESPnet2 toolkit [29] is used to implement E2E ASR model. ASR implementation steps are as follows. The first step is the data preparation step, where the train, valid and test set are obtained from splitting the entire downloaded raw dataset, and prepared them in the Kaldi format. Monitoring of the training progress is done by checking the validation score. Also, test and valid sets are used for the final speech recognition evaluation. Thus, the speech, corresponding text and speaker information are prepared in the Kaldi format. Secondly, speed perturbation [30] with standard factors of 0.9, 1.0 and 1.1 is used for data augmentation. Speed perturbation is performed and then saved the augmented data in the disk before training. After this step, there was an increase in the number of utterances from 3649 to 10947 of train set data. Another approach is to perform data augmentation during training, such as SpecAug [31]. Thirdly, processing of the data in wav.scp file with specified format such as flac is done for the efficient use of the data. Then, too long and too short speech utterances are removed from the train set as they are harmful for efficient training. But for testing and scoring, we still use the full data, which is important for fair comparison. Next the token list is generated from the train set. This is important for text processing. Here, a dictionary is prepared simply using the Manipuri characters. The sentencepiece toolkit developed by Google has been used. Then, the normalization of the data is done by estimating the mean and variance of the data. Also, collected the information of input and output lengths for the efficient mini batch creation. The ASR model is trained and the training log contains all information about the current experiment i.e., it tells about the training status. So, in case if the training job fails, this file will show the error messages. Finally, the testing or decoding step is done once the ASR model is trained and the WER and CER are calculated to check the performance of the model.

RESULTS AND DISCUSSION

The standard architecture is implemented in order to train the models. 80 dimensional log Mel filter banks are used as input features in the LSTM (base-line) and Conformer with log Mel features based models. The conformer using wav2vec2 model feature-based models uses wav2vec2 model features. The usage of 500 and 1,000 Byte-Pair Encoding (BPE) tokens as the output units has been studied. During training, the model's weights are modified using the Adam optimizer. The optimization method used to modify the model's weights in order to minimize the loss function has a step size that is determined by the learning rate. For the Conformer with log mel features model and the Conformer with wav2vec2 features model, the Adam optimizer's learning rate coefficient is set at 0.0005 and 0.0025, respectively. The warmup learning rate approach was used. Because of the task's complexity and the limited

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

amount of the labeled dataset, 20 epochs were used during training on a single GPU. The number of training instances that are simultaneously input into the model during training is determined by the batch size. Batch size of 20 with gradient accumulation every 3 steps were used in the experiment. CTC weight 0.3 is selected for decoding experiment. Subsequently, ten checkpoints with the highest valid accuracy are averaged during the decoding process. The loss with respect to the number of epochs plot is shown in Figure 3. This gives us better insight into how the training performance changes over the number of epochs. The plot shows the decreasing slope for all the models after each epoch. The conformer with wav2vec2 (unfreeze encoders) achieves the lowest loss value compared to other models.

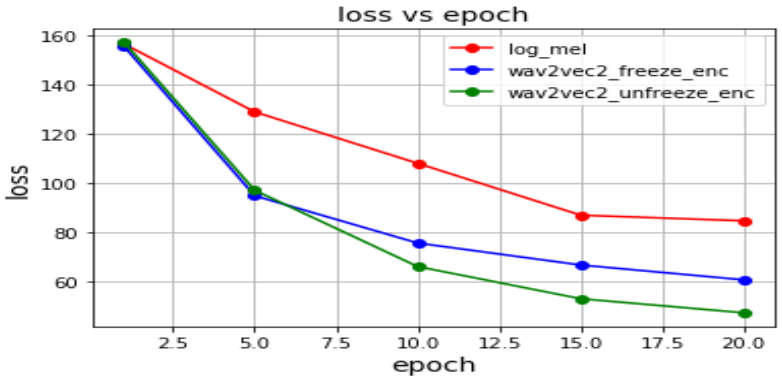

Figure 3. Loss vs number of epochs plot for Conformer Based Sytems during training.

