An Experiment on Speech-to-Text Translation Systems for Manipuri to English on Low Resource Setting

Loitongbam Sanayai Meetei[1], **Laishram Rahul**[2], **Alok Singh**[1], **Salam Michael Singh**[1],
Thoudam Doren Singh[1], and **Sivaji Bandyopadhyay**[1]

[1]Centre for Natural Language Processing (CNLP) & Dept. of CSE, NIT Silchar, India
[2]Dept. of CSE, SIT, Tumkur
{loisanayai,laishramrahulib,alok.rawat478,salammichaelcse,
thoudam.doren,sivaji.cse.ju}@gmail.com

Abstract

In this paper, we report the experimental findings of building Speech-to-Text translation systems for Manipuri→English on low resource setting which is first of its kind in this language pair. For this purpose, a new dataset consisting of a Manipuri-English parallel corpus along with the corresponding audio version of the Manipuri text is built. Based on this dataset, a benchmark evaluation is reported for the Manipuri→English Speech-to-Text translation using two approaches: 1) a pipeline model consisting of ASR (Automatic Speech Recognition) and Machine translation, and 2) an end-to-end Speech-to-Text translation. Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) and Time delay neural network (TDNN) Acoustic models are used to build two different pipeline systems using a shared MT system. Experimental result shows that the TDNN model outperforms GMM-HMM model significantly by a margin of 2.53% WER. However, their evaluation of Speech-to-Text translation differs by a small margin of 0.1 BLEU. Both the pipeline translation models outperform the end-to-end translation model by a margin of 2.6 BLEU score.

1 Introduction

In recent times, the advance in machine translation (MT) systems research jumped from textual modality to multi modality. The success of the several machine translation system for major languages based on statistical and neural approaches shed light towards building better translations systems for low resource languages as well. Of these, the statistical machine translation (SMT) (Koehn et al., 2003) and neural machine translation (NMT) models (Cho et al., 2014) started its journey from the traditional text-to-text translation which further expanded to the use of multiple modalities (Huang et al., 2016; Caglayan et al., 2016; Meetei et al., 2019; Gain et al., 2021) in the translation task. The usage of multiple modalities in MT uncovers new avenues for MT researchers. MT tasks where multiple modalities are utilized include using multiple-input modalities, for example, incorporating visual and text modalities (Meetei et al., 2021; Singh et al., 2021), translation between different input and output modalities such as Speech-to-Text translation (Ney, 1999; Weiss et al., 2017), etc. With these various methodologies of MT, the main goal is to obtain the most key information in a modality in generating the optimal sentence translation.

The Speech-to-Text (S2T) translation is the translation of a speech in a source language to a target language text. The Speech-to-Text translation task can be broadly addressed using two approaches: 1) with a pipeline strategy, which separates the different modalities into modality conversion, i.e., ASR, followed by text-to-text MT. 2) end-to-end (E2E) translation where the target text is directly generated from the speech in the source language. The Speech-to-Text (S2T) can find its application in our daily life by creating an ease form of communication for individuals with physical disabilities. It can also be used in reducing the turnaround of quick documentation, generating subtitles, etc.

Despite the fact that researchers are pushing the frontiers in machine translation and improving their capabilities, most of the work is focused on well-studied languages while work on low resource languages such as Manipuri is falling behind. Manipuri (also known as Meiteilon) is the official language of Manipur, a northeastern state of India. Manipuri is an extremely low resource language with a limited dataset available for the NLP (Natural Language Processing) tasks which is one of the primary reasons that hindered the development of NLP systems for the language.

Proceedings of the 18th International Conference on Natural Language Processing, pages 54–63
Silchar, India. December 16 - 19, 2021. ©2021 NLP Association of India (NLPAI)

