Research Article

An Effective Speech-To-Text Translation Using Attention Fusion of Multimodal Conformer and Transformer Approach

Avinash Kumar Reddy Kayyam[1], D Naga Malleswari[2]

[1]M.Tech. Scholar, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur,522502, Email: 2401050067cse@gmail.com
[2]Associate Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation (KL University), Green Fields, Vaddeswaram, A.P. – 522502, Email: nagamalleswary@kluniversity.in

Corresponding Author: Avinash Kumar Reddy Kayyam, mail id: 2401050067cse@gmail.com

Abstract

There are many applications such as education, online meetings, healthcare, and digital communication in which direct conversion of spoken language into required text is needed using Multilingual speech-to-text translation (MSTT), which focuses on eliminating intermediate steps to reduce complexity. In this process, there is always a challenge during translation in different languages, errors raised that reduce translation quality, and addressing acoustic variance. Tools such as mBART, Whisper, and wav2sec 2.0 illustrate Automatic speech recognition (ASR) and Machine Translation (MT) as complete pipeline models. In these approaches, existing translation methods and conventional models suffer from error propagation; ASR has issues with inaccuracies that degrade translation quality; and end-to-end mechanisms suffer from low-resource languages andlimited acoustic diversity. Other issues, such as limited adaptability, bias, and alignment issues between speech-to-text, were experienced by pretrained models. To eliminate these shortfalls, an integrated approach fuses XLS-R and WavLM pretrained models for speech recognition, integrates SpecAugment on Conformer, a gated cross-attention transformer decoder, hybrid connectionist temporal classification and cross-entropy loss, and learning-rate warm-up with the FedAdam optimizer. Evaluation results on the proposed model with existing models show superior BLEU score performance, WER reduction, and faster convergence under real-world communications.

Keywords: Pretrained models fusion, Transformer encoder, Decoder, Fusion, Metrics assessment, and Accuracy.

1. INTRODUCTION

Applications such as international business platforms, international education, international tourism, healthcare domains, and medical platforms require speech translation from diverse languages into the required translated text. In these domains, multilingual translation is necessary; hence, research focus is more on this context. This translation assists doctors and patients of different language backgrounds in healthcare; online translation of one language into the required language in education; customer support and assistants improve user experience would require accurate multi-language translation, and communication and media-related digital content generation, AI communication, and online meetings, which possess noise and low resource settings, would require efficient and accurate translation as a requirement.

The methods considered in the analysis include automatic pipelines of ASR and MT, then end-to-end transformer models, pretrained mechanisms like Whisper, XLSR, mBART, wav2sec 2, and coConformer approaches. Although ASR+MT involves automatic translation over scalable data sources, it lacks error injection, introduces direct mistakes, and lacks optimization mechanisms. When the encoder and decoder were performed as a two-step mechanism. Leads to inconsistency and information loss due to cascading pipelines. Conventional end-to-end transformers without intermediate modules translate text directly, ensuring advantages such as simplifying pipeline functionality with increased learning capacity on one side, and, on the other side, lacking issues such as long-sequence speech handling, inconsistencies due to low resource constraints, and translation quality reduction due to variability of accents. Few studies have demonstrated limitations in ensuring robustness, accent, and adaptation quality. Among pretrained tools,Whisper has high overhead and involves high mistranslations, although it handles under-resource conditions; eM-BART and wav2vewav2vec 2 require fine-tuning parameter and limited adaptability, although they improve capability. Another aspect of the translational method is the conformer model, in which operations applied are convolution and self-attention as local and global speech features, leading to enhanced learning capability, whereas shortfalls include being noise-sensitive, tone diversity issues, issues with synchronization, and insufficient robustness alignment. Overall, the drawbacks associated are semantic

inconsistency, complexity in handling diverse accents, error propagation in token alignments, and degraded performance.

To overcome issues raised by various existing approaches, it is required to develop an adaptive multilingual model with an encoder, decoder, conformer module, Gated mechanism with fused attention, and FedAdam as an optimizer to output expected translation results.

Hence, an enhanced model is proposed that consists of pretrained speech encoding of WavLM/XLS-R fusion, SpecAugment over the Conformer module, Cross Gated Attention Decoder, Cross-Entropy loss with Hybrid CTC, and Learning-rate warmup with FedAdam optimization. Each component provides specific benefits in converting one source to another target. Among the components' functionalities, the first component is WavLM and XLS-R fusion, where XLS-R supports large datasets and low-resource settings, whereas WavLM supports speaker tone, acoustic noise, and accent variation. Another component in the proposed system flow is the Conformer module, which combines convolution and self-attention: convolution learns local short-term phonetics, whereas self-attention captures global long-range details. To mask the time and frequency, SpecAugment is preferred, which reduces overfitting and eliminates any associated shortfalls. A gated cross-attention transformer decoder, as the third component in the flow, translates specific acoustic data into text sequences, where the decoder learns tokens using self-attention, whereas cross-attention aligns speech with target embeddings. A gated mechanism is used to suppress unnecessary details to improve consistency. An integrated Hybrid CTC and Cross-Entropy loss as a fourth component performs token alignment and sequence generation. The result of this component is aligning frames with tokens, excluding strict annotations, and eliminating errors. In addition, Cross Loss is used to focus on accurate target token alignment, and the use of hybrid loss focuses on consistency improvement, lowering Word Error Rate (WER), and achieving translation performance. Finally, FedAdam is used as the optimizer, improving stability and training speed by adaptively assigning weights via first- and second-order moments, with a learning-rate warmup to avoid unstable updates and improve model stability.

2. METHODS ANDE DISCUSSION

Issues associated with existing methods, such as errors due to weak generalization and semantic inconsistency in multilingual translation, were experienced by conventional translational methods, pretrained methods, ASR and MT pipelines, and advanced mechanisms. The studies demonstrated on these mechanisms were categorized into specific domains based on their theme, such as (i) Multilingual Speech Translation and End-to-End systems, (ii) Regional language and low resource environments, (iii) LLM, knowledge-enhanced, and Transformer-based, and (iv) Healthcare, Assistive, and Accessibility translations.