The comparative ASR results of the explored ASR set-up on the valid data and test data is summarized in Table 2. By examining the table, it can be noted that Conformer with wav2vec2 unfreeze model results are improved compared with the baseline model of 63.7% WER and 31.8% CER on test data. There is an improvement of 34.0% in WER and 23.5% in CER. Similarly, in comparison with Transformer based, the conformer based model showed better results with an improvement of 22.2% in WER and 17.0% in CER. Moreover, freezing the encoder layers of the Conformer with the wav2vec2 model during training did improve the results, but unfreezing the encoder layers gave superior results of 29.7% WER and CER of 8.3% on test data with an improvement of 7.7% in WER and 4.3% in CER.

Table 2: Comparison of ASR Results on the validation data and test data in terms of WER and CER (in percentage)

System Type	**Valid**		**Test**	
	WER (↓)	**CER(↓)**	**WER (↓)**	**CER(↓)**
LSTM Based with log mel(Baseline)	60.9	29.3	63.7	31.8
Transformer Based with log mel	50.3	25.6	51.9	25.3
Conformer Based with log mel	41.0	14.1	45.3	15.0
Conformer Based with wav2vec2 model and freeze encoders	35.7	12.0	37.4	12.6
Conformer Based with wav2vec2 model w/o freezing	**28.6**	**8.5**	**29.7**	**8.3**

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

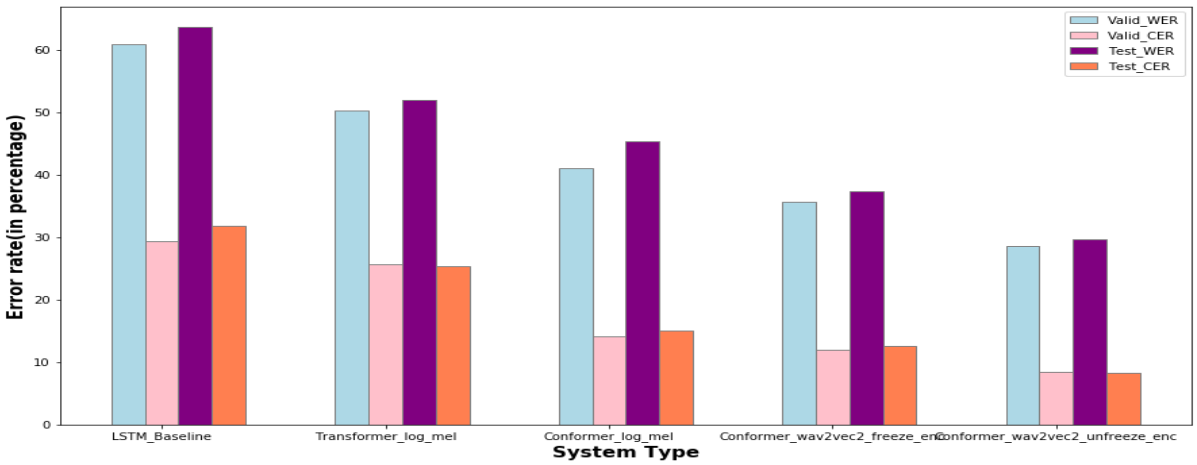

Figure 4. Comparison of model performance.

From Figure 4., it is evident that the choice of freezing or unfreezing the encoders significantly impacts the ASR results. When the encoder part is freezed, the training is faster but the unfreeze gave better results likely because of the nonsimilarity of English and Manipuri language. This shows that the well-trained models like wav2vec2 which was trained in high-resource language can be utilized in the ASR model training for low-resource language even when the languages are not closely related to each other. Table 3. gives the analysis of amount of each type of errors such as substitution error, deletion error and insertion error across different system types. The summation of these type of errors will actually give us the error rates. Figure 5. shows the error distribution of three possible types of errors. The highest error values were for the substitution error type compared to other error types, probably because of some mismatch due to noise in the speech corpus. There is only slight increase in insertion errors compared to deletion errors except for LSTM based model with higher deletion error.

