Enhancing Machine Translation with Multimodal Punctuation Features

PraSang प्रसंग പ്രസംഗം

ప్రసంగం பிரசங்கம்

Pranav Satheesan, Sarang Galada

Mentors: Harshita, Vennala

Problem Statement

Machine Translation (MT) performs well when the input text is properly punctuated. In Speech to Speech pipelines, punctuation is typically restored through post-processing after ASR decoding. But these approaches rely on textual post-processing ignoring speech cues and prosodic information. This projects explores whether combining speech-based punctuation features with textual punctuation can improve MT quality.

Objectives:

1. Build and train a multimodal (speech + text) punctuation restoration model.
2. Extract punctuation features from the trained restoration model.
3. Train an End-2-End Machine Translation model which fuses the multimodal (textual + speech) punctuation features during MT.

Data

Fleurs: Parallel corpora of **Speech** (audio) + **Transcription** (text), in 102 languages.

We chose: Source language **English-US**, Target language **Hindi-India**

EN_US (speech-text):

- Train: 2602
- Val: 394
- Test: 647

HI_IN (speech-text):

- Train: 2120
- Val: 239
- Test: 418

Data

Sample:

A tornado is a spinning column of very low-pressure air, which sucks the surrounding air inward and upward.

बवंडर बहुत कम दबाव वाली हवा का एक घुमता हुवा स्तंभ होता है, जो आसपास की हवा को अंदर और ऊपर की तरफ खींचता है.

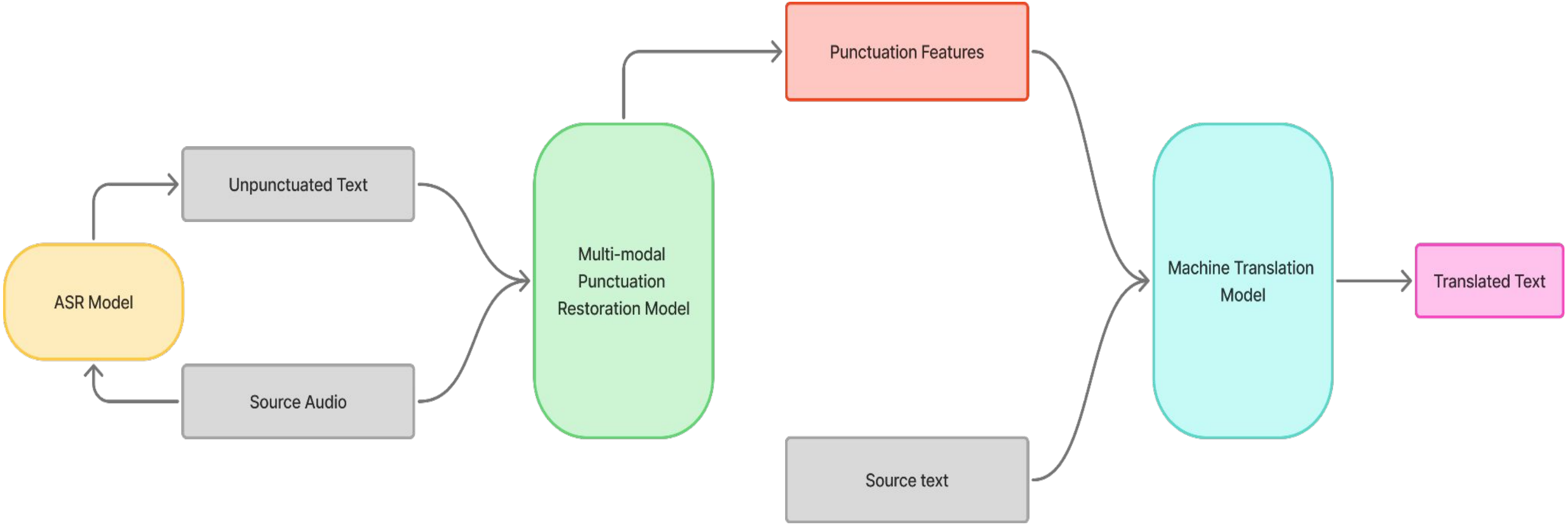
(Only a subset (eg. 1281 ids in training set) overlap for Eng-Hindi !)

Data

Punctuation stats after cleaning (training set):

Lang	**. or ।**	**,**	**-**	**:**	**;**	**?**	**!**
English	2538	2397	350	64	34	17	12
Hindi	2317	2014	414	61	28	19	7

Architecture Diagram



Methodology

- **Acoustic and Lexical feature extraction**:
  - Extract lexical features from unpunctuated text, using IndicTrans2 encoder
  - Extract acoustic features from audio, using Wav2Vec 2.0
  - For speech-text forced alignment, extract word timestamps from audio, using Whisper ASR

- **Multimodal fusion and Punctuation prediction**:
  Training a learnable Acoustic encoder (Conv + LSTM), Fusion network and Prediction head.

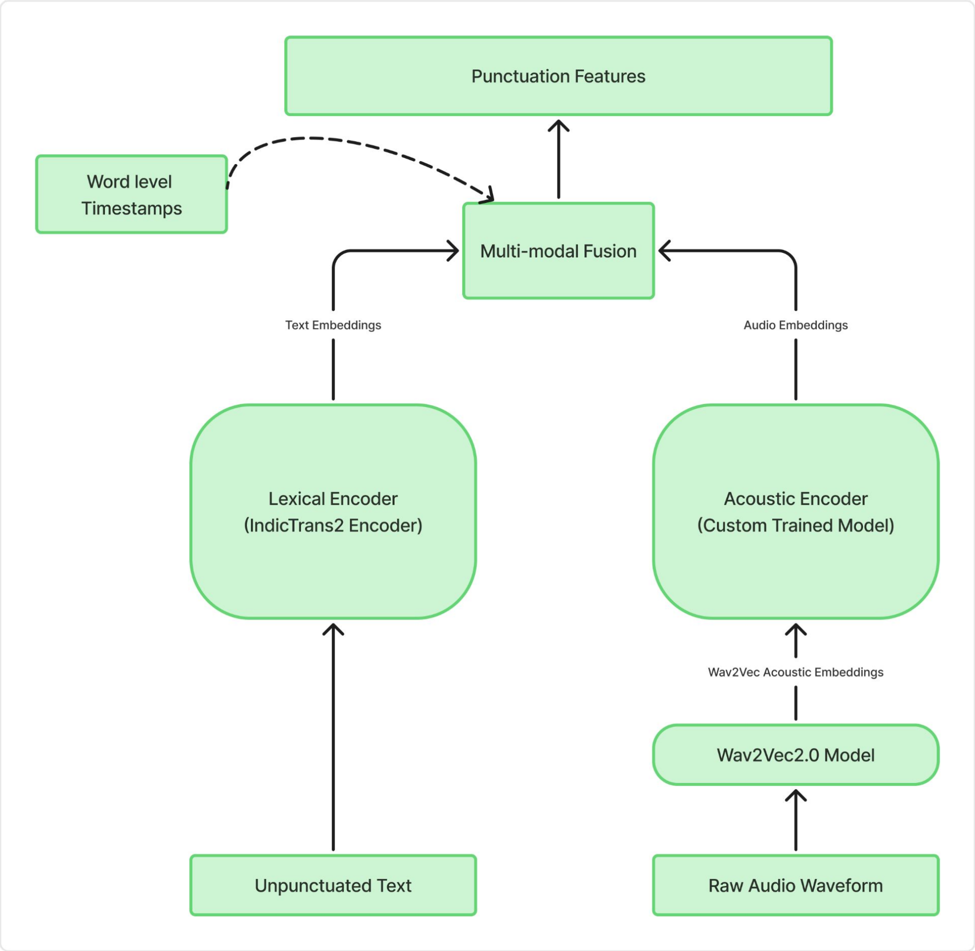
Methodology

★ **Punctuation Feature Fusion during MT**: Fusion of learned punctuation features into the latent encoder output of the MT model, followed by decoding into target language.

★ **Evaluation**: Measure improvements in translation quality using standard metrics (BLEU, METEOR), with and without multimodal features, and in comparison to the baseline MT.