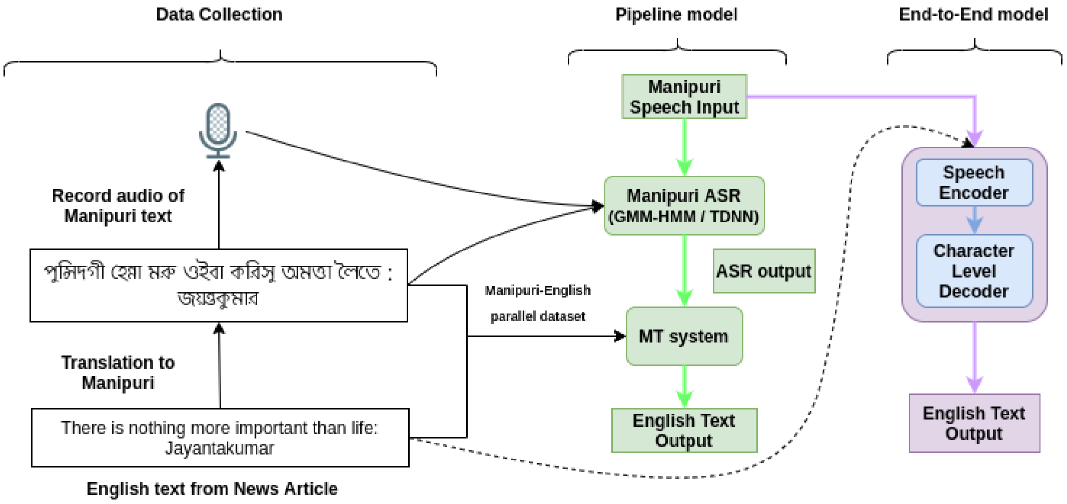
Data Collection
Pipeline model
End-to-End model
Manipuri Speech Input
Record audio of Manipuri text
Manipuri ASR (GMM-HMM / TDNN)
Speech Encoder
Character Level Decoder
পুন্সিদগী হেন্না মরু ওইবা করিসু অমত্তা লৈতে : জয়ন্তকুমার
ASR output
Manipuri-English parallel dataset
MT system
Translation to Manipuri
English Text Output
English Text Output
There is nothing more important than life: Jayantakumar
English text from News Article

Figure 1: Manipuri→English S2T translation models

This work aims to promote Speech-to-Text translation of an extremely low resource language by presenting a benchmark evaluation on a manually collected speech dataset. This work makes the following contributions:

- We build the first Manipuri→English S2T translation dataset.

- Comparison between a pipeline and end-to-end S2T translation model on the collected corpus is reported as the benchmark evaluation.

The rest of this paper is presented as follows: The prior relevant research is discussed in Section 2, followed by the framework of our model in Section 3. Section 4 and Section 5 explain the setup of our system and analysis of our results. The conclusion and future work are summarized in Section 6.

2 Related Works

Early attempts to address S2T translation follows a pipeline approach of two independent models: ASR and MT systems (Ney, 1999; Matusov et al., 2005). The approach utilized the hypothesis of ASR as an input to the MT model to generate the target-language text. Initial work on direct Speech-to-Text translation includes (Bérard et al., 2016; Duong et al., 2016; Bansal et al., 2017). Using a small French-English synthetic dataset from 7 speakers, Bérard et al. (2016) carried out an end-to-end S2T translation. The author reported that their system to be capable of generalizing to a new speaker effectively. Bansal et al. (2018) carried out an end-to-end S2T translation in low resource settings by training with smaller subsets of 160 hours labeled data. The author reported a BLEU score of 5.3 and 29.4 when trained with 20 hours and 160 hours, respectively.

Some of the work in the development of speech technology for the Manipuri language includes Rahul et al. (2013); Patel et al. (2018); Devi et al. (2021). Patel et al. (2018) reported a WER of 19.28% on a GMM-HMM and WER of 13.57% on a Deep Neural Network-HMM (DNN-HMM) acoustic model systems. The speech corpus used in the experiment comprised around 61 hours. Works on MT for Manipuri-English language pair are reported using various techniques such as Example-based MT (Singh and Bandyopadhyay, 2010a), SMT (Singh and Bandyopadhyay, 2010b; Singh, 2013), and unsupervised NMT (Singh and Singh, 2020). In a comparative study of SMT and NMT systems on the Manipuri-English language pair, the authors (Rahul et al., 2021; Singh and Singh, 2021) reported NMT system to perform better than the SMT system. To date, there is no work in S2T translation for Manipuri-English language pair. In order to fill this gap, a Manipuri-English S2T translation is developed using a small dataset in our work.

3 Methodologies

Figure 1 illustrates the methodology of our work. As the initial step of our work, English text dataset is collected from news articles, which is translated to Manipuri language. In the next step, speech is recorded for the Manipuri text. The overall collected dataset is then used to train the pipeline and End-to-End S2T translation models.

