Enhancing Machine Translation From Multimodal Punctuation Prediction

Pranav Satheesan
pranavsatheesan2001@gmail.com

Sarang Galada
sarang.galada@gmail.com

Abstract

Machine Translation (MT) systems depend on well-structured input, including proper punctuation, to produce coherent and semantically accurate translations. However, in speech-to-speech translation pipelines, automatic speech recognition (ASR) often outputs unpunctuated text, leading to degraded MT performance. Traditional punctuation restoration methods operate as a post-processing step relying solely on textual features, overlooking prosodic cues inherent in the speech signal.

In this work, we build a multimodal punctuation restoration model that jointly leverages acoustic and lexical information to more accurately predict punctuation. We then extract learned multimodal punctuation embeddings from this model and integrate them into a machine translation system. Our approach explores whether infusing these punctuation-aware representations during translation decoding can improve translation quality.

1 Introduction

In speech-to-speech machine translation (S2S MT), the quality of translation is closely tied to the structure of the input text—particularly its punctuation. Punctuation marks serve as important cues for syntactic boundaries, discourse segmentation, and sentence-level semantics, all of which directly affect translation fluency and accuracy. However, automatic speech recognition (ASR) systems, which serve as the front-end of most S2S MT pipelines, typically produce raw, unpunctuated transcripts. This loss of structure often propagates errors downstream, resulting in degraded machine translation performance.

To mitigate this, current pipelines commonly employ *punctuation restoration* as a separate post-processing step after ASR decoding. These models rely purely on textual features and operate under the assumption that the lexical content alone suffices to infer punctuation. This assumption, however, ignores *prosodic information*—such as pauses, pitch changes, and timing—which are rich in signals related to sentence boundaries and speaker intent. As a result, existing punctuation models may underperform in ambiguous or low-resource contexts where textual cues are insufficient.

This project is driven by two hypotheses:

1. **Multimodal Speech plus Text Punctuation:** Incorporating acoustic (speech) features alongside textual features can improve punctuation restoration performance by capturing prosodic signals absent from text.(Sunkara et al., 2020)

2. **Machine Translation Enhancement:** Learned punctuation representations, when integrated into the MT model during decoding, can improve translation quality by injecting structural and semantic cues into the translation process.

To test these hypotheses, we follow a three-stage pipeline:

- First, we train a *multimodal punctuation restoration model* that fuses aligned acoustic and lexical inputs to predict punctuation.

- Second, we extract *punctuation embeddings* from this model that encode contextual cues from both modalities.

- Finally, we train an end-to-end MT system that incorporates these embeddings during decoding, aiming to enhance translation performance with richer source-side structure.

2 Methodology

Our multimodal punctuation restoration system leverages both acoustic and lexical information to

accurately predict punctuation marks in unpunctuated text. The core architecture comprises an acoustic encoder, a lexical encoder, a forced alignment module, a multimodal feature fusion mechanism, and a prediction head. The overall process is designed to seamlessly integrate with a machine translation (MT) system to improve its performance.

2.1 Acoustic Encoder

To capture rich phonetic and prosodic information from the audio, we employ a pre-trained Wav2Vec2.0 model as our acoustic encoder. Wav2Vec2.0(Baevski et al., 2020), trained on a vast corpus of speech data, has demonstrated remarkable capabilities in learning robust speech representations.

Feature Extraction: The pre-trained Wav2Vec2.0 model extracts 768-dimensional frame-level acoustic features. These features are extracted at a resolution of one feature vector every 20 milliseconds, providing a dense representation of the speech signal over time.

Dimensionality Reduction and Contextualization: To prepare these high-dimensional features for subsequent processing and to capture local temporal dependencies, the 768-dimensional features are first passed through a 1D convolutional layer. This layer reduces the dimensionality from 768 to 512, using a kernel size of 5 and padding of 2 to preserve the input sequence length. The output of the convolutional layer is then fed into a unidirectional Long Short-Term Memory (LSTM) network. The LSTM, with an output dimension of 1024, processes the sequential acoustic features, enabling it to learn long-range dependencies and generate context-aware acoustic representations for each frame. The unidirectional nature is chosen to mimic the natural flow of speech.

2.2 Lexical Encoder

The lexical encoder is responsible for generating context-aware representations of the unpunctuated text.

Tokenization: Unpunctuated text input is tokenized using the IndicTrans2 (Gala et al., 2023) tokenizer. IndicTrans2 is a transformer-based model specifically designed for Indian languages, ensuring effective handling of their unique linguistic characteristics, including rich morphology and diverse scripts.