2.1 Multilingual Speech Translation and End-to-End systems

This domain of studies focuses on converting spoken content from one language into another while reducing dependency on intermediate ASR and MT pipelines, whereas end-to-end approaches focus on achieving semantic consistency, reducing latency, and reducing error propagation across multiple language diversities. Dash, P., Babu, S., Singaravel, L. et al. (2025) [1] note that in a nation like India, where multiple languages exist, translating from one language to another in this bilingual context is not suitable. Hence, automatic translation systems would encode human speech into the required language without errors. To perform translation, uses generative AI pretrained transformers that predict subsequent words in a language and learn language structure and semantics. An LSTM is integrated to capture temporal dependencies in generating accurate translations. By conducting research on 50 Tamil-English sample translations in a bilingual context, it achieves 83% on short terms and 74% on long sentences. As per [2], the study states that a tool named Bellfish that translates between two languages requires expertise and innovation. Although traditional speech-to-speech translation has existed, it lacks scalability and high performance. A novel approach namely multilingual multimodal machine massive translation M4T SEAMLESS is developed that supports translations on 101 languages in both speech and text, whereas single model approaches won't support 101 languages in which speech to speech, and text to speech support 36 languages, and other translations support 96 languages. Nivedita Sethiya and Chandresh Kumar Maurya (2025) [3] denote that traditional ST translations use ASR for spoken words identification and MT for translated text in the required language, suffer from training costs, rich resource environments, and error propagation. To overcome these, forms an end-to-end ST module that eliminates intermediates and directly translates from source to target. This study demonstrates existing methods, challenges, datasets used, and future directions for efficient translation. Changhan Wang, Anne Wu, and Juan Pino (2025) [4] provide a large multilingual speech-to-text translation

dataset, namely CoVoST2, with characteristics supporting low-resource languages, multilingual training and scaling, zero-shot transfer learning, and an end-to-end ST model. This process involves ASR, MT, and ST baselines that convert 21 languages to English, and from English to 15 languages. Feng, X., Zhao, Y., et al. (2024) [5] dictate that converting speech to text using multimodal contexts, due to limited resources, is challenging. The integration of Speech translation and machine translation in multitasking, from one source type to another type, improves translation accuracy. An adaptive multitasking approach is proposed by assigning weights to tasks during speech-to-text translation. Disparities exist between speech-to-text translation over heterogeneous modalities, which are reduced by transport optimality from end-to-end processes. This model was applied to English-to-German, French-to-French, and Tibetan-to-Chinese. Li, Zichao, Xu, Yanyan, et al. (2026) [7] demonstrate that direct S2S became a realistic solution for cross-lingual domains, suffering from translational quality at the decoder stage. To overcome this, a novel method, namely a dual-pass encoder, is designed in which the speech encoder operates on acoustic representations and the text encoder on semantics. This manages cross-model flow with optimal resource allocation. This uses ASR and MT tasks that make use of discrete units as intermediaries. This framework uses pretraining and natural models for ensuring accuracy and efficiency, over datasets such as s2sGiga and CVSS-C. Sarkar, Balaram, Karande, Pranav, et al. (2026) [14] depict those existing approaches of speech-to-text translation, such as MT and ASR, struggle with latency, error propagation, and memory issues. Adaptive E2E ST experiences deterioration due to new language pairs, also called catastrophic failure. Hence, a continual learning E2E ST approach combines PLS, augmentation, and RS, where RS aggressively samples, and PLS results in overfitting. Combined RS and PLS, called a continual bi sampler, overcomes the issue of forgetting and is applied to 4 pairs of MUST-C and mTEDx datasets. Tran, Chung, Sakti, Sakriani, et al. (2026) [18] depict various translations, but text-to-speech (T2S) is not widely explored and lacks acoustic and linguistic diversity. SAM translator with mBART is defined as a model that combines two properties: self-paced learning focuses on efficiency and stability, and mixture of experts provides selective experts for cross-linguistic variation modeling. Low-resource constrained environments were also handled with augmented self-training. Over the CVSS-C dataset, the derived model performs well and improves the BLEU score by 3% in high-resource and 10% in low-resource environments. Evaluations on different combinations observed some improvement over other models. Weng, Zhenzi, Wang, Zijing et al. (2026) [20] propose a robust speech-to-text translation system called Ross S2T, which is provided by a deep semantic encoder, performs semantic exchange from end to end, and leads to performance without affect. To handle impairments in corrupted speech, a GAN deep semantic encoder and an aided compensator are defined to extract hidden information in speech and extract deep semantic features as well as outperform other models. Aggarwal, Kritika, Lohmor, Anchal, et al. (2026) [21] performed the development of soft linguistic communication in a global environment, in which language obstacles were minimized using a combination of ASR for diverse dialects, MT for accurate translation, and T2S tasks. Transformed models were used to increase accuracy to 98.89% and stability.

2.2 Regional language and low-resource environments: This set of studies describes linguistically diverse communities that lead to the development of translation models with limited datasets, emphasizing aspects such as multilingual adaptation, transfer learning, and efficient dataset construction. Mandiya, Reagan E., Kasereka, Selain K., et al. (2026) [10] denote that Lingala is largely influenced by other cultures but is neglected and has limited datasets, requiring the creation ofion of large datasets for this language. Transformer- and LSTM-based models perform bidirectional French-to-Lingala text-to-text and audio-to-text translations. BERT is a bidirectional encoder that is superior due to its strong, capable contextual representation, with Whisper as the first ASR in the pipeline. Effectiveness is defined in terms of BLEU, which is increased for the final BERT pipeline compared to the standard Deep Learning Transformer approach and integrated Whisper methods. Prama, Tabia Tanzin, Oni, Mangsura Kabir, et al. (2026) [11] describe Sylheti as a language used by 11 million people in regions of Bangladesh and Sikkim, which is largely neglected. Hence, language translation between Sylheti and Bangla became focused on 5002 paired collections, requiring neural machine translation (NMT) over a collected dataset of 10k words from Bangla and Sylheti, clauses, and sentences. The important aspects performed in this study are translation from Sylheti to Bangla and sentiment analysis. Mulukutla, Rama Rajeswari, Gangashetty, Suryakanth V., et al. (2026) [13] demonstrate various models used to translate Telugu to English, including cascaded pipeline, ASR, MT, TTS, and Hybrid DNN+MM. The hybrid approach uses Telugu utterances of 2628 in Kaldi and achieves a low WER score of 0.7%. This provides an effective regional language translator, using Speech to Text, Google MT, and Gemini AI Vertex, in which synthesis and distinguished items were specified. Sudheeksha, Kanchan, Dishika K., et al. (2026) [23] denote an end-to-end process for converting TuLu to English for speech-to-speech translation, which involves

language barriers. To minimize ASR, it uses speech-to-speech, text-to-text translation, and text-to-speech using an embedding layer, LSTM, and time-distributed layer with SoftMax over a dataset of 3k samples, 150 iterations, ensuring 78% accuracy and 1% loss. To ensure strong semantic and syntactic dependencies, uses categorical cross-entropy and the Adam optimizer.