3.1 Language Resources

To build the dataset for our experiment, we collected news articles reported in English from a local daily newspaper[1]. The collected English text is machine translated to Manipuri followed by manual post-editing of the MT output and training the MT system with the incremental approach (Meetei et al., 2020). Following the development of the parallel dataset, speech is recorded for each of the Manipuri sentences by the native speakers of Manipur. The total number of participants for speech records is five: one male speaker and four female speakers. There is no overlapping of utterances among the participants. The recorded speech is post-processed, where the quality of speech records are verified manually. Any invalid speech found is rerecorded to collect quality speech records for the experiment. The overall collected dataset comprises of:

- 3500 Manipuri-English parallel text datasets, and
- around 5 hrs 30 minutes of speech record of the Manipuri text.

3.2 Speech Feature Extraction

For any Speech-to-Text system, extracting the audio signal components that can be used to determine linguistic content is important. Mel-frequency cepstral coefficients (MFCCs), the most popular, extensively utilized cepstral feature for ASR, is used as the audio feature for the ASR system and the E2E Speech-to-Text translation system.

3.3 Pipeline translation model: ASR and MT

Our pipeline S2T translation model consists of two independent models:

- Automatic Speech Recognition, and
- Neural Machine Translation (NMT)

In our work, we built two separate pipeline systems using GMM-HMM and TDNN Acoustic models, which is followed by a shared NMT system. The ASR output is fed to the NMT system to generate the target language.

3.3.1 Automatic Speech Recognition (ASR)

The objective of an automatic speech recognition system is to predict the most likely discrete symbol sequence from a given input acoustic speech vector O, out of all valid sequences in a target language T. Taking input speech sequence as a set of observation O= $(o_1, o_2, ...o_n)$ and the symbol to be predicted represented by S = $(s_1, s_2, ...s_n)$, the aim of the ASR model is:

$$\hat{S} = argmaxP(O|S)P(S). \qquad (1)$$

where P(S) is the prior probability for the sequence S, and the observation likelihood, P(O|S) is the likelihood of the acoustic input sequence O given the sequence S, computed using HMM.

The acoustic model based on deep neural networks is trained with time delay neural network, TDNN (Peddinti et al., 2015).

3.3.2 Neural Machine Translation (NMT)

A Neural Machine Translation (NMT) is built for the MT system in the pipeline model. For a source sentence, $\mathbf{S} = \{s_1, \ldots, s_n\}$, NMT, an encoder-decoder sequence-to-sequence technique, jointly models the conditional probability $p(\mathbf{T}|\mathbf{S})$ to translate a target sequence, $\mathbf{T} = \{t_1, \ldots, t_m\}$.

Following the attention mechanism (Bahdanau et al., 2014; Luong et al., 2015), a bi-LSTM (Sutskever et al., 2014) is used as an encoder. At time step t, the encoder state is represented by the concatenation of the forward hidden state, $\vec{h_i}$, and backward hidden state, $\overleftarrow{h_i}$. As each word in the output sequence is decoded, the attention mechanism learns where to focus attention on the input sequence.

3.4 End-to-End S2T translation model

Our end-to-end S2T translation model follows Bérard et al. (2018) architecture, an attentive encoder-decoder model. The speech encoder takes audio features, X= $(x_1, x_2, ..., x_{T_x}) \in \mathbb{R}^{T_x \times N}$ as an input sequence. The audio features are fed into two non-linear $(tanh)$ layers, which generate $\acute{N}$ size features. The new feature sequence length is reduced by a factor of 4 using two 2D convolutional layers with stride (2; 2), which is then passed to a three stacked bidirectional LSTMs (Schuster and Paliwal, 1997). The decoder generates target-language sequences at the character level. The character-level decoder is composed of a conditional LSTM with the global attention mechanism (Bahdanau et al., 2014).

[1] Imphal Free Press https://www.ifp.co.in/

	sentences	duration (in min)
train	3300	~314
dev	100	~9
test	100	~8

Table 1: Manipuri→English Speech-to-Text translation dataset setup

.

4 Experimental Setup

In this section, we present the different Speech-to-Text translation experiments conducted, including the dataset and experimental setup.

The training, development, and test data sets for Manipuri→English S2T translation models are summarized in Table 1.