Contextual Embedding: The tokenized text is then passed through the encoder component of the IndicTrans2 model. This encoder, typically a multi-layer transformer encoder, processes the sequence of tokens, generating context-aware lexical representations. These representations capture the semantic and syntactic relationships between words, crucial for accurate punctuation prediction.

2.3 Forced Alignment

To establish a precise correspondence between the acoustic and lexical features, we perform forced alignment of the audio and text.

Word-Level Timestamp Extraction: We utilize OpenAI Whisper (Radford et al., 2022) for forced alignment. Whisper, a robust automatic speech recognition (ASR) model, can extract highly accurate word-level timestamps when run with verbose=True. This verbose mode provides detailed segment and word-level timing information, allowing us to accurately map spoken words to their corresponding acoustic segments in the audio. This alignment is critical for synchronizing the features from the acoustic and lexical encoders, enabling a truly multimodal approach.

2.4 Multimodal Punctuation Restoration

The aligned speech and text features are then fused to create comprehensive multimodal representations, which are subsequently used by a prediction head to classify punctuation marks.

Feature Alignment: The word-level timestamps obtained from forced alignment are used to align the frame-level acoustic features with their corresponding word tokens. For each word, the acoustic features within its start and end timestamps are aggregated (e.g., by averaging or pooling) to form a single word-level acoustic representation.

Feature Fusion: Two distinct methods are explored for fusing the aligned word-level acoustic and lexical features:

Norm and Add: In this method, the acoustic and lexical feature vectors are first normalized (e.g., using L2 normalization) to bring them to a comparable scale. Subsequently, the normalized vectors are element-wise added. This approach assumes that both modalities contribute additively to the overall representation and aims to combine their information while mitigating scale differences.

Concat and Project: This method involves concatenating the acoustic and lexical feature vectors. The concatenated vector is then projected into a lower-dimensional space using a learned linear transformation (e.g., a fully connected layer). This

approach allows the model to learn complex interactions between the modalities and potentially discover more effective combined representations.

Prediction Head: The fused multimodal representations are fed into a prediction head, which is a fully connected linear layer with softmax activation function. This head classifies the punctuation mark (e.g., comma, full stop, question mark, none) to be inserted after each word. The system is trained to predict the presence or absence of punctuation and the specific type of punctuation at each word boundary.

2.5 Integration into Machine Translation

The learned punctuation embeddings are designed to enhance the performance of a downstream machine translation system.

Feature Enhancement: The output of our punctuation restoration module, specifically the rich multimodal representations (before the final classification), can be considered as "punctuation embeddings." These embeddings, encoding information about the likely presence and type of punctuation, are then fused with the encoder outputs of the IndicTrans2 model. This fusion can be achieved through various mechanisms, such as concatenation followed by another projection layer, or through an attention mechanism where the punctuation embeddings guide the attention process within the MT encoder. By providing explicit punctuation cues to the MT model's encoder, we hypothesize that the translation quality will be significantly improved, particularly in terms of fluency, naturalness, and preservation of sentence structure in the target language.

3 Workflow

The overall pipeline for punctuation-aware machine translation is illustrated in Figure **??**. The system operates in the following stages:

Audio Input and ASR Decoding: The source audio is first passed through an automatic speech recognition (ASR) model (e.g., Whisper) to produce unpunctuated text. Multimodal Punctuation Restoration: The ASR-generated unpunctuated text and the original source audio are then input to the Multimodal Punctuation Restoration Model. This model fuses acoustic features (from audio) and lexical features (from text) to predict appropriate punctuation. The output is a set of learned punctuation features. Machine Translation with Fusion: These punctuation features are then integrated with the encoder representations of the source text in a Machine Translation (MT) model (e.g., IndicTrans2). This enriched input helps the decoder generate more fluent and accurate translated text. This end-to-end approach leverages both lexical and prosodic cues to enhance translation quality, especially in noisy or low-resource speech settings.

4 Experiments:

1. **Concat and project:** In this fusion approach, lexical features (1024-d) from the text encoder and acoustic features (1024-d) from the speech encoder are concatenated to form a 2048-dimensional combined representation. This high-dimensional vector is then projected back into a 1024-dimensional space to form the final multimodal representation. This strategy preserves raw feature details from both modalities while allowing the model to learn shared latent patterns.

Evaluation Metrics	
Accuracy	0.9394
Precision (w)	0.9344
Recall (w)	0.9394
F1 Score (w)	0.9366
AUC ROC	Undefined (check class coverage)

Table 1: Evaluation metrics for the classification model.