2.3 LLM, knowledge-enhanced, and Transformer-based: This domain provides methods to ensure robustness over cross-lingual understandings using mechanisms such as transformer architectures, LLM-driven translation, attention mechanisms, semantic reasoning and alignment, distillation, contextual learning, and advanced encoder-decoder pipelines. Jorge Iranzo-Sánchez, Jaume Santamaría-Jordà, et al. (2025) [6] denote the need for LLMs in ST, MT models, and traditional encoder-decoder mechanisms from a general perspective. From a medical domain perspective, Conversion from English to French, German, and Slovene takes place over ESO data of recorded videos using ASR and MT models. This translation adapts the oncology domain and uses three models: a bilingual encoder-decoder, a large pretrained multilingual encoder-decoder, and a multilingual LLM, for an enhanced translation index. Sataer, Yikemaiti, Gao, Miao, et al. (2026) [8] state that LLMs provide impressive capability in language translations due to incorporation of linguistic knowledge. The two prompts, such as linguistically augmented and chain of linguistics, improve classification accuracy using the hierarchical Subject Verb Object (HSVO) schema. The results evaluated are predominantly for the HSVO approach over dependency trees, constituency trees, and semantic role labeling tags. Zhou, Yue, Yuan, Yuxuan, et al. (2026) [9] note that knowledge distillation applies to the teacher module whose output affects the student module, especially transferring partial knowledge where the student cannot get a full translation. Hence, multi-grained knowledge distillation makes use of adaptive word-level and cross-hidden-state distillations, producing outputs to learn sequence-level and hidden representations by overcoming many challenges. Other aspects ensured by multi-stage distillation outperformed other state-of-the-art methods on a dataset, namely MuST-C. Liu, Henglyu, Chen, Andong, et al. (2026) [19] denote advancements in LLM systems, while existing methods that focus on aligning input and output modalities may overlook deeper semantic representations. This issue is eliminated by adaptive inner STA, which aligns text and speech at selected layers using Optimal Transport (OT) to quantify fine-grained features. A cross-modal approach is initiated to identify the best layers that align and perform training. This Adaptive inner speech-text alignment outperforms other existing models due to inner layer speech-text alignment. Li, Yi, Zhao, Rui, et al. (2026) [22] demonstrate that speech signals include paralinguistic cues, which alter identical sentences to provide different meanings. Existing methods that possess paralinguistic cues may degrade performance. To overcome these issues, PlaST is proposed to use novel dual branches, such as a style extractor and a speech encoder, that independently introduce paralinguistic and linguistic notations. Hybrid transport optimization is used on linguistic representation, and paralinguistic information is refined using attention-based retrieval components. On benchmarks such as standard linguistic CoVoST-2 and paralinguistic sets such as ContraProst, the proposed model performs better than other ST models.

2.4 Healthcare, Assistive, and Accessibility translations: This categorization of studies focuses on communication with individuals with disabilities, language barriers, and communication impairments who require translation technologies for healthcare, accessibility support, and disability assistance. Bawankar, Anurag, Sharma, Aniket, et al. (2026) [12] explore signed language for many people with impairments in India, and the lack of efficient tools to handle and translate is challenging. Hence, AI-assisted technologies outperform recognition using CNN and generate translations using transformers and 3D rendering over other mechanisms. A large dataset is constructed that robustly performs in the translation of signs to visuals. Rajalakshmi S., Nishaanth R., et al. (2026) [15] denote that providing a reader for individuals with impairments can be overcome by combining domains such as computer vision, natural language processing, translation, and speech synthesis technologies into a unified approach. Hence, an integrated set of YOLOv8, Pix2Tex, Tesseract, BLIP2, IndicTrans2, and gTTS is used to synthesize text-to-image conversion. For languages, this model enhances multilingual translations effectively by evaluating Mean Opinion Scores above 4.5. Kostiuk, Yevhen, Tonja, AtnafuLambebo, et al. (2026) [16] denote converting hate speech into non-hate speech over Spanish words and tried with many existing models. BLEU metric is evaluated using average metrics per hold and cross-hold validation, and it computes to 0.236 per 4-hold experimentation. Review of research on this aspect leads to more effective models that reduce hate words in online platforms. Mao, Rongjie, Zhu, Yuncheng, et al. (2026) [17] demonstrate that autism spectrum disorder detection, in which speech and language play a key role, hence data mining and automated tools were used. These included automated aids like LENA, Kaldi, and Praat, and foundational ASR methods like Whisper and self-encoders like Wave2Vec 2.0. Various

ML approaches like SVM, data mining approaches like clustering, DL models like CNN, and transformer models like LLMs were used. Challenges are noted, and future directions are recommended in this study.

3. METHODOLOGY

In this section, the proposed methodology is defined to overcome drawbacks represented in the literature review, and the functionality is demonstrated in Figure 1, Algorithm A1, and significant modules from 3.1 to 3.11.