4.1 Pipeline S2T Translation Models

The system set up of independent ASR and MT systems of pipeline S2T translation model are as follows:

4.1.1 ASR systems

The transcript of the Manipuri text is written in Bengali script. Words in Manipuri have exact grapheme-to-phoneme mapping. A grapheme-to-phoneme list for the Manipuri ASR system is prepared by using the Bengali to Roman script transliteration module of (Meetei et al., 2021). The acoustic features fed to the GMM-HMM model consists of 13-dimensional MFCC, and 3-dimensional pitch features for speaker adaptation, namely Probability of Voicing (POV)-weighted mean subtraction over 1.5 second windows, Normalized Cross Correlation Function (NCCF)-derived POV feature, and delta pitch calculated on raw log pitch. While TDNN acoustic models are trained using 40-dimensional MFCC with 100-dimensional i-vectors and 3-dimensional pitch features. We utilized a 3-gram model trained with SRILM (Stolcke, 2002) for decoding. The ASR systems are built using the Kaldi toolkit (Povey et al., 2011).

4.1.2 NMT systems

Two NMT systems are trained using different dataset set up:

- NMT_{in}: NMT model trained with the in-domian dataset (Table 1).
- NMT_{g}: NMT model trained by combining the in-domain and additional parallel Manipuri-English text dataset. The additional dataset is acquired from TDIL[2], data scrapped from vikaspedia [3] which are then manually aligned and the work from (Meetei et al., 2020). Overall, the domain of the dataset is from tourism, agriculture, medical and news articles. The total training dataset size is 23126 (3300 in-domain and 19823 additional parallel sentences).

As an encoder, a two-layer bi-LSTM with 512 hidden units is used, and the batch size is set to 32. With a learning rate of 0.001 and Adam optimizer (Kingma and Ba, 2014), we train the system utilizing early stopping, where training is halted if a model does not progress on the validation set for more than 15 epochs.

4.2 End-to-End S2T Translation Model

End-to-End S2T translation models are implemented in `PyTorch` (Paszke et al., 2019) with `fair-seq` toolkit[4]. We utilize "`T-Sm`" architecture (Wang et al., 2020) with default hyper-parameters and train with Adam optimizer and a learning rate of 0.002. Early stopping is used to halt the training when the system does not improve for 15 epochs on the development set.

4.3 Evaluation Metrics

The word error rate (WER), which is the ratio of word insertion, deletion, and substitution errors in a transcript to the total number of uttered words, is used to evaluate our ASR systems. The final hypothesis of S2T are evaluated with BLEU (Papineni et al., 2002). BLEU is a precision-based automatic metric used to evaluate the quality of machine-translated text.

5 Results and Analysis

In this section, we illustrate the results of our Manipuri→English pipeline and end-to-end S2T translation models. Along with the automatic metric evaluation, we carried out an in-depth qualitative analysis and human evaluation of our translation systems.

5.1 Automatic Metrics based Evaluation

The ASR systems are evaluated in terms of word error rate (WER), and the final hypothesis of translation from the pipeline and end-

[2] `https://tdil-dc.in/`

[3] `https://vikaspedia.in/`

[4] `https://github.com/pytorch/fairseq`

	Acoustic Model	WER	MT	BLEU	Translation Model
Pipeline	GMM-HMM	27.69	NMT_{in}	6.1	PipeHmmIN
			NMT_g	4.6	PipeHmmG
	TDNN	**25.16**	NMT_{in}	**6.2**	PipeTdnnIN
			NMT_g	4.1	PipeTdnnG
E2E	-	-	-	3.6	E2E

Table 2: Manipuri→English Speech-to-Text translation results

.