Part 1: Punctuation Restoration Model

Multimodal Punctuation Restoration Model



Custom Architecture (Acoustic Encoder):

Input: Raw audio waveform

Step 1: Wav2Vec2.0 Encoder

★ Converts audio into frame-level representations

★ Outputs: **768-dimensional** features (one every 20ms)

Step 2: 1D Convolution Layer

★ **Input Dim**: 768 → **Output Dim**: 512

★ **Kernel Size**: 5, **Padding**: 2

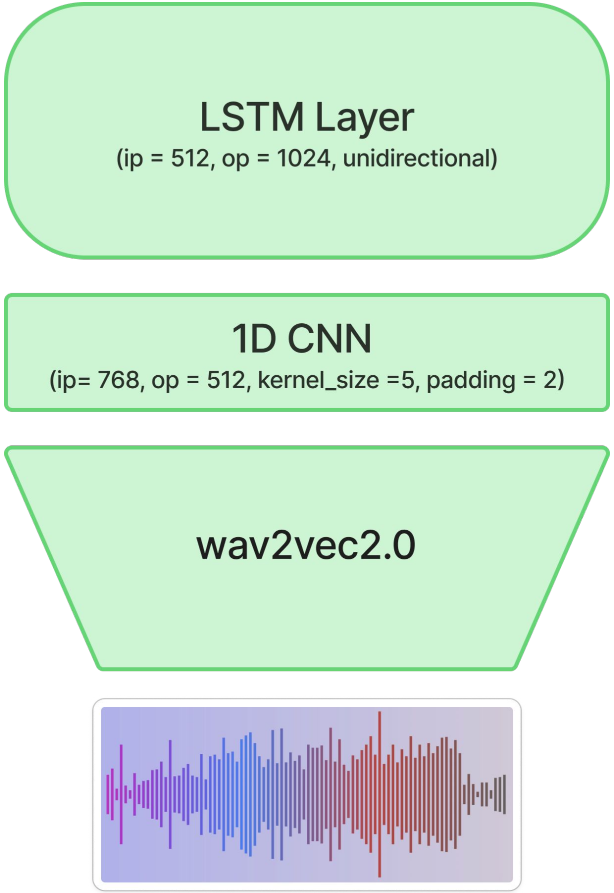
★ Captures **local temporal patterns** and smoothes features

Step 3: Unidirectional LSTM Layer

★ **Input Dim**: 512 → **Output Dim**: 1024

★ Models **long-range temporal dependencies**

★ Processes sequence **left-to-right only** (causal)



LSTM Layer

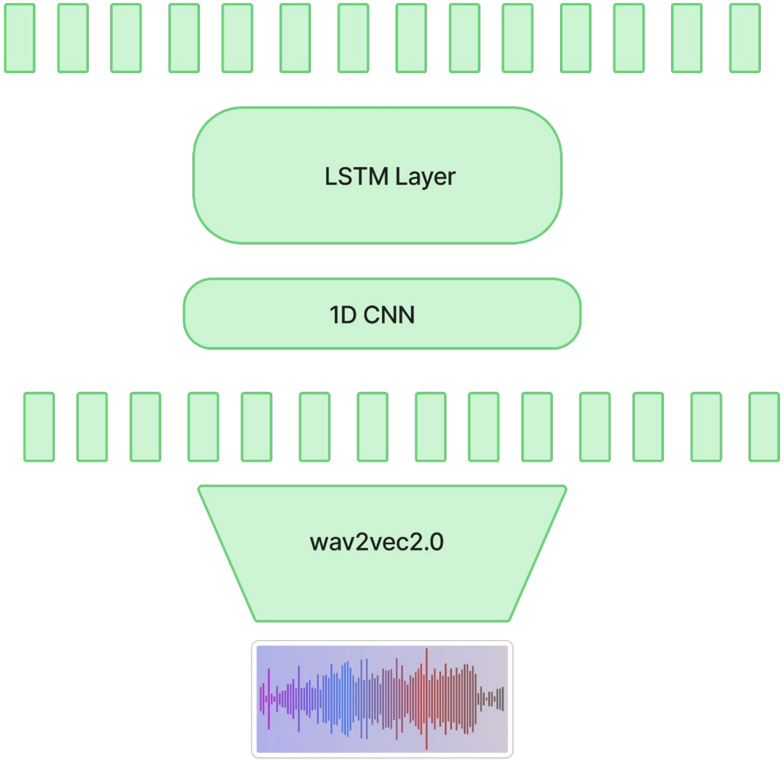
(ip = 512, op = 1024, unidirectional)

1D CNN

(ip= 768, op = 512, kernel_size =5, padding = 2)

wav2vec2.0

Custom Architecture (Acoustic Encoder)



Acoustic representations
shape: 150×1024

(ip = 512, op = 1024,
unidirectional)

(ip= 768, op = 512,
kernel_size =5, padding = 2)

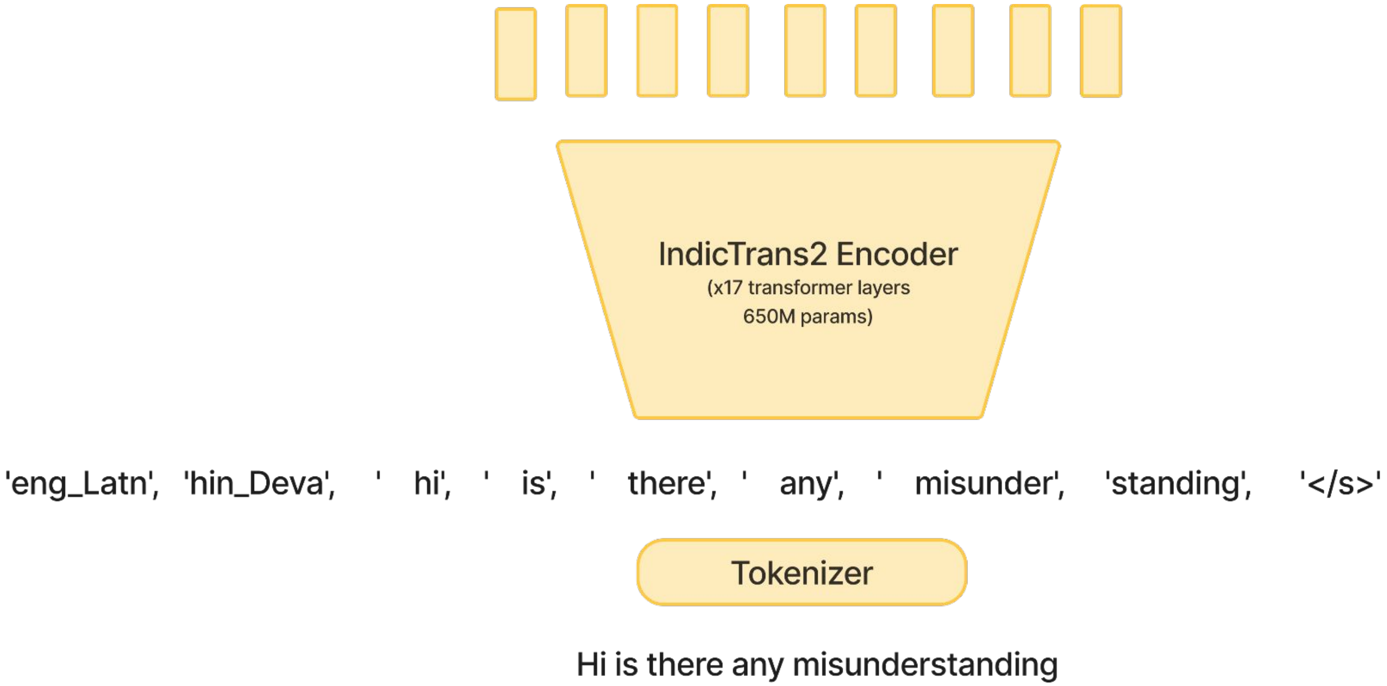
Wav2vec representations
shape: 150×768

(sampling = 16kHz
50 frames/ sec = 20ms per frame)

3 seconds

https://arxiv.org/pdf/2006.11477

IndicTrans2 Model (MT Lexical Encoder)



Lexical representations

shape: 9×1024

IndicTrans2 Encoder

(x17 transformer layers

650M params)

'eng_Latn', 'hin_Deva', ' hi', ' is', ' there', ' any', ' misunder', 'standing', '</s>'

9 tokens

Tokenizer

Hi is there any misunderstanding

5 words

https://arxiv.org/pdf/2305.16307

Forced Alignment

What is Forced Alignment?

★ A process that aligns spoken audio with its textual transcript at word or phoneme level.

★ Maps **when** each word or phoneme is spoken in an audio file.

Input: Audio file (.wav file) + (optionally) corresponding transcript

Output: Time-aligned annotations: word start & end timestamps.