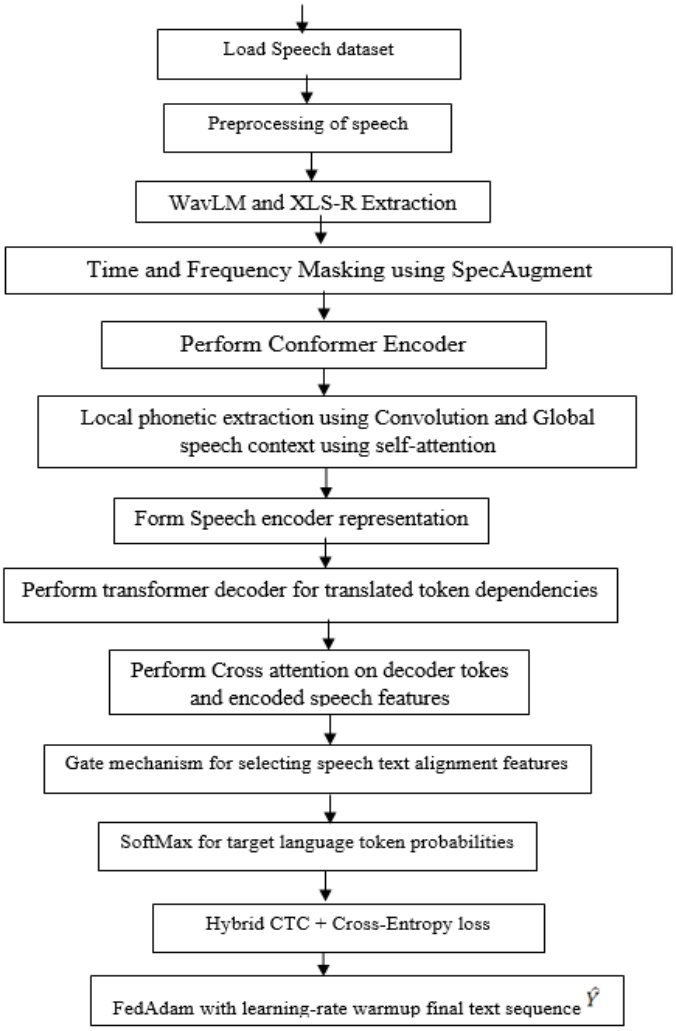


Figure 1. Flowchart of Hybrid multilingual speech to text translation method

Figure 1 demonstrates the flow of activities from the input given to processing each activity, and outputs text for the given speech. This flow is implemented in algorithm A1, which demonstrates a multilingual speech-to-text translation system that translates multilingual speech to text sequences in an end-to-end pipeline. Initially, speech utterances and corresponding text pairs were loaded, then preprocessing is performed to improve quality and minimize acoustic disturbances. Then, pretrained WavLM and XLS-R encoders are fused to capture contextual and phonetic data from diverse languages. To enhance time and frequency masking in low-resource environments, SpecAugment is used. Then, augmented features are propagated to the Conformer encoder, where local phonetic features are captured by convolution layers, whereas global long-range dependencies are captured by self-attention learning. These representations were passed to the transformer decoder, which produces translated text. Then, cross-attention is applied over encoded speech and decoder tokens. A gated mechanism is defined for selective speech-to-text features before SoftMax, which generates target token probabilities. Then, Translation loss

and Hybrid CTC are applied for translation accuracy and optimization of sequence alignment. FedAdam optimizer is applied during training to update model parameters. Finally, the proposed model generates text from speech and is evaluated using measures such as BLEU score, Word Error Rate (WER), and accuracy, which define the effectiveness of the model.

A1: Algorithm Hybrid_multilingual_speechtoText_translation(Speech_dataset[][]):
Input:$\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}\}$ be multilingual speech, and $\mathrm{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{m}\}$ be target text sequence
Output: Text sequence $\widehat{\mathrm{Y}}$
Step1: Load text pairs and speech utterances
Step2: Apply preprocessing on speech for noise removal, resampling, normalization, and silence trimming
Step3: Extract multilingual acoustic features from WavLM and XLS-R and denoted as F_{wavlm} = WavLM(X) and F_{xlsr} = XLSR(X)
Step4: Fuse outputs of pretrained speech representations F_{fusion} = Concatenate(F_{wavlm} , F_{xlsr})
Step5: For time and frequency masking, perform SpecAugment
Step6: Propagate augmented features to Conformer Encoder using H_{conf} = Conformer(F_{fusion})
Step7: Local and global extractions using
H_{local} = Convolution(H_{conf})
H_{global} = SelfAttention(H_{conf})
Step8: Perform final encoder representation using H_{enc} = Fusion(H_{local} , H_{global})
Step9: Initialize previous translated tokens to transformer decoder
Step10: Compute cross attention between speech features and decoder tokens using C_t= CrossAttention(Q,K,V)
Step11: Define Gated alignment standard using G_t = Sigmoid(Wg[C_t ; D_t] + bg)
Step12: Generate contextual alignment representation using $H_{align} = G_t \odot C_t + (1-G_t) \odot D_t$
Step13: Apply SoftMax for target token probabilities using $P(y_t)$ = SoftMax(Wo H_{align} + bo)
Step14: Perform model namely Hybrid CTC + Cross-Entropy loss $L_{total} = \lambda L_{CTC} + (1-\lambda) L_{CE}$
Step15: Apply Adam to update model parameters
Step16: Generate final text sequence using $\hat{Y} = \mathrm{argmax}(P((y_t))$
Step17: Evaluate model performance in terms of BLEU, WER, and Accuracy

Significant modules with justified mathematical formula were demonstrated are as follows

3.1 Speech Processing:

In this, the quality and consistency of multilingual speech signals to be improved. Initially, Let $X = \{x_1, x_2, \ldots, x_n\}$ be multilingual speech, and $Y = \{y_1, y_2, \ldots, y_m\}$ be target text sequence Eq.(1)

Translation accuracy is influenced due to speaker variability, background noise, amplitude fluctuations, and recording environments. Hence normalization is applied for ensuring stable feature learning and faster convergence during training. Performing Speech normalization is as $x_{norm} = \frac{x-\mu}{\sigma}$ Eq. (2) where x is input speech signal, μ denotes mean of speech samples, and σ is standard deviation

3.2 **WavLM and XLS-R Feature Fusion:**

Although pretrained model WavLM captures acoustic and contextual speech notations, while XLS-R denotes multilingual speech across different languages. Their fusion enables to simultaneously learn language-independent phonetic structures and multilingual semantic patterns, for improving translation performance.The fused pretrained multilingual speech representation is denoted as $F_{fusion} = [F_{wavlm}; F_{xlsr}]$ Eq.(3)

Where F_{wavlm}= WavLM acoustic representation, F_{xlsr}= XLS-R multilingual representation, and [;] denotes concatenation operation

3.3 SpecAugment:

Generalization and robustness of the model is improved by masking portions of speech spectrograms, in which time masking removes specific temporal places and prevents overfitting whereas frequency masking hides specific frequency bands to make resistant.