transcript1	অহুমশুবা কৱাথা কুম্মৈ হৌদোকখ্রে *Third Kwatha Festival inaugurated*
GMM-HMM	৩ শুবা কৱাথা কুম্মৈ হৌদোকখ্রে
TDNN	অহুমশুবা কৱাথা কুম্মৈ হৌদোকখ্রে
transcript2	দ্রাইভরশিংগী য়ুনিয়ননা বন্দ য়েথোকখ্রে *Drivers union suspends bandh*
GMM-HMM	দ্রাইভরশিংগী য়ুনিয়ননা বন্দ য়েথোকখ্রে
TDNN	দ্রাইভরশিংগী য়ুনিয়ননা ভাবন য়েথোকখ্রে
transcript3	ওল জিরিবাম রোড ত্রান্সপোর্ট দ্রাইভরস য়ুনিয়ননা বন্দ য়েথোকখ্রে *All Jiribam Road Transport Drivers Union suspends bandh*
GMM-HMM	ওল জিরিবাম ভোট ট্রান্সপোর্ট দ্রাইভর য়ুনিয়ননা বন্দ য়েথোকখ্রে
TDNN	ওল জিরিবাম ভোট ট্রান্সপোর্ট দ্রাইভর য়ুনিয়ননা বাল য়েথোকখ্রে
transcript4	লৈবাক ৩১ লানলবা মতুংদা ইন্দিয়ান বাইকরশিং ইম্ফাল য়ৌরকখ্রে *Indian bikers reach Imphal after crossing 31 countries*
GMM-HMM	লৈবা ৩১ লানলবা মতুংদা ইন্দিয়ান বাইকরশিং ইম্ফাল য়ৌরকখ্রে
TDNN	লৈবা ৩১ লানলবা মতুংদা ইন্দিয়ান বাইকরশিং ইমফাল য়ৌরকখ্রে

Table 3: Sample input-output of Manipuri Automatic Speech Recognition systems

to-end systems is measured in terms of BLEU score using SacreBLEU (Post, 2018). Table 2 shows the automatic evaluation score of the ASR (GMM-HMM and TDNN) output and the translation output. The signature of the SacreBLEU is : $BLEU + case.mixed + numrefs.1 + smooth.exp + tok.13a + version.1.5.1$.

- **ASR:** TDNN model outperforms the GMM-HMM model significantly by achieving an improvement of 2.53% WER.
- **Translation**: The pipeline model with TDNN ASR and NMT_{in} achieve the highest BLEU score.

From the results in Table 2, it is observed that the evaluation of the target language translations from the output of the ASR systems using a shared NMT system differ by a small margin. The TDNN pipeline model achieve a 0.1 to 0.5 BLEU score more than the GMM-HMM pipeline model.

Comparing the evaluation scores of the translation hypothesis from the pipeline and End-to-End models, it is clear that the pipeline models outperforms the End-to-End model significantly by a margin of 2.6 BLEU score. The result also shows that the usage of additional out of domain data where the size of the dataset is substantially larger than the in-domain dataset size has negative effect on the BLEU score. A likely cause is the use of development and test dataset from the in-domain dataset.

5.2 Qualitative Analysis of Manipuri ASR Systems

Table 3 shows some sample input-output of Manipuri ASR systems where we analyse the robustness of the systems on selected words in the reference transcript highlighted in green.

In **transcript1**, "অহুমশুবা" (~ "ahumsuba" meaning *third*) is generated in its numerical format "৩ শুবা" (~ "3 suba" meaning 3^{rd}) by GMM-HMM ASR system while the TDNN ASR system generate it in its actual format. Though, both the format has same speech feature, TDNN ASR system performs better in n-gram match.