Tools: Montreal Forced Aligner, OpenAI-Whisper (verbose = True)

ASR Model - OpenAI Whisper

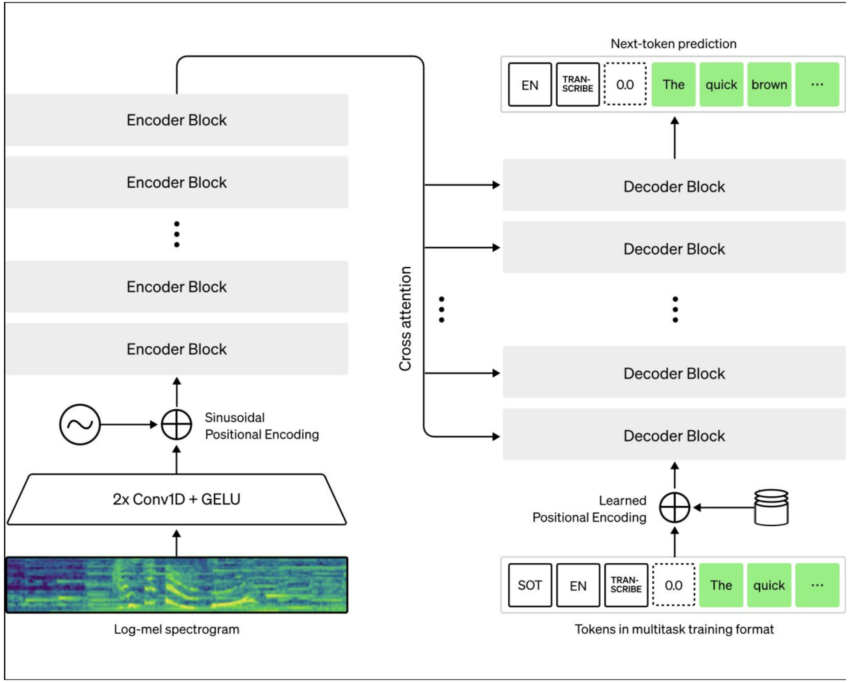
What is Whisper?
Whisper is an automatic speech recognition (ASR) system developed by OpenAI, designed to transcribe and translate spoken language with high accuracy and robustness.

Key Features:

- Multilingual and multitask: supports transcription and translation across many languages.
- Transformer-based encoder-decoder architecture.
- Trained on 680,000+ hours of diverse, supervised audio data.
- Robust to accents, background noise, and disfluencies.
- Open-source and widely adopted for speech-related tasks.

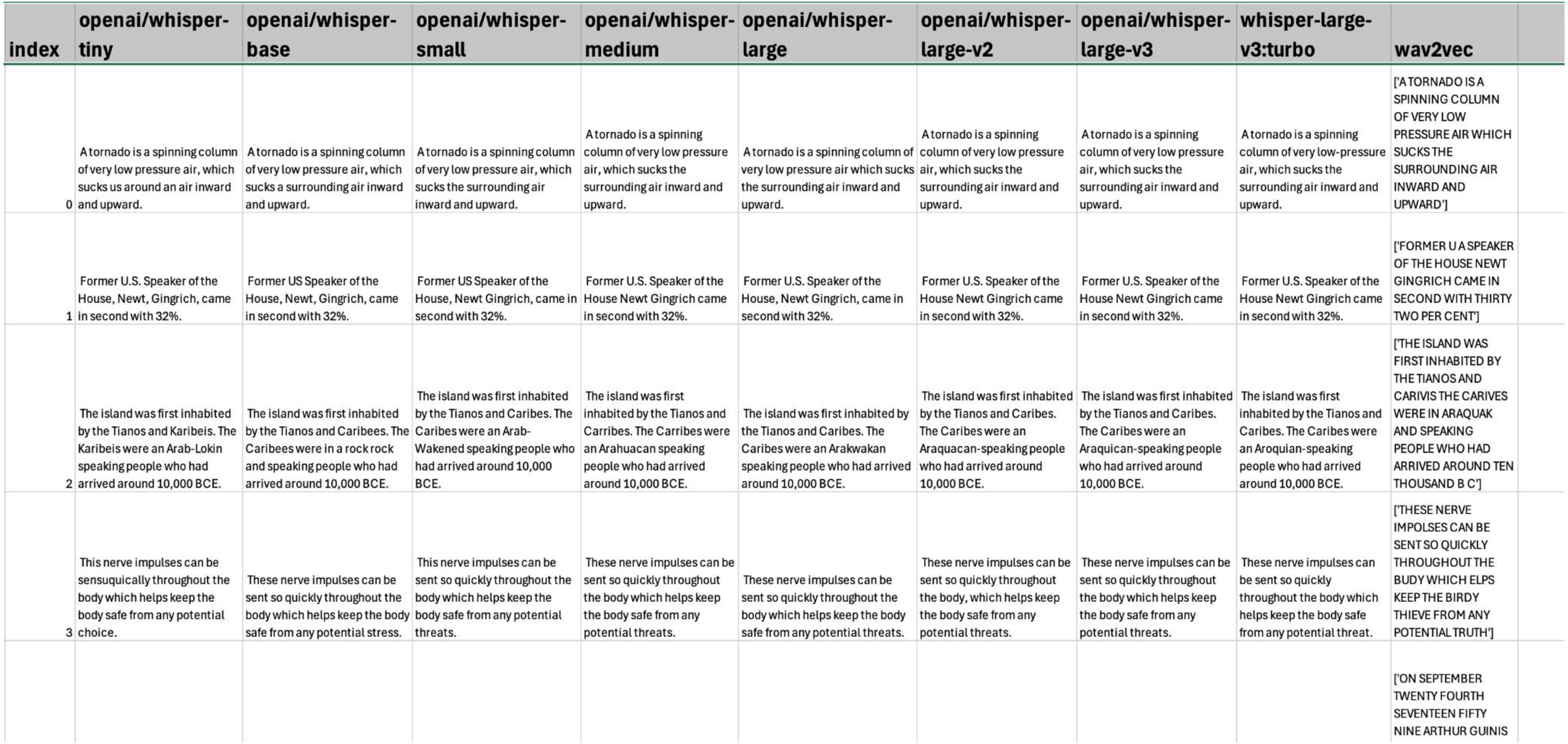
Relevance to This Project:

- Used to generate unpunctuated ASR transcripts.
- Provides realistic input for downstream punctuation restoration and machine translation models.
- Enables building an end-to-end speech-to-translation pipeline with minimal domain-specific tuning.