This activity involves masking of time as well as frequency masking over preprocessed sample, in which Time masking is defined by $M_t = X[:, t_0 : t_0 + t]$ Eq.(4)

Frequency masking is denoted as $M_f = X[f_0 : f_0 + f, :]$ Eq.(5) Where t is masked time interval, and f= masked frequency band.

3.4 Conformer Encoder:
The combination of convolution operations and self-attention mechanisms denotes Conformer Encoder where short term local phonetic by convolutions and global long range context information by self-attention.The combination of convolution and self-attention is denoted as $H_{conf} = Convolution(X) + SelfAttention(X)$ Eq.(6)

3.5 Self Attention Mechanism: To identify important relationships between speech tokens, requires self- attention mechanism. Use of Query (Q), Key (K), and Value (V) matrices is to determine token relevance, whereas scaling using $d_k$stabilizes gradient learning.

$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ Eq.(7) Where $Q$is Query matrix, $K$is Key matrix, $V$is Value matrix, and d_k= denotes dimensionality scaling factor

3.6 Cross Attention fusion: To generate interaction between decoder-generated tokens and encoded speech items, cross attention is defined in which decoder highlights relevant speech regions and predicts target token, improves translation consistency.

$C_t = Attention(D_t, H_{enc}, H_{enc})$ Eq.(8) Where $D_t$is decoder hidden representation, and H_{enc}denotes encoded speech representation.

3.7 Gated Alignment: Information extraction between speech representations and decoder states is done by Gated mechanism, in which sigmoid gate dynamically learns speech-text alignment weights, leads to robustness. The computation of gate is $G_t = \sigma(W_g[C_t; D_t] + b_g)$ Eq.(9)

Computation of aligned representation is $H_{align} = G_t \odot C_t + (1 - G_t) \odot D_t$ Eq.(10) Where σ is Sigmoid activation, and $\odot$ is element-wise multiplication

3.8 SoftMax for probability computation for token: Conversion of decoder outputs into probability distributions, and outputs token with highest probability, enable text generation using SoftMax classifier.

$P(y_t) = Softmax(W_o H_{align} + b_o)$ Eq.(11) where$P(y_t)$ is probability of target token, and W_o, b_o denotes output weights and bias.

3.9 Hybrid CTC + Cross-Entropy Loss: Integration of Connectionist Temporal Classification (CTC) and Cross-Entropy (CE), would optimize jointly on sequence alignment against token prediction, ensures speech to text alignment improvement and token prediction accuracy.

Hybrid loss is defined as $L_{total} = \lambda L_{CTC} + (1 - \lambda)L_{CE}$ Eq.(11) where λ is Balancing coefficient, $(1-\lambda)$ is Weight assigned to Cross-Entropy loss, L_{CTC} is Connectionist Temporal Classification loss for sequence alignment, and L_{CE}= denotes Cross-Entropy loss for target token prediction

Cross-Entropy loss is computed by $L_{CE} = -\sum_{i=1}^{N} y_i \, log(\hat{y}_i)$ Eq.(13) where N is Total number of target tokens, y_i is Ground-truth target token label for the i^{th}sample, $\hat{y}_i$ denotes Predicted probability of the correct target token, and $log(\hat{y}_i)$ denotes Logarithmic probability of predicted token

CTC loss is defined as $L_{CTC} = -log\, P(Y \mid X)$ Eq.(14) where $P(Y \mid X)$ denotes Probability of predicting target text sequence Y given input speech sequence X.

3.10 Perform FedAdam Optimization:
Perform adjusting learning rates adaptively using first and second-order gradient moments, correcting bias, improves stability, updates parameters, and learning rate enables faster convergence and reduced training loss. This is made possible using FedAdam.

First moment update is computed by $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$ Eq.(15) where m_{t-1} denotes Previous first moment estimate, β_1 is Exponential decay rate for first moment estimation, $(1-\beta_1)$ is Weight assigned to current gradient , and g_t= Current gradient of loss function with respect to parameters at iteration t

Second moment update is defined by $v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2$ Eq.(16) where v_{t-1} is Previous second moment estimate , β_2 denotes Exponential decay rate for second moment estimation, g_t^2 is Square of current gradient, and $(1-\beta_2)$ denotes Weight assigned to squared gradient

Bias correction is defined as $\widehat{m}_t = \frac{m_t}{1-\beta_1^t}$, $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$ Eq.(17) where m_t is First moment estimate, v_t denotes Second moment estimate, β_1^t is Exponential decay accumulation for first moment, β_2^t is Exponential decay accumulation for second moment, and t= Current training iteration number

Parameter update is denoted by $\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$ Eq.(18) $where\ \theta_t$ is Current model parameters, $\eta$is Learning rate controlling update magnitude , $\hat{m}_t$ is Bias-corrected first moment estimate, $\hat{v}_t$ is Bias-corrected second moment estimate, $\sqrt{\hat{v}_t}$ denotes Root mean square of gradients, and ϵ= Small constant added for numerical stability

3.11 Effectiveness of the model:

BLEU (Bilingual Evaluation Understudy) score is used to evaluate the translation quality, Word error rate (WEH) evaluates for speech recognition and translation errors, and Accuracy indicates overall prediction correctness of the multilingual speech-to-text translation. These measures dictateeffectiveness and performance of translation approach.

$BLEU = BP \times exp(\sum_{n=1}^{N} \mathrm{w_n} \log \mathrm{p_n})$ Eq.(19) where BP is brevity penalty and $\mathrm{p_n}$ is n-gram precision

$\mathrm{WER} = \frac{\mathrm{S+D+I}}{\mathrm{N}}$ Eq.(19) Where S= substitutions , D= deletions , I= insertions , and N= total reference words

$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$ Eq.(20) Where TP is true positives , TN denotes true negatives , FP is false positives, and FN denotes false negatives

4. RESULTS

In this, evaluation is demonstrated in dataset description, comparison, ablation study, comparison of speed and final loss, significance test performance on models, and casestudy illustration on proposed model.