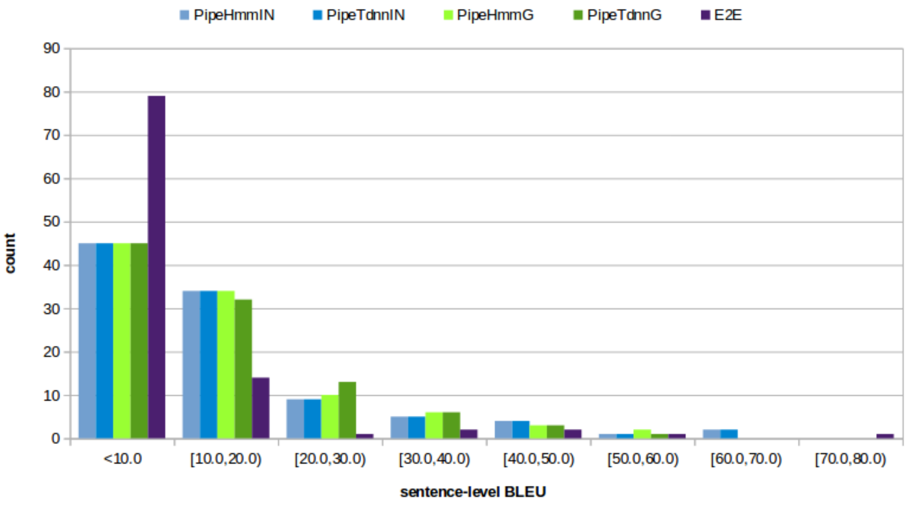

Figure 2: Sentence level BLEU evaluation

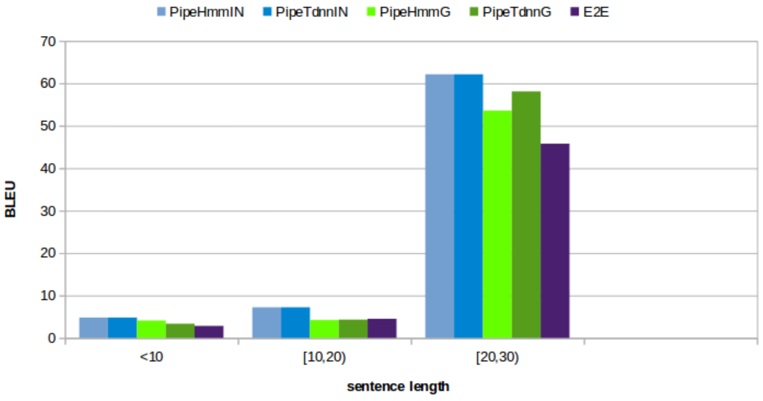

Figure 3: Sentence length BLEU evaluation

The samples **transcript2** to **transcript4** shows some of the examples where ASR systems generates incorrect transcript words (highlighted in "red") of reference words (highlighted in "green"). From the sample results, it is observed that the ASR systems suffer when the word contains the phoneme "b" ("বন্দ" ~ "bandh", "ভাবন" ~ "bhavan", "বাল" ~ "bal").

A single phoneme in Manipuri could be represented by different graphemes in the Bengali script. One such case is shown in **transcript3** where the ASR systems generate the word "ত্রান্সপোর্ট" (~ "transport") as "ট্রান্সপোর্ট" (~ "transport"). In **transcript4**, the word "ইমফাল" (~ "imphal") is a correct representation of the word "ইম্ফাল" (~ "imphal") where the joined characters are written separately.

As the automatic evaluation metrics are computed at the word level, the cases highlighted in **transcript3** and **transcript4** often led to low evaluation score.

5.3 Sentence Level Evaluation

An analysis of the Manipuri→English S2T translation system is carried out by computing the BLEU score at the sentence level. Figure 2 shows the analysis based on the number of sentences with respect to the BLEU score. While the analysis in Figure 3 shows the performance of the systems with shorter and longer sentences based on the length of the reference sentence.

In Figure 2, the majority of the translations from the E2E model are observed to score a BLEU score of less than 10, while less than half of the translations from the pipeline model scored less than 10. It is interesting to note that the highest sentence level BLEU is achieved by the E2E model even though the overall performance of the pipeline model outperforms the E2E model significantly. A likely cause of the poor performance of end-to-end S2T translation system is the small size of the dataset. The result in Figure 3 shows that the systems perform well with longer sentences [20,30)

source1	আর.তি.আই. এক্টকী মতাংদা খংমিন্নবগী থৌরম পাংথোকখ্রে
reference1	Awareness programme on RTI Act held
GMM-HMM ASR	আর তি আই ঈ কী মতাংদা খংমিন্নবগী থৌরম পাংথোকখ্রে
PipeHmmIN	Awareness programme on Mudra Dayal held held
PipeHmmG	Awareness programme on foot held
TDNN ASR	আর তি আই এক্টকী মতাংদা খংমিন্নবগী থৌরম পাংথোকখ্রে
PipeTdnnIN	Awareness programme on tobacco Dayal control held
PipeTdnnG	Awareness programme on Act held at Moreh
E2E	Awareness programme on RTI Act held at Manipur Press Club , Majorkhul
source2	পাওমীশিংগী মীফম মনুংদা ৱারেপ্পা নাবানা মেডিয়াগী মীওইশিংদা ৱা ঙাংখি
reference2	Wareppa Naba speaks to media persons during press meet
GMM-HMM ASR	পাওমীশিংগী মীফম মনুংদা ৱারেপ্পা নাবানা মেডিয়াগী মীওইশিংদা ৱা ঙাংখি
PipeHmmIN	Ng Ibobi speaks to media persons during press conference
PipeHmmG	Ibobi speaks during media persons during press conference
TDNN ASR	পাওমীশিংগী মীফম মনুংদা ৱারেপ্পা নাবানা মেডিয়াগী মীওইশিংদা ৱা ঙাংখি
PipeTdnnIN	Ng Ibobi speaks to media persons during press conference
PipeTdnnG	Ibobi speaks during media persons during press conference
E2E	Ng . Uttam speaks to media persons during press conference
source3	অপুনবা ইরৈপাক্কি মহৈরোই শিনপাংলুপকী মীহুৎশিংনা মেডিয়াদা ৱা ঙাংলি
reference3	Representatives of Apunba Ireipakki Maheiroi Sinpanglup speaking to the media
GMM-HMM ASR	অপুনবা ইরৈপাক্কি মহৈরোইশিং পান লুপকী মীহুৎশিংনা মেডিয়াদা ৱা ঙাংলি
PipeHmmIN	Representatives of Apunba Ireipakki Maheiroi Sinpanglup speaking to media
PipeHmmG	Representatives of Ukhrul woman speaking during the inaugural ceremony
TDNN ASR	অপুনবা ইরৈপাক্কি মহৈরোই শিনপাংলুপকী মীহুৎশিংনা মেডিয়াদা ৱা ঙাংলি
PipeTdnnIN	Representatives of Apunba Ireipakki Maheiroi Sinpanglup speaking to media
PipeTdnnG	Representatives of Ukhrul woman speaking during the inaugural ceremony
E2E	Representatives of Apunba Ireipakki Maheiroi Sinpanglup speaking to the media