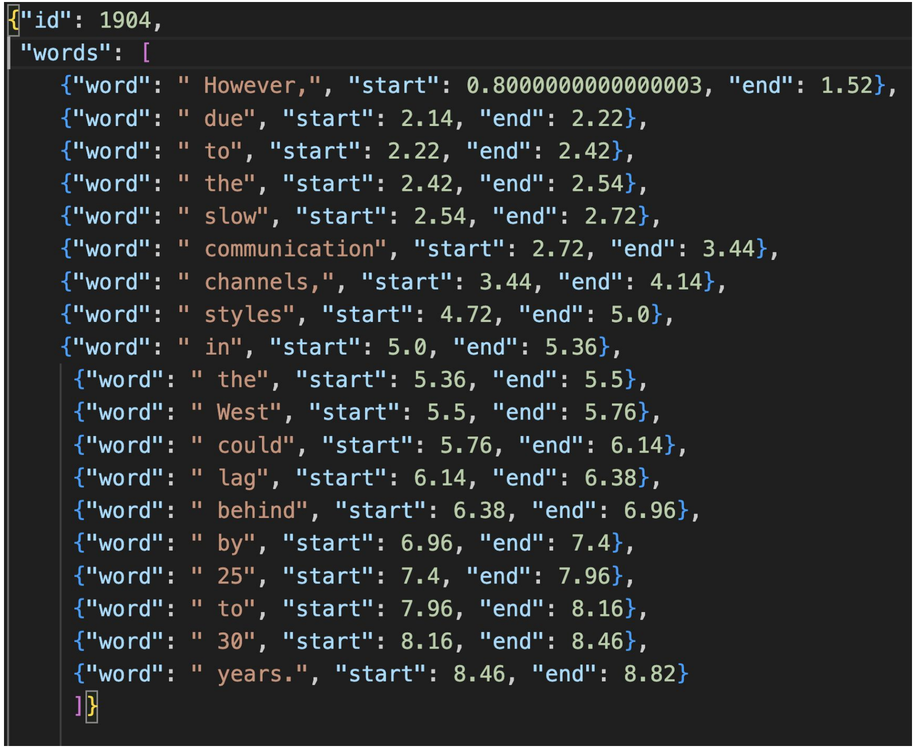
https://arxiv.org/pdf/2212.04356

Comparison of ASRs: Selection -> whisper medium



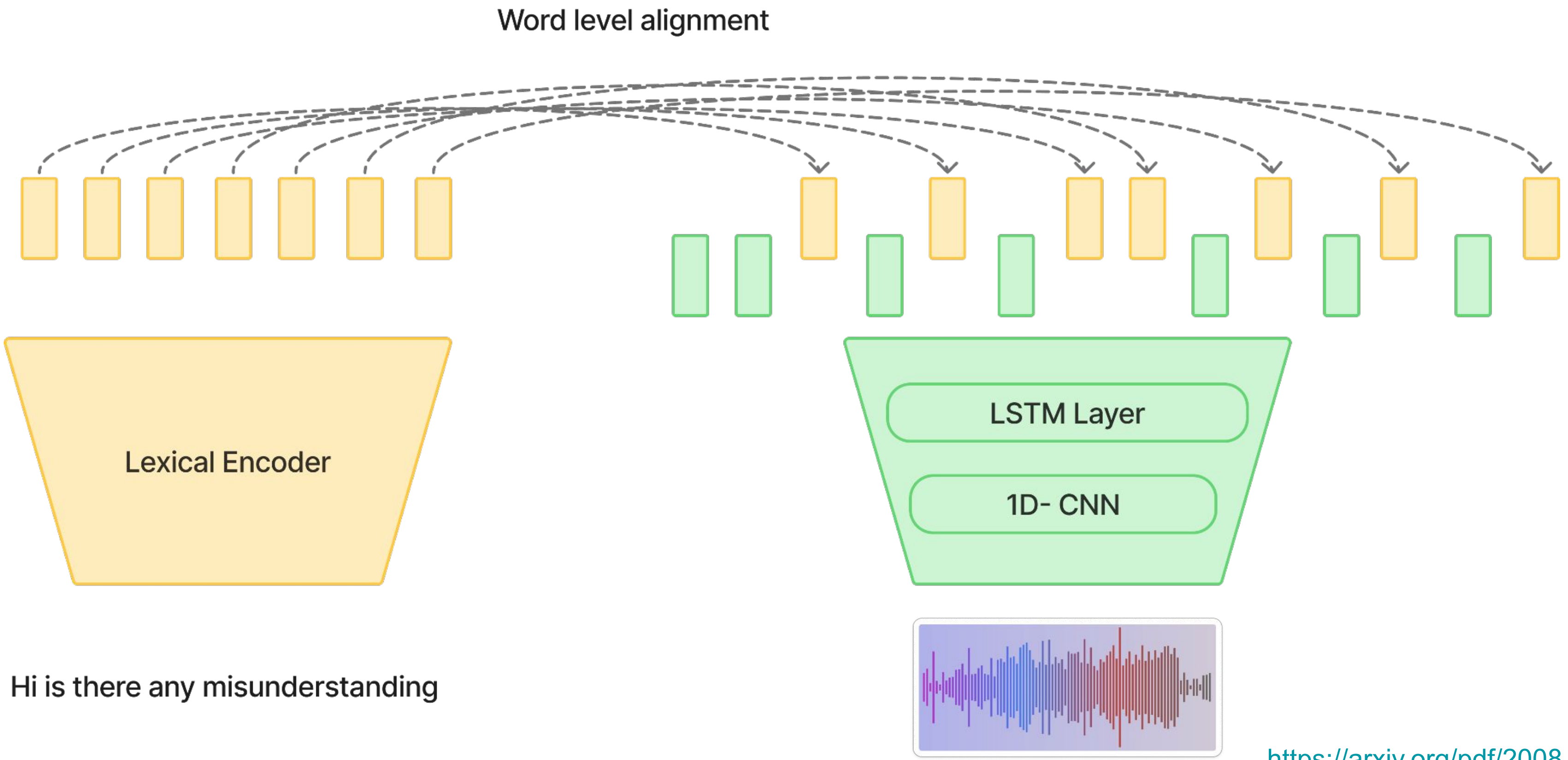
index	openai/whisper-tiny	openai/whisper-base	openai/whisper-small	openai/whisper-medium	openai/whisper-large	openai/whisper-large-v2	openai/whisper-large-v3	whisper-large-v3:turbo	wav2vec
0	A tornado is a spinning column of very low pressure air, which sucks us around an air inward and upward.	A tornado is a spinning column of very low pressure air, which sucks a surrounding air inward and upward.	A tornado is a spinning column of very low pressure air, which sucks the surrounding air inward and upward.	A tornado is a spinning column of very low pressure air, which sucks the surrounding air inward and upward.	A tornado is a spinning column of very low pressure air which sucks the surrounding air inward and upward.	A tornado is a spinning column of very low pressure air, which sucks the surrounding air inward and upward.	A tornado is a spinning column of very low pressure air, which sucks the surrounding air inward and upward.	A tornado is a spinning column of very low-pressure air, which sucks the surrounding air inward and upward.	['A TORNADO IS A SPINNING COLUMN OF VERY LOW PRESSURE AIR WHICH SUCKS THE SURROUNDING AIR INWARD AND UPWARD']
1	Former U.S. Speaker of the House, Newt, Gingrich, came in second with 32%.	Former US Speaker of the House, Newt, Gingrich, came in second with 32%.	Former US Speaker of the House, Newt Gingrich, came in second with 32%.	Former U.S. Speaker of the House Newt Gingrich came in second with 32%.	Former U.S. Speaker of the House, Newt Gingrich, came in second with 32%.	Former U.S. Speaker of the House Newt Gingrich came in second with 32%.	Former U.S. Speaker of the House Newt Gingrich came in second with 32%.	Former U.S. Speaker of the House Newt Gingrich came in second with 32%.	['FORMER U A SPEAKER OF THE HOUSE NEWT GINGRICH CAME IN SECOND WITH THIRTY TWO PER CENT']
2	The island was first inhabited by the Tianos and Karibeis. The Karibeis were an Arab-Lokin speaking people who had arrived around 10,000 BCE.	The island was first inhabited by the Tianos and Caribees. The Caribees were in a rock rock and speaking people who had arrived around 10,000 BCE.	The island was first inhabited by the Tianos and Caribes. The Caribes were an Arab-Wakened speaking people who had arrived around 10,000 BCE.	The island was first inhabited by the Tianos and Carribes. The Carribes were an Arahuacan speaking people who had arrived around 10,000 BCE.	The island was first inhabited by the Tianos and Caribes. The Caribes were an Arakwakan speaking people who had arrived around 10,000 BCE.	The island was first inhabited by the Tianos and Caribes. The Caribes were an Araquacan-speaking people who had arrived around 10,000 BCE.	The island was first inhabited by the Tianos and Caribes. The Caribes were an Araquican-speaking people who had arrived around 10,000 BCE.	The island was first inhabited by the Tianos and Caribes. The Caribes were an Aroquian-speaking people who had arrived around 10,000 BCE.	['THE ISLAND WAS FIRST INHABITED BY THE TIANOS AND CARIVIS THE CARIVES WERE IN ARAQUAK AND SPEAKING PEOPLE WHO HAD ARRIVED AROUND TEN THOUSAND B C']
3	This nerve impulses can be sensuquically throughout the body which helps keep the body safe from any potential choice.	These nerve impulses can be sent so quickly throughout the body which helps keep the body safe from any potential stress.	This nerve impulses can be sent so quickly throughout the body which helps keep the body safe from any potential threats.	These nerve impulses can be sent so quickly throughout the body which helps keep the body safe from any potential threats.	These nerve impulses can be sent so quickly throughout the body which helps keep the body safe from any potential threats.	These nerve impulses can be sent so quickly throughout the body, which helps keep the body safe from any potential threats.	These nerve impulses can be sent so quickly throughout the body which helps keep the body safe from any potential threats.	These nerve impulses can be sent so quickly throughout the body which helps keep the body safe from any potential threat.	['THESE NERVE IMPOLSES CAN BE SENT SO QUICKLY THROUGHOUT THE BUDY WHICH ELPS KEEP THE BIRDY THIEVE FROM ANY POTENTIAL TRUTH']
									['ON SEPTEMBER TWENTY FOURTH SEVENTEEN FIFTY NINE ARTHUR GUINIS