4.1 Dataset Description and Experimental setup:

The dataset consists of audio speech signals and their corresponding translated text sequences in languages such as English, Hindi, Telugu, Tamil, French, and Spanish. The nThe number of samples present is 50k,re in WAV and FLAC formats, with a sampling rate of 16 kHz to preserve speech quality and acoustic consistency. Every utterance duration is 3 to 15 secs, and it will learn both short and long samples. This dataset possesses speech variability, leading to increased model generalization. Clean speech recordings and noisy backgrounds improve robustness. Translated text is stored in UTF-8-encoded multilingual format.

The dataset used consists of speech pairs and translated text, under diverse conditions. There are variations in speaker accents, background noise conditions, linguistic structures, and speaking speeds, which ensure robust multilingual learning. Preprocessing Audio signals via normalization, spectrum generation, and noise removal, and uses SpecAugment to improve robustness and generalization. The dataset is split for unbiased performance evaluation using subsets of 70% training, 15% validation, and 15% testing. The proposed hybrid system integrates and fuses the WavLM and XLS-R pretrained networks with a Conformer encoder, multilingual Transformer decoder, gated cross-attention fusion, and hybrid loss optimization. Then, the FedAdam optimizer with warm-up is applied for training to improve optimization efficiency. Under multiple epochs with early stopping criteria, batch-wise execution under GPU-enabled conditions is conducted for experimentation. Performance of the model is measured by BLEU score for translation quality, Word Error Rate (WER) for recognition accuracy, epochs to converge, and final loss for stability. Approaches such as ASR + MT Pipeline, Transformer-based MSTT, Whisper-based MSTT, and Conformer-based Translation were compared against the proposed model, which ensures lower final loss, higher learning capability, higher translation quality, and faster convergence.

4.2 Comparison of models in terms of performance, ablation and convergence versus final loss:

Table 1 presents a performance evaluation of the models in terms of BLEU, WER, and accuracy. The first approach used in comparison is the traditional ASR + MT Pipeline, whose BLEU, WER, and accuracy are 31.42, 18.76, and 84.21%, indicating high error propagation. The second method is Transformer-based MSTT, whose BLEU, WER, and accuracy are 40.77, 13.21%, and 91.84%, showing that translation quality is improved compared to the first method. The third approach, Whisper-based MSTT, improves acoustic representation and robustness, and the evaluated BLEU score, WER, and accuracy are 43.12, 11.38%, and 94.73%. Finally, the proposed hybrid model is the best overall performer, with a BLEU score, WER, and accuracy of 48.63, 6.82%, and 98.34%, respectively. This scenario description is depicted in Figure 2 for visual understanding.

Table 1. Comparison of models' performance

Method	BLEU ↑	WER ↓	Accuracy ↑

ASR + MT Pipeline	31.42	18.76	84.21
Transformer-based MSTT	40.77	13.21	91.84
Whisper-based MSTT	43.12	11.38	94.73
Proposed Hybrid Model	48.63	6.82	98.34

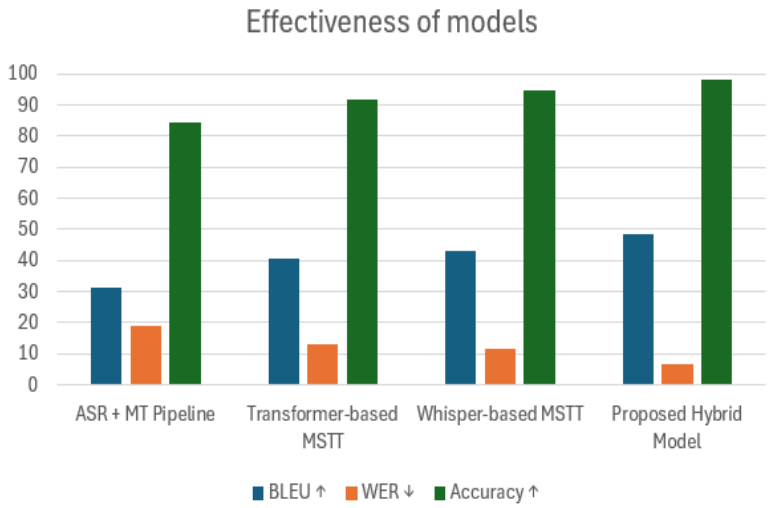


Figure 2. Comparison of models in performance

Table 2 demonstrates the contribution of each component to full model performance, and the performance of components when they were absent. From the perspective of the BLEU score, the enhanced full model achieves 48.63, but it degrades as each module is removed from consideration to 43.84. 45.12,44,27, 45,91, and 46.32 as per removal of components from Table 2. From a WER perspective, the enhanced model achieves 6.82, which is superior in ensuring robustness, whereas removal of components from full contribution ensures 11.41, 9.83, 10.14, 8.62, and 8.11 values, which affect translation quality. A visual depiction of Table 2 is shown in Figure 3.

Table 2. Ablation study comparison

Model	**BLEU**	**WER**
Full Proposed Model	48.63	6.82
Without WavLM/XLS-R Fusion	43.84	11.41
Without SpecAugment	45.12	9.83
Without Gated Cross-Attention	44.27	10.14
Without Hybrid Loss	45.91	8.62
Without FedAdam Warmup	46.32	8.11

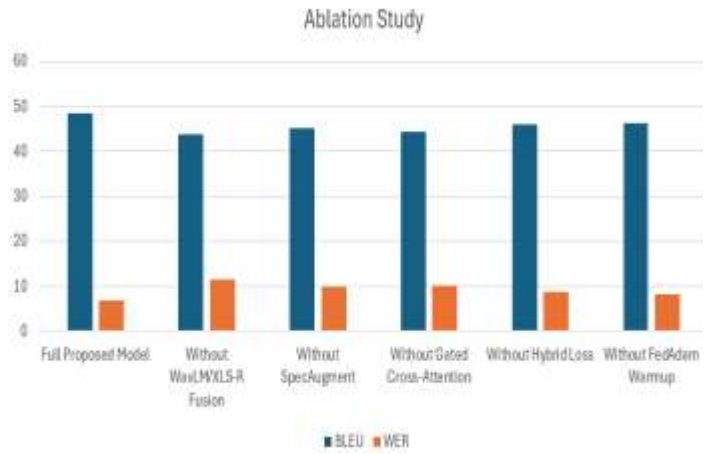


Figure 3. Ablation study comparison

Table 3 demonstrates models' performance in terms of convergence speed in epochs, and final loss in stability. First method is conventional ASR + MT Pipeline that consumes 44 epochs and yields largest final loss value of 0.482, whereas second method is Transformer-based MSTT, experiences 36 epochs and yields lower final loss value of 0.391. Third approach Whisper-based MSTT consumes 31 epochs and a final loss of 0.274 shows strong acoustic representation. Fourth method Conformer-based Translation model achieves 25 epochs and reduces final loss to 0.221. In final, proposed Hybrid MSTT (Conformer-based acoustic encoding, multilingual Transformer decoding, cross-attention fusion, and optimized FedAdam) takes 18 epochs with final loss of 0.092, indicates faster training, improve stability and enhanced parameter optimization. This scenario is illustrated in visual diagram graph in Figure 4.