	No punct	**Full stop**	**Comma**	**Semi Colon**
No punct	5569	21	107	3
Full stop	41	289	19	0
Comma	160	27	123	2
Semi Colon	5	0	1	0

Table 2: Confusion matrix for the classification model.

	Precision	**Recall**	**F1-Score**
No punct	0.96	0.98	0.97
Full stop	0.86	0.83	0.84
Comma	0.49	0.39	0.44
Semi Colon	0.00	0.00	0.00
Accuracy		0.94	
Macro avg	0.58	0.55	0.56
Weighted avg	0.93	0.94	0.94

Table 3: Classification report showing precision, recall, and F1-score for each class.

2. **Concat and bottleneck projection:** Here, lexical and acoustic features (each 1024-d)

are concatenated into a 2048-dimensional representation, which is then passed through a feed-forward projection layer to reduce dimensionality to 512-d. A final transformation projects this to a 1024-d multimodal representation. This introduces non-linearity and compression, helping the model learn more abstract, informative features.

Evaluation Metrics	
Accuracy	0.9398
Precision (w)	0.9404
Recall (w)	0.9398
F1 Score (w)	0.9401
AUC ROC	Undefined (check class coverage)

Table 4: Evaluation metrics for the classification model.

	No punct	Full stop	Comma	Semi Colon
No punct	5524	17	159	0
Full stop	39	299	11	0
Comma	134	17	161	0
Semi Colon	3	0	3	0

Table 5: Confusion matrix for the classification model.

	Precision	Recall	F1-Score
No punct	0.97	0.97	0.97
Full stop	0.90	0.86	0.88
Comma	0.48	0.52	0.50
Semi Colon	0.00	0.00	0.00
Accuracy		0.94	
Macro avg	0.59	0.59	0.59
Weighted avg	0.94	0.94	0.94

Table 6: Classification report showing precision, recall, and F1-score for each class.

3. **Norm and Add Fusion:** In this technique, both lexical and acoustic features are independently normalized and then summed elementwise to obtain the multimodal representation. This method enforces scale compatibility and avoids dimensional expansion, while allowing effective integration of complementary information from both modalities. The output is a 1024-dimensional feature vector aligned per word. However, this method gives relatively poorer performance.

Evaluation Metrics	
Accuracy	0.9120
Precision (w)	0.9095
Recall (w)	0.9120
F1 Score (w)	0.9007
AUC ROC	Undefined (check class coverage)

Table 7: Evaluation metrics for the classification model.

	No punct	Full stop	Comma	Semi Colon
No punct	5495	185	20	0
Full stop	61	285	3	0
Comma	190	95	27	0
Semi Colon	5	1	0	0

Table 8: Confusion matrix for the classification model.

	Precision	Recall	F1-Score
No punct	0.96	0.96	0.96
Full stop	0.50	0.82	0.62
Comma	0.54	0.09	0.15
Semi Colon	0.00	0.00	0.00
Accuracy		0.91	
Macro avg	0.50	0.47	0.43
Weighted avg	0.91	0.91	0.90

Table 9: Classification report showing precision, recall, and F1-score for each class.

5 Future Work:

The multimodal punctuation restoration module has been successfully developed, combining lexical and acoustic features to produce enriched punctuation-aware representations. The next phase of the project involves integrating these learned features into the machine translation (MT) pipeline. This step is critical for evaluating the actual impact of multimodal punctuation on downstream translation quality.

To achieve this, we plan to experiment with the following three training strategies for the MT model:

- Train only the last few decoder layers while keeping the rest of the model frozen. This allows the decoder to adapt to the enhanced input without disrupting the pretrained structure.
- Train only the decoder, using a frozen encoder to isolate the decoder's ability to leverage punctuation-informed features.

- Train the full encoder-decoder model end-to-end, enabling comprehensive optimization but requiring more compute and careful regularization.

These approaches will help us assess the best way to integrate punctuation-aware features into the translation model effectively. Future experiments will also explore multilingual extensions and evaluate the generalizability of this approach across language pairs.

Acknowledgments

The authors are thankful to the IASNLP 2025 organizers and team at IIIT Hyderabad and our mentors Harshita and Vennela.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Monica Sunkara, Srikanth Ronanki, Dhanush Bekal, Sravan Bodapati, and Katrin Kirchhoff. 2020. Multimodal semi-supervised learning framework for punctuation prediction in conversational speech. *Preprint*, arXiv:2008.00702.