Table 4: Manipuri→English Speech-to-Text translation sample input-output

while the sentences with length below 20 score a BLEU score less than 10.

With only very few samples achieving a BLEU score above 50, it is clear that a massive effort is required for the development of Manipuri→English Speech-to-Text translation systems.

5.4 Qualitative Analysis of Manipuri→English S2T Systems

Sample input and output from the pipeline models and E2E Speech-to-Text translation model are shown in Table 4. The grammatical error or incorrect word(s) in the output from our systems are highlighted in "blue".

In the first sample, despite preserving the information moderately, the fluency scale of translation output with the Pipeline-GMM-HMM is worse compared to the other systems. One of the main reason is error propagation from the ASR model where word "এক্টকী" (~ "act-ki") is incorrectly generated as "ঈ কী" (~ "e-ki"). Furthermore, E2E translation model generate additional non-relevant information even though the output sentence is fluent. In the second sample, the named entity word "Wareppa Naba" (a name of a person) is incorrectly generated and is replaced by another name of a person (i.e., Ng Ibobi, Ibobi and Ng . Uttam). In different languages, there are cases where multiple words in one langugae is represented by a single word in another language. One such case is highlighted in the second sample where both the words meet and conference which are synonyms is represented by a single word in Manipuri "মীফম" (~ "mifam"). This often results to low BLEU score as the evaluation metric is computed at the word level n-gram matching and doesn't consider synonyms. The third sample shows the handling of long Manipuri multi-word named entity "অপুনবা ইরৈপাক্কি মহৈরোই শিনপাংলুপকী" (~ "Apunba Ireipakki Maheiroi Sinpanglup-ki"), where the suffix "-ki" is used to denote the possessive noun. It is observed that the multi-word named entity is translated correctly despite the slight variation in the output of

Score	Adequacy	Fluency
1	No information is preserved	Incomprehensible
2	Small amount of information is preserved	Disfluent
3	Moderately preserved information	Non-native
4	All information is preserved	Flawless sentence, and all are correct in terms of grammatical rules

Table 5: Adequacy Fluency scale

	Adequacy	Fluency
PipeHmmIN	1.6	2
PipeHmmG	1.5	1.63
PipeTdnnIN	**1.63**	1.98
PipeTdnnG	1.3	1.93
E2E	1.23	**2.83**

Table 6: Human Evaluation of Manipuri→English S2T Translation Systems.

the ASR system (GMM-HMM). This is the impact of the system trained on in-domain training dataset. However, the translation is not in the line with the NMT system trained on mixed domain training dataset (NMT_g) as the probability distribution got skewed towards the add-on dataset.

5.5 Adequacy and Fluency Analysis of Translation Outputs

Fluency analysis provide evaluation based mainly on grammatical rules. Adequacy indicates information preserved. Adequacy and fluency are measured on a scale of 1 to 4 and the meaning of the various scales are summarized in Table 5. To measure adequacy and fluency, human evaluation on the test dataset from each S2T translation system is carried out. The adequacy and fluency ratings reported by our human evaluators are shown in Table 6.