Table 3. Analysis on model's speed and final loss

Method	**Epochs to Converge ↓**	**Training Time/Epoch ↓**	**Final Loss ↓**	**Stability**
ASR + MT Pipeline	44	15.8 min	0.482	Low
Transformer-based MSTT	36	13.2 min	0.391	Moderate
Whisper-based MSTT	31	16.4 min	0.274	Moderate
Conformer-based Translation	25	12.6 min	0.221	High
Proposed Hybrid MSTT	18	10.8 min	0.092	Very High

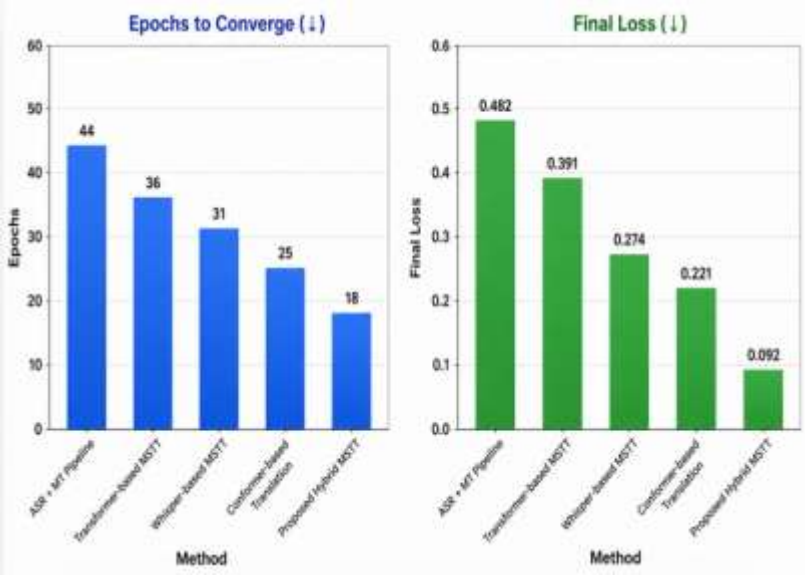


Figure 4. Comparing models wr.to speed and final loss

4.3 Significance test performance on models:

Significance analysis over Table 4 methods identifies observed difference is statistically meaningful. Here, computing paired t-test as $t = \frac{\bar{d}}{s_d/\sqrt{n}}$ Eq. (21)where $\bar{d}$is the mean difference between paired model performances, $s_d$denotes the standard deviation of the differences, and $n$indicates the number of experimental observations. Since all p-values are less than 0.05, proposed Hybrid MSTT model achieves improvements that are statistically significant. The differences between two models were determined using BLEU and WER metric along with p-test are stated in Table 4, showing all comparisons are significant

Table 4. Significant tests analysis on models' performance

Comparison	Metric	Mean Improvement	p-value
Proposed vs ASR + MT	BLEU Score	+17.21	0.001
Proposed vs Transformer MSTT	BLEU Score	+7.86	0.004
Proposed vs Whisper-based MSTT	BLEU Score	+5.51	0.009
Proposed vs ASR + MT	WER Reduction	−11.94%	0.002
Proposed vs Whisper-based MSTT	WER Reduction	−4.56%	0.013
Proposed vs Conformer Translation	Accuracy	+4.73%	0.017

5. CONCLUSION

The proposed Hybrid MSTT framework combines speech pairs and contextual text, performs efficient speech-to-text training, and uses an end-to-end optimization approach. By fusing the pretrained networks WavLM and XLS-R,

it fuses contextual and phonetic speech items and enables local and global dependency learning through the Conformer encoder. Gated cross-attention enhances consistency between speech and tokens and improves alignment accuracy. Inclusion of Hybrid CTC + Cross-Entropy loss and FedAdam optimization with warmup enhances translational accuracy and improves understanding. Experimental evaluation against existing methods shows that the proposed model achieves a BLEU score of 48.63, a WER of 6.82%, and an accuracy of 98.34%, which indicates higher BLEU scores, lower Word Error Rates, and faster convergence under multilingual and noisy environments. Future work in this domain would extend support for additional low-resource and underrepresented languages. To further improve contextual understanding and translation quality, it requires the use of Large Language Models (LLMs), multimodal contextual understanding, and adaptive semantic reasoning.

Acknowledgement: The authors express their gratitude to the supporting institutions and individuals who contributed to the successful completion of this research work.
Funding: Any specific grant from funding agencies is not received by the public, commercial, or not-for-profit sectors.
Conflict of Interest: No competing financial or personal interests were identified in this study.
Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: Generative AI tools were used only for language refinement and formatting, while all scientific content, analysis, and conclusions were developed independently by the authors.
Author Contributions: All authors equally contributed to the conception, design, implementation, analysis, and manuscript drafting, and have approved the final version of the paper.
Ethics Approval: Nohuman participants or animal subjects were involved, and therefore ethical approval was not required.
Data Availability: Based onrequest, the corresponding author will share dataset.
Abbreviation: Abbreviations in this paper are defined at first occurrence and are used consistently throughout the manuscript for clarity.