Table 3: Error Distribution Details

System Type	**Substitution**	**Deletion**	**Insertion**
LSTM_log_mel(Baseline)	37.7	10.2	4.0
Transformer_log_mel	37.0	3.4	4.9
Conformer_log_mel	30.6	2.9	3.9
Conformer_wav2vec2 freeze	26.9	2.9	3.7
Conformer_wav2vec2 unfreeze	24.1	2.1	3.5

Journal of Information Systems Engineering and Management
2025, 10(56s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

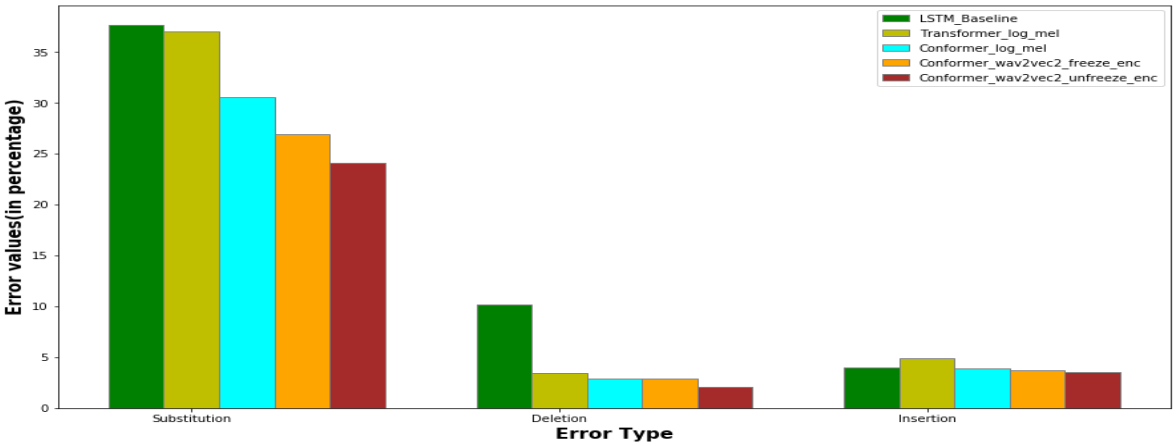

Figure 5. Error distribution of Substitution, Deletion and Insertion type of errors in each models.

CONCLUSION

This work presents Conformer based E2E ASR in Manipuri utilizing the well-trained wav2vec2 models and compares the result with the baseline LSTM based E2E ASR model using ESPnet2 framework. The experiments performed on publicly available ULCA Manipuri dataset have shown that the proposed model gave superior results of 29.7% WER and CER of 8.3% on test data compared to both LSTM and Conformer with log mel features. This work can be the first benchmark for this particular dataset and speech researchers may explore ways to improve the results. Future research directions would include applying unsupervised approach on E2E ASR and its evaluation. Unsupervised ASR can learn from unlabeled audio without needing the labeled transcriptions for training. This might be useful for low-resourced languages. Additionally, other SSL models may also be explored with large-sized dataset for validation on noisy and robust ASR systems. While this work provides valuable insights, there are limitations to this study. The training of the model was done on the speech utterances from news domain. Its performance in out-of-domain speech remains invalidated.

REFRENCES

[1] Hinton G, Deng L, Yu D, Dahl G E, Mohamed A R, Jaitly N et.al 2012 Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*. 29(6):82-97.

[2] Graves A and Jaitly N 2014 Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning* (pp. 1764-1772). PMLR.

[3] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et.al 2014 Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[4] Bahdanau D, Chorowski J, Serdyuk D, Brakel P and Bengio Y 2016 End-to-end attention-based large vocabulary speech recognition. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4945-4949). IEEE.

[5] Chan W, Jaitly N, Le Q and Vinyals O 2016 Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4960-4964). IEEE.

[6] Prabhavalkar R, Rao K, Sainath T N, Li B, Johnson L and Jaitly N 2017 A Comparison of sequence-to-sequence models for speech recognition. In *Interspeech* (pp. 939-943).