- Among our translation systems, the pipeline model (PipeTdnnIN) achieves the highest adequacy score. The adequacy score of all the systems are observed to be in correlation with our automatic evaluation.

- The fluency score is observed to be non-correlated with the automatic evaluation scores. In terms of fluency, the end-to-end model achieved the highest score. This indicates that despite not preserving the information of the source language, the system is able to generate a fluent text.

6 Conclusion and Future work

In this work, a comparative study of the conventional pipeline model and end-to-end model of S2T translation on an extremely low-resource Manipuri-English language pair is presented. We also made a comparison of two acoustic models: GMM-HMM and TDNN, for the ASR module. An improvement of 2.53% WER is observed in the ASR model with TDNN compared to GMM-HMM. The TDNN ASR model is observed to be more robust than the GMM-HMM model in terms of n-gram match. The ASR output is fed to a shared NMT system (trained with the in-domain or the additional out of domain dataset) in our pipeline model. In comparison, the translation hypothesis of the pipeline models are comparable in terms of the BLEU score. However, using an NMT system trained with a dataset from mixed domain results to the decrease in the automatic evaluation score. Though the end-to-end S2T translation has various advantages over traditional pipeline models, the limited size of our dataset led to the end-to-end S2T model's low performance compared to the pipeline model. An extensive collection of parallel S2T translation training data is generally required to train such an end-to-end S2T translation model.

In future, we plan to increase the size of the dataset along with the collection of other forms of modalities such as images. We also plan to explore various Speech-to-Text machine translation models to enhance the performance.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. *arXiv preprint arXiv:1803.09164*.

Sameer Bansal, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. *arXiv preprint arXiv:1702.03856*.

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, Laurent Besacier, and Christophe Servan. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost Van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? *arXiv preprint arXiv:1605.09186*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *arXiv preprint arXiv:1406.1078*.

Thangjam Clarinda Devi, Leihaorambam Sarbajit Singh, and Kabita Thaoroijam. 2021. Vowel-based acoustic and prosodic study of three manipuri dialects. In *Advances in Speech and Music Technology*, pages 425–433. Springer.

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.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Iitp at wat 2021: System description for english-hindi multimodal translation task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 161–165.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Franz J Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Evgeny Matusov, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *Ninth European Conference on Speech Communication and Technology*.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. Wat2019: English-hindi translation on hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021. Low resource multimodal neural machine translation of english-hindi in news domain. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 20–29.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, Sivaji Bandyopadhyay, Mihaela Vela, and Josef van Genabith. 2020. English to manipuri and mizo post-editing effort and its impact on low resource machine translation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 50–59.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 517–520. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Tanvina Patel, DN Krishna, Noor Fathima, Nisar Shah, C Mahima, Deepak Kumar, and Anuroop Iyengar. 2018. An automatic speech transcription system for manipuri language. In *INTERSPEECH*, pages 2388–2389.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Laishram Rahul, Loitongbam Sanayai Meetei, and HS Jayanna. 2021. Statistical and neural machine translation for manipuri-english on intelligence domain. In *Advances in Computing and Network Communications*, pages 249–257. Springer.

Laishram Rahul, Salam Nandakishor, L Joyprakash Singh, and SK Dutta. 2013. Design of manipuri keywords spotting system using hmm. In *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pages 1–3. IEEE.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Salam Michael Singh, Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2021. Multiple captions embellished multilingual multi-modal neural machine translation. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 2–11.

Salam Michael Singh and Thoudam Doren Singh. 2020. Unsupervised neural machine translation for english and manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78.

Salam Michael Singh and Thoudam Doren Singh. 2021. Statistical and neural machine translation systems of english to manipuri: A preliminary study. In *Soft Computing and Signal Processing*, pages 203–211. Springer.

Thoudam D Singh and Sivaji Bandyopadhyay. 2010a. Manipuri-english example based machine translation system,". *International Journal of Computational Linguistics and Applications (IJCLA), ISSN*, pages 0976–0962.

Thoudam Doren Singh. 2013. Taste of two different flavours: Which manipuri script works better for english-manipuri language pair smt systems? In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 11–18.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010b. Manipuri-english bidirectional statistical machine translation systems using morphology and dependency relations. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 83–91.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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.

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*.