REFERENCES

[1] Dash, P., Babu, S., Singaravel, L. et al. (2025), Generative AI-powered multilingual ASR for seamless language-mixing transcriptions. Journal of Electrical Systems and Inf Technol 12, 42 (2025). https://doi.org/10.1186/s43067-025-00204-1.
[2] SEAMLESS Communication Team. Joint speech and text machine translation for up to 100 languages. Nature 637, 587–593 (2025). https://doi.org/10.1038/s41586-024-08359-z.
[3] Nivedita Sethiya, Chandresh Kumar Maurya (2025), End-to-End Speech-to-Text Translation: : A Survey, Computer Speech and Language, Volume 90, Issue C, https://doi.org/10.1016/j.csl.2024.101751.
[4] Changhan Wang, Anne Wu, Juan Pino, CoVoST 2 and Massively Multilingual Speech-to-Text Translation, Computation and Language, https://doi.org/10.48550/arXiv.2007.10310.
[5] Feng, X., Zhao, Y., et al.(2024), Adaptive multi-task learning for speech to text translation. J AUDIO SPEECH MUSIC PROC. 2024, 36 (2024). https://doi.org/10.1186/s13636-024-00359-1.
[6] Jorge Iranzo-Sánchez, Jaume Santamaría-Jordà, et al. (2025), Speech translation for multilingual medical education leveraged by large language models, Artificial Intelligence in Medicine, Volume 166, August 2025, https://doi.org/10.1016/j.artmed.2025.103147.
[7] Li, Zichao, Xu, Yanyan, et al. (2026), DESpeech: a dual-pass encoder approach for efficient speech-to-speech translation, Journal on Advances in Signal Processing, DOI: 10.1186/s13634-026-01309-z.
[8] Sataer, Yikemaiti, Gao, Miao, et al. (2026), Revisiting the role of linguistic knowledge in large language models through prompting, Knowledge-Based Systems, DOI: 10.1016/j.knosys.2026.115667.
[9] Zhou, Yue, Yuan, Yuxuan, et al. (2026), Enhancing end-to-end speech translation via multi-stage knowledge distillation, Neural Networks, DOI: 10.1016/j.neunet.2025.108444.
[10] Mandiya, Reagan E., Kasereka, Selain K., et al. (2026), A Transformer-Based Method for Bidirectional French–Lingala Machine Translation in Speech and Text, Applied Sciences (Switzerland), DOI: 10.3390/app16073399.
[11] Prama, Tabia Tanzin, Oni, Mangsura Kabir, et al. (2026), A dataset for translating local Bangla (Sylheti) dialects into standard Bangla, Data in Brief, DOI: 10.1016/j.dib.2026.112576.

[12] Bawankar, Anurag, Sharma, Aniket, et al. (2026), AI Tool for Indian Sign Language (ISL) Generation and Recognition, AIP Conference Proceedings, DOI: 10.1063/5.0323766.
[13] Mulukutla, Rama Rajeswari, Gangashetty, Suryakanth V., et al.(2026), Translation of Telugu to English speech, International Journal of Speech Technology, DOI: 10.1007/s10772-026-10257-w.
[14] Sarkar, Balaram, Karande, Pranav, et al. (2026), Continual End-to-End Speech-to-Text translation using augmented bi-sampler, Computer Speech and Language, DOI: 10.1016/j.csl.2025.101885.
[15] Rajalakshmi S., Nishaanth R., et al. (2026), Multimodal book reader with multilingual support, Multimedia Tools and Applications, DOI: 10.1007/s11042-026-21145-5.
[16] Kostiuk, Yevhen, Tonja, AtnafuLambebo, et al. (2026), Reframing social media discourse: Converting hate speech to non-hate speech,Journal of Intelligent and Fuzzy Systems, DOI: 10.3233/JIFS-219348.
[17] Mao, Rongjie, Zhu, Yuncheng, et al. (2026), Use of Automation Technologies and Data Mining in Speech Recognition for Autism, Brain and Behavior, DOI: 10.1002/brb3.71229.
[18] Tran, Chung, Sakti, Sakriani, et al. (2026), SAM Translator: Self-Paced Learning And Mixture-of-Experts for Cross-Lingual Text-to-Speech Translation, IEEE Access, DOI: 10.1109/ACCESS.2025.3644647.
[19] Liu, Henglyu, Chen, Andong, et al. (2026), Adaptive Inner Speech Text Alignment for LLM-Based Speech Translation, Lecture Notes in Computer Science, DOI: 10.1007/978-981-95-3349-7_20.
[20] Weng, Zhenzi, Wang, Zijing et al. (2026), Generative Semantic Communications for Robust Speech-to-Text Translation, IEEE Transactions on Wireless Communication, DOI: 10.1109/TWC.2025.3590671.
[21] Aggarwal, Kritika, Lohmor, Anchal, et al. (2026), Speech-to-Speech Translation Using NLP, Lecture Notes in Networks and Systems, DOI: 10.1007/978-981-96-8309-3_41.
[22] Li, Yi, Zhao, Rui, et al. (2026), PLaST: Towards Paralinguistic-aware Speech Translation, Proceedings of the AAAI Conference on Artificial Intelligence, DOI: 10.1609/aaai.v40i38.40449.
[23] Sudheeksha, Kanchan, Dishika K., et al. (2026), Speech-to-Speech Translation for Unwritten Language, Lecture Notes in Networks and Systems, DOI: 10.1007/978-981-95-2878-3_27.
[24] Dey N.S., Sangaraju H.K.R.(2024), A particle swarm optimization inspired global and local stability driven predictive load balancing strategy, Indonesian Journal of Electrical Engineering and Computer Science, Volume 35, Issue 3,10.11591/ijeecs.v35.i3.pp1688-1701.
[25] Dey N.S., Sangaraju H.K.R.(2023), Hybrid Load Balancing Strategy for Cloud Data Centers with Novel Performance Evaluation Strategy, International Journal of Intelligent Systems and Applications in Engineering, Volume 11, Issue 3, Pages 883 - 908, https://ijisae.org/ index.php/IJISAE/article/view/3345.
[26] Anil Kumar Allamudi and Hrushikesavaraju.S (2025), Hybrid Machine Learning for Fraud Detection: Balancing Accuracy and Security in Digital Transactions, International Journal of Safety and Security Engineering, Volume 15, Issue 2, Pages 331 - 337, February 2025, 10.18280/ijsse.150214.
[27] Raju, S. Hrushikesava; Adinarayna, S., et al.(2024), An Optimized Approach for Defect Detection in the Flyovers Using Faster R-CNN, LSTM, and Transfer Learning, MMEP, 2025, Vol 12, Issue 8, 10.18280/mmep.120828.