[7] Watanabe S, Hori T, Kim S, Hershey J R and Hayashi T 2017 Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*. 11(8):1240-1253.

[8] Sainath T N, He Y, Li B, Narayanan A, Pang R, Bruguier A et.al 2020 A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6059-6063). IEEE.

[9] Bachate R P and Sharma A 2019 Automatic speech recognition systems for regional languages in India. *International Journal of Recent Technology and Engineering*. 8(2):585-592.

[10] Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C et.al 2016 Deep speech 2: End-to-end

speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR.

[11] Li J, Lavrukhin V, Ginsburg B, Leary R, Kuchaiev O, Cohen J M et.al 2019 Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.

[12] Kim S, Hori T and Watanabe S 2017 Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4835-4839). IEEE.

[13] Zhang Y, Pezeshki M, Brakel P, Zhang S, Bengio C L Y and Courville A 2017 Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv preprint arXiv:1701.02720*.

[14] Lee S H, Lee S and Song B C 2021 Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*.

[15] O'Neill P K, Lavrukhin V, Majumdar S, Noroozi V, Zhang Y, Kuchaiev O et.al 2021 Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*.

[16] Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C et.al 2016 Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR.

[17] Zhang Y, Qin J, Park D S, Han W, Chiu C C, Pang R, et.al 2020 Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.

[18] Panayotov V, Chen G, Povey D and Khudanpur S 2015 Librispeech: an asr corpus based on public domain audio books. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5206-5210). IEEE.

[19] Liu C, Zhang Q, Zhang X, Singh K, Saraf Y and Zweig G 2019 Multilingual graphemic hybrid ASR with massive data augmentation. *arXiv preprint arXiv:1909.06522*.

[20] Patel T, Krishna D N, Fathima N, Shah N, Mahima C, Kumar D et.al 2018 Development of Large Vocabulary Speech Recognition System with Keyword Search for Manipuri. In *Interspeech* (pp. 1031-1035).

[21] Meetei L S, Rahul L, Singh A, Singh S M, Singh T D and Bandyopadhyay S 2021 An Experiment on Speech-to-Text Translation Systems for Manipuri to English on Low Resource Setting. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)* (pp. 54-63).

[22] Singh N K, Chanu Y J and Pangsatabam H 2023 Mecos: A Bilingual Manipuri-English Spontaneous Code-Switching Speech Corpus for Automatic Speech Recognition. *Available at SSRN 4397841*.

[23] Chadha H S, Gupta A, Shah P, Chhimwal N, Dhuriya A, Gaur R et.al 2022 Vakyansh: ASR Toolkit for Low Resource Indic languages. *arXiv preprint arXiv:2203.16512 https://github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus*.

[24] Sak H, Senior A W and Beaufays F 2014 Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

[25] Zeyer A, Bahar P, Irie K, Schlüter R and Ney H 2019 A comparison of transformer and lstm encoder decoder models for asr. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 8-15). IEEE.

[26] Karita S, Chen N, Hayashi T, Hori T, Inaguma H, Jiang Z et.al 2019 A comparative study on transformer vs rnn in speech applications. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 449-456). IEEE.

[27] Gulati A, Qin J, Chiu C C, Parmar N, Zhang Y, Yu J et.al 2020 Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

[28] *A. Baevski, Y. Zhou, A. Mohamed and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations", Proc. Adv. Neural Inf. Process. Syst., vol. 33, pp. 12449-12460, 2020.*

[29] Watanabe S, Boyer F, Chang X, Guo P, Hayashi T, Higuchi Y et.al 2021 The 2020 espnet update: new features, broadened applications, performance improvements, and future plans. In *IEEE Data Science and Learning Workshop (DSLW)* (pp. 1-6). IEEE.

[30] Ko T, Peddinti V, Povey D and Khudanpur S 2015 Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*.

[31] Park D S, Chan W, Zhang Y, Chiu C C, Zoph B, Cubuk E D et.al 2019 Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.