Word Timestamps



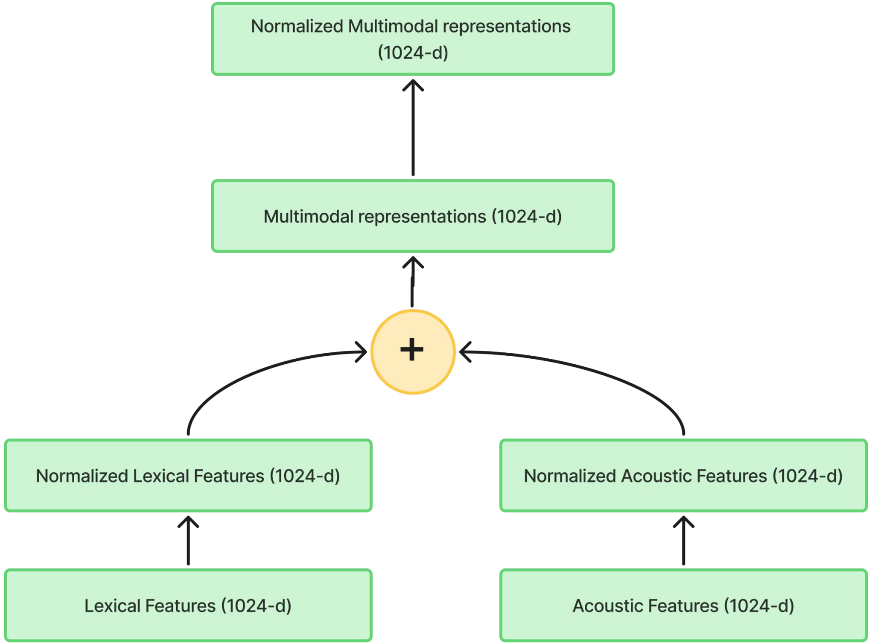
```
{"id": 1904,
 "words": [
    {"word": " However,", "start": 0.8000000000000003, "end": 1.52},
    {"word": " due", "start": 2.14, "end": 2.22},
    {"word": " to", "start": 2.22, "end": 2.42},
    {"word": " the", "start": 2.42, "end": 2.54},
    {"word": " slow", "start": 2.54, "end": 2.72},
    {"word": " communication", "start": 2.72, "end": 3.44},
    {"word": " channels,", "start": 3.44, "end": 4.14},
    {"word": " styles", "start": 4.72, "end": 5.0},
    {"word": " in", "start": 5.0, "end": 5.36},
     {"word": " the", "start": 5.36, "end": 5.5},
     {"word": " West", "start": 5.5, "end": 5.76},
     {"word": " could", "start": 5.76, "end": 6.14},
     {"word": " lag", "start": 6.14, "end": 6.38},
     {"word": " behind", "start": 6.38, "end": 6.96},
     {"word": " by", "start": 6.96, "end": 7.4},
     {"word": " 25", "start": 7.4, "end": 7.96},
     {"word": " to", "start": 7.96, "end": 8.16},
     {"word": " 30", "start": 8.16, "end": 8.46},
     {"word": " years.", "start": 8.46, "end": 8.82}
     ]}
```

Multimodal Fusion: Speech + Text embeddings

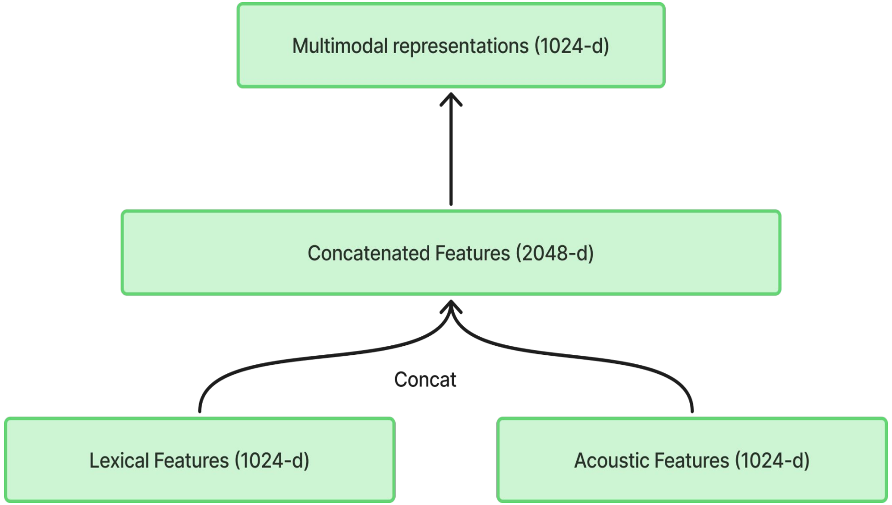


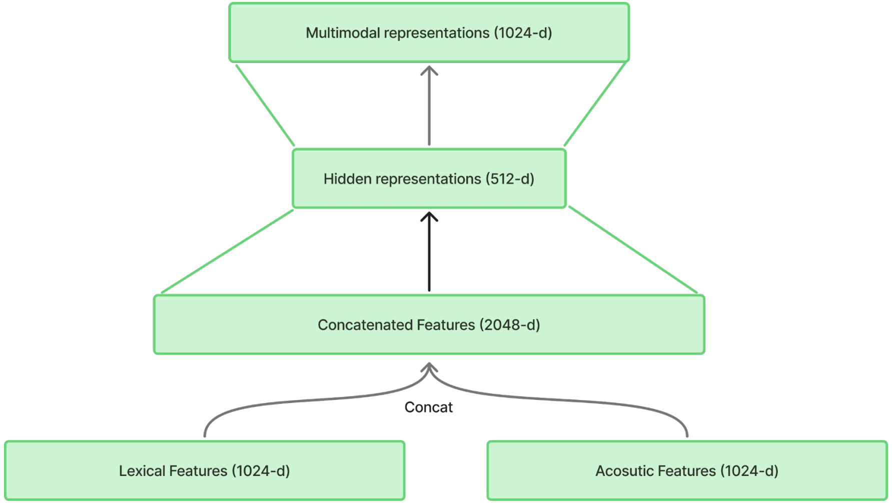
https://arxiv.org/pdf/2008.00702

Fusion Approaches - Norm and Add


Normalized Multimodal representations (1024-d)
Multimodal representations (1024-d)
+
Normalized Lexical Features (1024-d)
Normalized Acoustic Features (1024-d)
Lexical Features (1024-d)
Acoustic Features (1024-d)

Fusion Approaches - Concat and Project

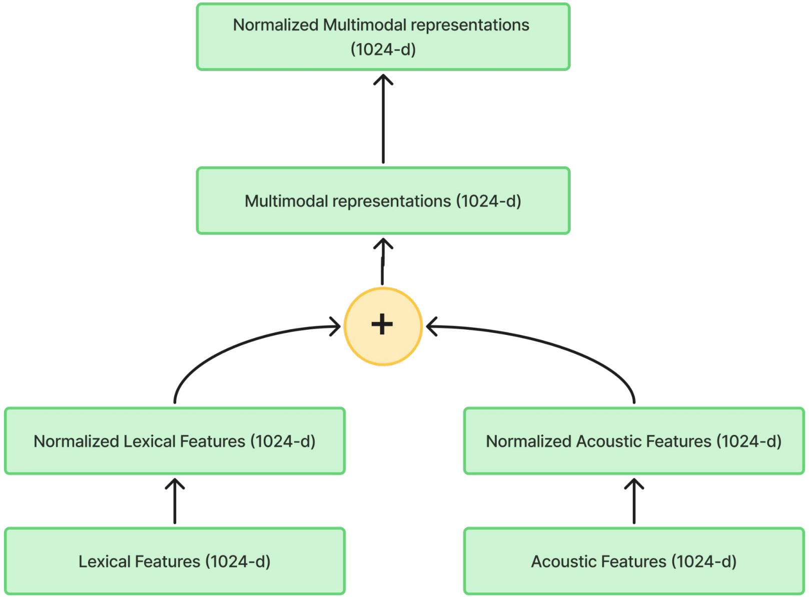

Multimodal representations (1024-d)
Concatenated Features (2048-d)
Concat
Lexical Features (1024-d)
Acoustic Features (1024-d)


Multimodal representations (1024-d)
Hidden representations (512-d)
Concatenated Features (2048-d)
Concat
Lexical Features (1024-d)
Acosutic Features (1024-d)

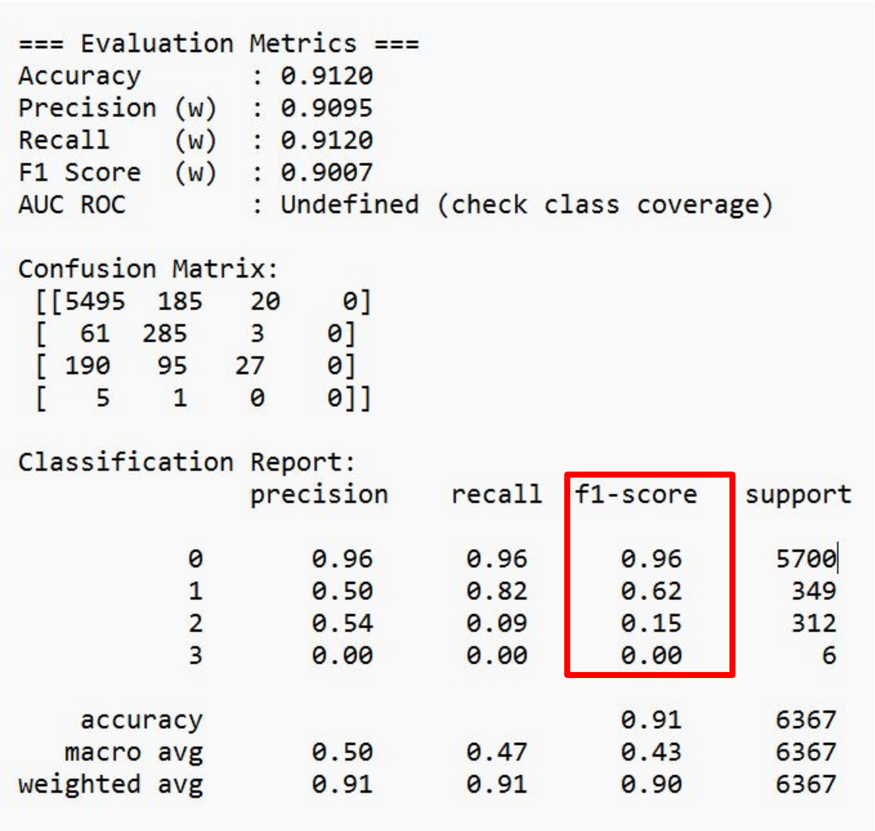
Experiments and Evaluation

Fusion techniques:

Norm and Add



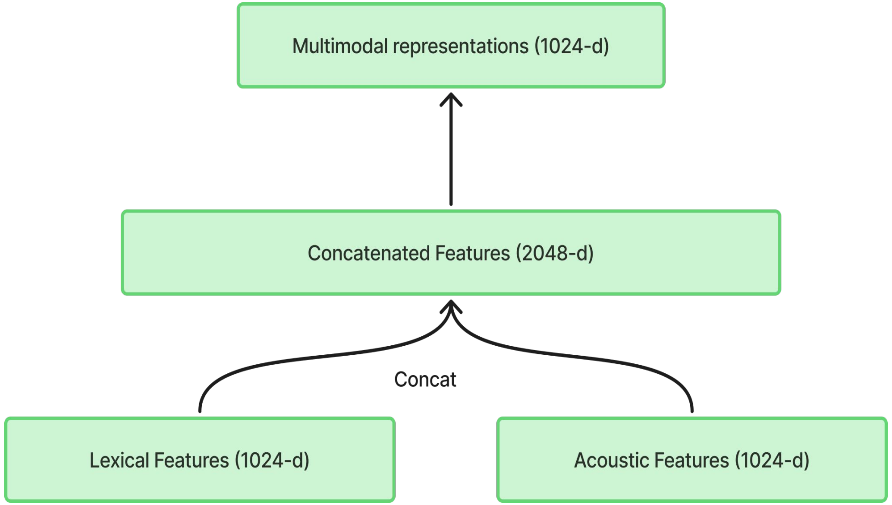
Normalized Multimodal representations (1024-d)
Multimodal representations (1024-d)
+
Normalized Lexical Features (1024-d)
Normalized Acoustic Features (1024-d)
Lexical Features (1024-d)
Acoustic Features (1024-d)

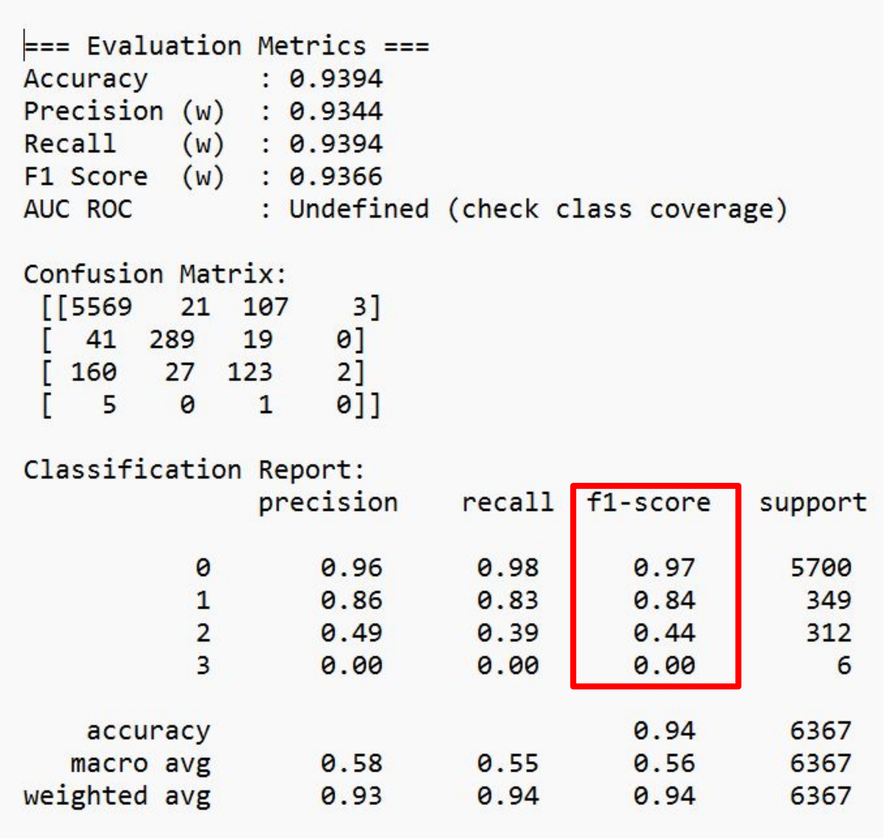


```
=== Evaluation Metrics ===
Accuracy          : 0.9120
Precision (w)     : 0.9095
Recall    (w)     : 0.9120
F1 Score  (w)     : 0.9007
AUC ROC           : Undefined (check class coverage)

Confusion Matrix:
 [[5495  185   20    0]
 [   61  285    3    0]
 [  190   95   27    0]
 [    5    1    0    0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96      5700
           1       0.50      0.82      0.62       349
           2       0.54      0.09      0.15       312
           3       0.00      0.00      0.00         6

    accuracy                           0.91      6367
   macro avg       0.50      0.47      0.43      6367
weighted avg       0.91      0.91      0.90      6367
```

Fusion techniques:

Concat and Project - 1

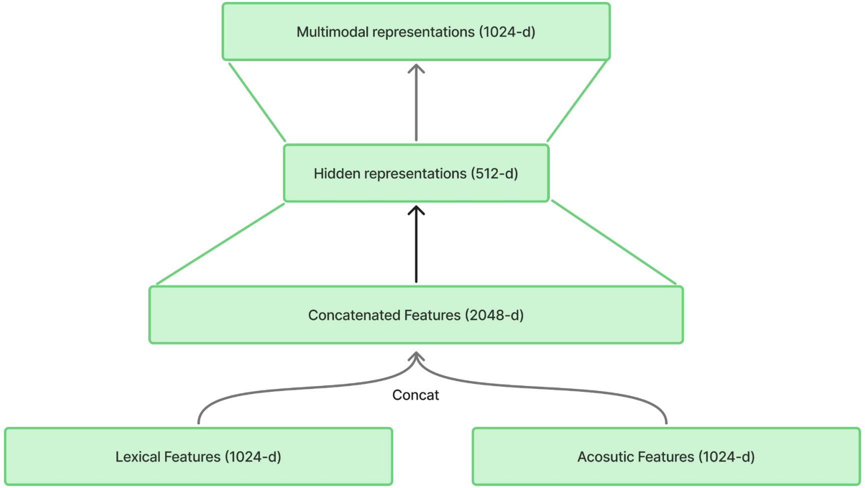

Multimodal representations (1024-d)
Concatenated Features (2048-d)
Concat
Lexical Features (1024-d)
Acoustic Features (1024-d)

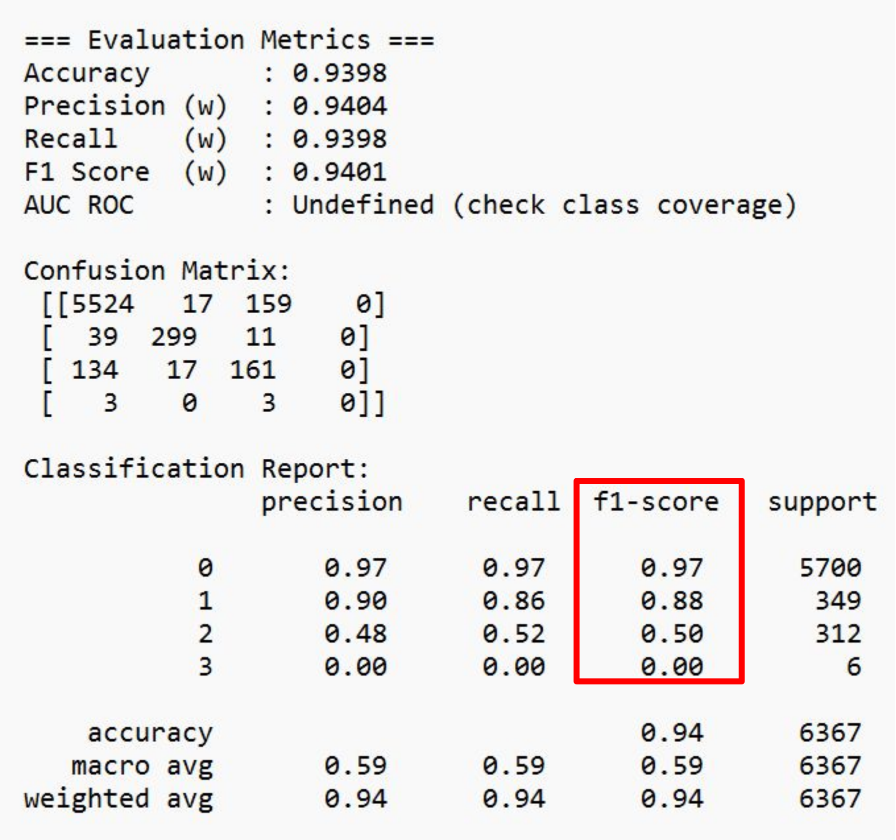


```
=== Evaluation Metrics ===
Accuracy           : 0.9394
Precision (w)      : 0.9344
Recall    (w)      : 0.9394
F1 Score  (w)      : 0.9366
AUC ROC            : Undefined (check class coverage)

Confusion Matrix:
 [[5569   21  107    3]
 [  41  289   19    0]
 [ 160   27  123    2]
 [   5    0    1    0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      5700
           1       0.86      0.83      0.84       349
           2       0.49      0.39      0.44       312
           3       0.00      0.00      0.00         6

    accuracy                           0.94      6367
   macro avg       0.58      0.55      0.56      6367
weighted avg       0.93      0.94      0.94      6367
```

Fusion techniques:

Concat and Project - 2


Multimodal representations (1024-d)
Hidden representations (512-d)
Concatenated Features (2048-d)
Concat
Lexical Features (1024-d)
Acosutic Features (1024-d)



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

Confusion Matrix:
 [[5524   17  159    0]
 [   39  299   11    0]
 [  134   17  161    0]
 [    3    0    3    0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97      5700
           1       0.90      0.86      0.88       349
           2       0.48      0.52      0.50       312
           3       0.00      0.00      0.00         6

    accuracy                           0.94      6367
   macro avg       0.59      0.59      0.59      6367
weighted avg       0.94      0.94      0.94      6367
```

In comparison..

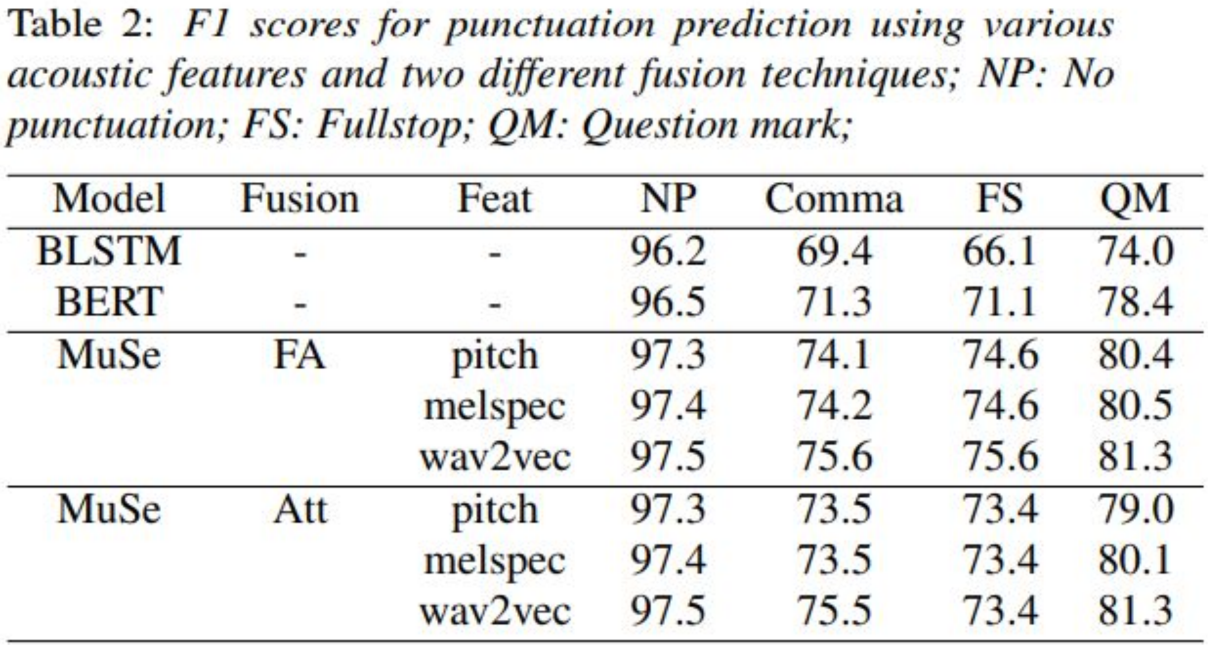
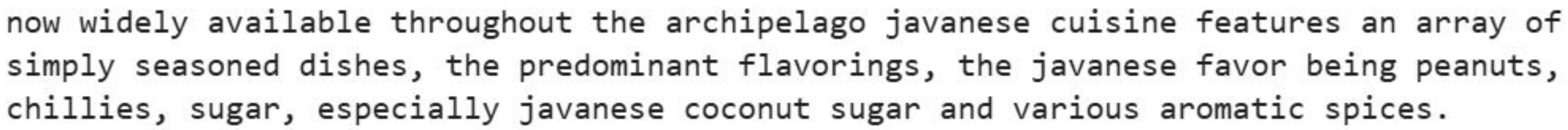


Table 2: *F1 scores for punctuation prediction using various acoustic features and two different fusion techniques; NP: No punctuation; FS: Fullstop; QM: Question mark;*

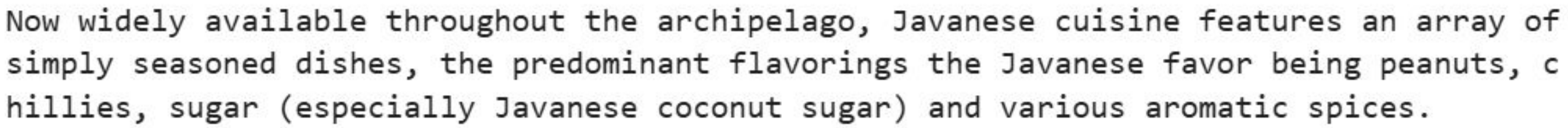
Model	Fusion	Feat	NP	Comma	FS	QM
BLSTM	-	-	96.2	69.4	66.1	74.0
BERT	-	-	96.5	71.3	71.1	78.4
MuSe	FA	pitch	97.3	74.1	74.6	80.4
		melspec	97.4	74.2	74.6	80.5
		wav2vec	97.5	75.6	75.6	81.3
MuSe	Att	pitch	97.3	73.5	73.4	79.0
		melspec	97.4	73.5	73.4	80.1
		wav2vec	97.5	75.5	73.4	81.3

M Sunkara et. al., 2020

Reference Examples

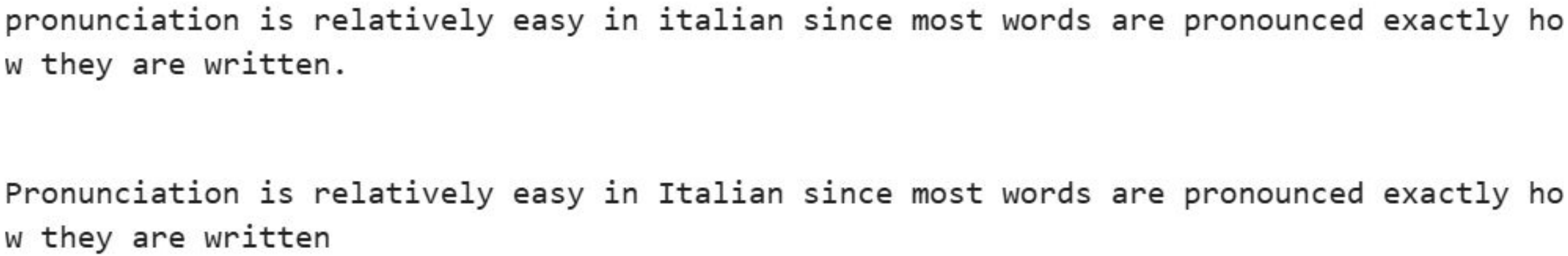


```
now widely available throughout the archipelago javanese cuisine features an array of
simply seasoned dishes, the predominant flavorings, the javanese favor being peanuts,
chillies, sugar, especially javanese coconut sugar and various aromatic spices.
```



```
Now widely available throughout the archipelago, Javanese cuisine features an array of
simply seasoned dishes, the predominant flavorings the Javanese favor being peanuts, c
hillies, sugar (especially Javanese coconut sugar) and various aromatic spices.
```

Reference Examples



```
pronunciation is relatively easy in italian since most words are pronounced exactly ho
w they are written.

Pronunciation is relatively easy in Italian since most words are pronounced exactly ho
w they are written
```

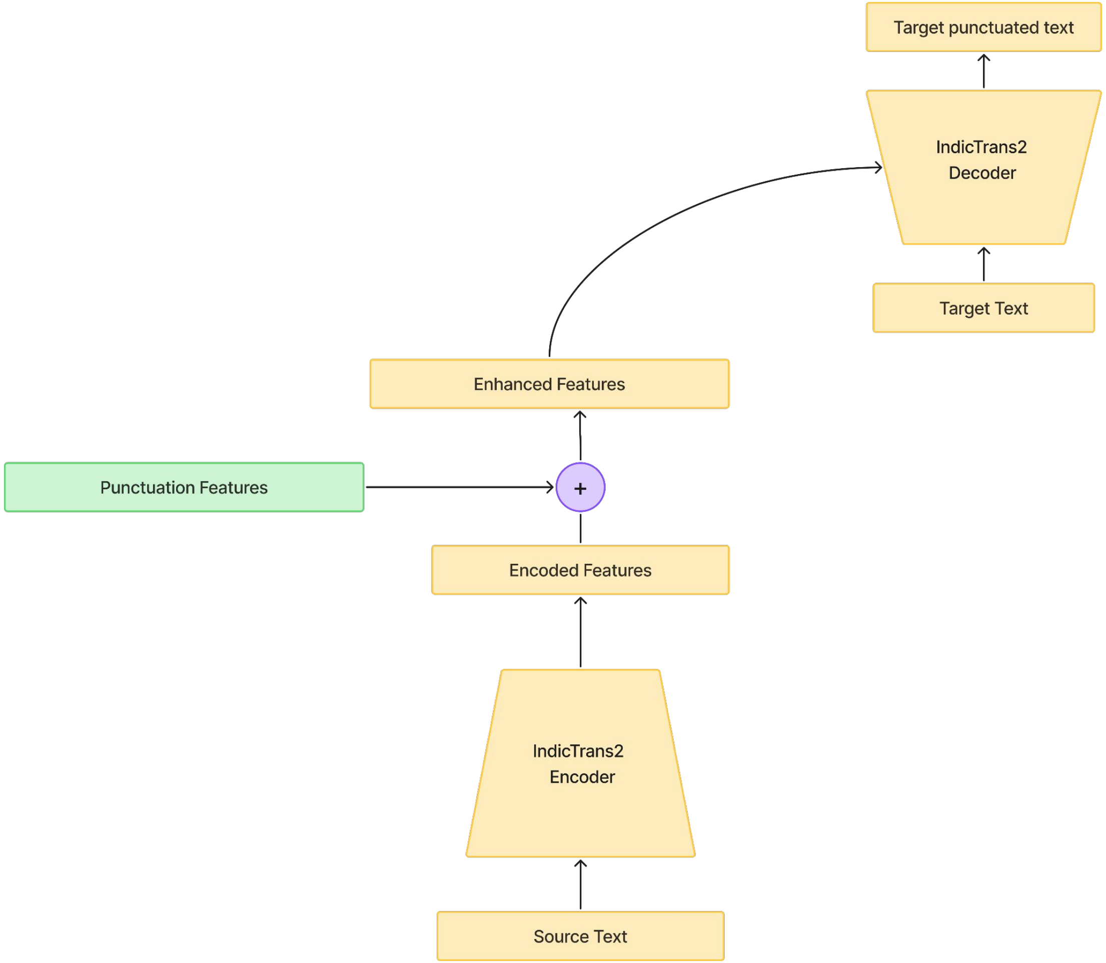
Part 2: Fused Machine Translation

Architecture

Work in progress. Debugging

Three approaches:

1. Train only last few decoder layers and keep rest of the model frozen (ongoing)
2. Train only decoder model and keep encoder frozen
3. Train full MT model



Challenges

- ★ Data cleaning and preprocessing! – unpaired cross-lingual entries, deduplication, ASR noise, length mismatch
- ★ Forced alignment! – trial and error with MFA and Whisper
- ★ Computational constraints – Broke down into cascaded steps
- ★ Manipulating new custom models such as IndicTrans 2
- ★ Experiments to converge on suitable architectures, hyperparameters

References, Code, Hyperlinks

Code:

https://github.com/pranav-satheesan/IASNLP_MM_punctuation.git

Report:

https://www.overleaf.com/project/685db3b8bb1754f61a5bc7d4

Datasets & Model Weights:

https://drive.google.com/drive/u/1/folders/1lEgtbcHJd6N5e0sbd-cMwY4mVz0NOzZz

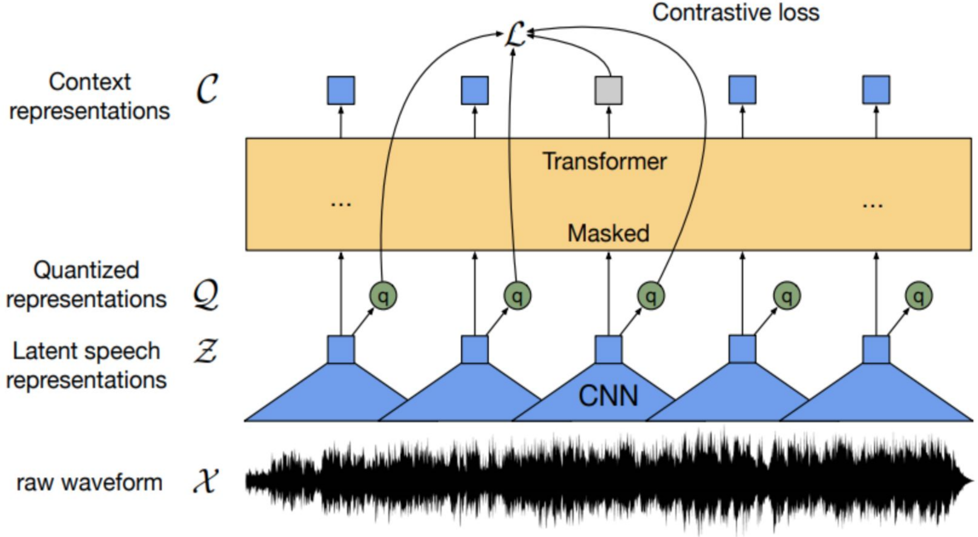
Reference Papers:

https://doi.org/10.48550/arXiv.2008.00702
https://arxiv.org/abs/2006.11477

Thank You

Wav2vec2.0 Model (Acoustic Encoder)

- Suppose our input is a one-second audio file.
- The model first downsamples the audio input using the CNN feature encoder to a shorter sequence of hidden-states, where there is one hidden-state vector for every 20 milliseconds of audio.
- For one second of audio, we then forward a sequence of 50 hidden-states to the transformer encoder. (The audio segments extracted from the input sequence partially overlap, so even though one hidden-state vector is emitted every 20 ms, each hidden-state actually represent 25 ms of audio.)
- The transformer encoder predicts one feature representation for each of these hidden-states, meaning we receive a sequence of 50 outputs from the transformer.
- Each of these outputs has a dimensionality of 768. The output sequence of the transformer encoder in this example therefore has shape (768, 50).
- As each of these predictions covers 25 ms of time, which is shorter than the duration of a phoneme, it makes sense to predict individual phonemes or characters but not entire words. CTC works best with a small vocabulary, so we'll predict characters.


Contrastive loss
$\mathcal{L}$
Context representations
$\mathcal{C}$
Transformer
...
...
Masked
Quantized representations
$\mathcal{Q}$
q
Latent speech representations
$\mathcal{Z}$
CNN
raw waveform
$\mathcal{X}$